|CURIEUX|

ACADEMIC REVIEW

Economics & Business

2026 Issue Part 2

Editing Staff

Chief-In-Editor

George Hasnah

Lead Editor

Amanuel Geremew

Ashley Du

Assisting Editors

Jing Yan Huang

Nidhi Potineni

Sophia Kong

Tony Yuan

Ridit Aggarwal

Manu Sharma

Rishi Gadikota

Eric Wang

Table Of Contents

Page 6: Controlling Markets: Democratic Socialism's Impact on Long-Term Macroeconomic Growth By Mateo Tamayo Alvarez

Page 15: An Agent-Based Simulation Framework for Modeling Household Debt Accumulation and Structural Wealth Inequality: A Case Study of Thailand By Jenny Loungjornpibool [1] and Komsan Suriya [2]

Page 37: The Global Market for the "Caviar of the East"Ethics, Regulation, and Sustainability By Ilinca Harja-Iliescu

Page 56: Improving Option Pricing Accuracy Using a Hybrid Machine Learning and Black-Scholes Model By Vyshal Sreenivas

Page 66: Election Cycles and Short-Term Fiscal Policy: A Comparative Study Between Parliamentary and Presidential Systems By Ginevra Colapinto

Page 80: Education Quality and Youth Unemployment: Evidence from a 30-Year Cross-Country Panel By Samit Mandhare

Page 95: Young Women and Grocery-Brand Marketing Tactics By Macy Hart

Page 116: Do Mega-Events like the LA28 Olympics Create New Long-Term Benefits for Host Cities when Considering Economic, Social, and Environmental Impacts Together? By Matthew Chiang

Page 127: Evaluating the Effects of Spain's National Public Transport Pass on Congestion-Related Economic Losses in Madrid and Barcelona By Priya Barke

Page 148: China's Pet Economy: Development, Digital Growth, and Current Challenges By Xinhe Wang

Page 156: Volatility Regimes and the Dispersion–Return Relationship: Testing for Herding in U.S. Equities During the 2007–2009 Crisis By Hrishik Agarwal

Page 170: The Game Theory Behind the Evolution of the Korean Entertainment Industry By Isaac Ahn

Page 178: The Price of Performance: The Inefficiency in Baseball's Payrolls By Jyotish Jotwani

Page 186: A Fragile Foundation: How Inequality, Financial Collapse, and Government Failure Shaped the Great Depression By Joseph Akilov

Page 192: Impact of the Russia–Ukraine War on Food Inflation in West Asian Countries: The Roles of Wheat-Import Concentration and Self-Sufficiency By Shayan Shamsi Pour Siahmazgi and Ariel Goldman

Page 200: Median Income Regression Model By Siddhant Jha

Page 206: To What Extent Does Participation in FBLA Correlate with Financial Literacy among High School Students? By Abhyuday Shashi

Page 215: To What Extent Does Impulse Buying and Social Influence Lead to Irrational Spending in the Fashion Industry? By Alexandra Filip

Page 234: Supply Chain Risk Management: Case Studies and Strategies for Resilience By Aarav Agarwal

Page 249: The Effect of the COVID-19 Pandemic on U.S Financial Companies By Pitt Sinonpat

Page 265: Is It Possible and Desirable to Raise Taxation on the Top 1%? By Jayden Tran

Page 269: Economic Aspects of Sustainable Solutions In Data Centres By Isabel Shiju

Page 279: Perceived AI-Likelihood, Disclosure, and Consumer Responses to AI-Generated Fashion Imagery: A Between-Subjects Pilot Study By Gabriella Messina

Page 295: What Are the Best Economic Predictors for Armed Conflict and Peace? By Wang Qisheng

Page 310: The Shahed Drone and the Economics of Modern Warfare By Siddhanth Dayanand

Page 319: Did Republican Counties Face Greater Retaliatory Tariff Exposure in 2025? By Jubili Bamzai-Wokhlu

Page 337: Quiet on the Set: A Task-Based Analysis of the Impact of AI on Workers in the US Motion Picture and Video Industries By Tyler Frank [1]*, Mentor By Kenneth Frank[2], Jane Park[3]

Page 357: Inflation as a Structural Barrier to Our Global Economy: Lessons from Argentina, The United States, and Countless Geopolitical Shocks By Keya Kumar

Page 368: Exchange Rate Fluctuations and Domestic Gold Prices in India (2015–2025): A Year-on-Year Analysis By Aditya Singh

Page 375: When the Proxy Loses Its Referent Mileage, Battery Health, and Signal Reconstruction in the Used Electric Vehicle Market By Jiwon Chung

Page 387: Comparative Empirical Descriptive Study of Unemployment Trends in the United States and Germany, 2005–2024 By Zayaan Ali Pallimalil

Page 398: How Have Fintech Platforms Changed the Accessibility and Behaviour of Retail Investors in Stock Markets? By Parv Anand

Page 409: Denim Wars: How Celebrity Endorsement Influences People and Shapes Advertising and Brand Culture By Mariia Tarasiuk

Page 419: Beyond Translation: Barriers to Tax Compliance Among Limited English Proficient Immigrant Taxpayers By Isaac Kwun

Page 431: AI-Driven Recommendations and Consumer Financial Decision-Making: A Literature Review and Survey-Based Investigation By Aarna Bafna

Page 455: A Comparative Analysis of ATP Masters 1000 Event Marketing in China and the United States By Isaac Zhang

Page 463: The Architecture of Time: Horizons, Constraints, and the Quality of Business Decision-Making By Riley Popovich

Page 477: Beyond the Affordable Housing Label: Rent-Based Classification, Cyclicality, Growth, and Valuation in U.S. Multifamily REITs By Kavin Krishnarajah

Page 492: College Debt: Institutional Factors By Lalith Durbhakula

Page 511: The Impact of the African Growth and Opportunity Act (AGOA) on Sub-Saharan African Economies By Binh Minh Tran

Page 522: Beyond the Wage Gap: Explaining the Persistence of the Gender Wealth Gap in the United States By Hudson Waiting

Page 527: AI and the Changing Role of Human Workers By Raunit Sharma

Page 533: The Market Potential and Commercial Value of Accessible Design: Business Opportunities in Products and Services for People with Disabilities By Yuzheng Dong

Page 545: The Effects of the Padel Industry on Pakistan's Sporting Industry By Khazain Hassan Mansha

Controlling Markets: Democratic Socialism's Impact on Long-Term Macroeconomic Growth By Mateo Tamayo Alvarez

Abstract

Democratic socialism isn't just a political ideology; it is an economic system that pairs socialist ideologies in economics with democratic political systems. In macroeconomic terms, democratic socialism emphasizes government intervention, higher taxation, a welfare state, and many other methods. This literature review surveys work on how democratic socialism impacts long-term macroeconomic growth. Overall, the research debates the effects of democratic socialism on macroeconomic growth as well as the political impacts. Some research concluded that democratic socialism inhibits economic growth. Other research suggests that democratic socialism contributes to economic growth. In addition, some studies look at the political effects of democratic socialism.

Introduction

High taxes are often seen as negative. Yet in the Scandinavian countries, which are some of the happiest in the world, taxes are relatively high (World Bank Open Data; World Happiness Report). The high taxes found in the Scandinavian countries are a mechanism of a political and economic philosophy known as democratic socialism. In macroeconomic terms, democratic socialism emphasizes government intervention, higher taxation, and a welfare state, among other things. Democratic socialism has silently risen as an alternative to both the liberty of free-market capitalism and the control of communism, acting as a sort of middle ground that offers the best of both worlds. While still offering the economic benefits of capitalism, democratic socialism also has the power to control industries and regulate the market. Democratic socialism is a political philosophy as much as it is an economic policy plan. This system often has support of backlash towards its political side.

How does democratic socialism affect long-term macroeconomic growth? This literature review surveys work on this question. Overall, the research debates the effects of democratic socialism on macroeconomic growth as well as the political impacts. Some research concluded that democratic socialism inhibits economic growth. Other research suggests that democratic socialism contributes to economic growth. In addition, some studies look at the political effects of democratic socialism.

This literature review has three sections. The first and second look at the negative and positive economic effects respectively. The third looks at the political implications. The research spans topics from welfare spending to regulation and productivity. Findings from these suggest that taxation and welfare spending lead to higher growth. These three sections serve to show opposing views and different methods on the macroeconomic and political performance of democratic socialism. While there are pieces that seemingly contradict each other, it's important to look at democratic socialism considering its possible effectiveness in some countries, including Norway and Sweden, and its rising popularity in capitalist nations like the US.

Negative Economic Effects

There are numerous studies that suggest democratic socialism has negative effects on a country's macroeconomic performance. Various thinkers concluded that this system reduces productivity and connected this finding to periods of slow economic growth and production. Others have found that core socialist policies such as regulation have led to astronomical decreases in GDP that amount to trillions of dollars.

Some research suggests that transfer payments hinder economic growth. The cross sectional approach taken by Helms and Jay argues that transfer payments are one way in which democratic socialist policies inhibit growth. In their research, they gathered data ranging from 1965-1979 to examine the effect on economic growth when local and state taxes were being used to fund certain projects and necessary societal industries. Their findings indicate that the raising of taxes, most notably property taxes, in order to fund transfer payments, does have a negative effect on growth. They used models from previous studies and came up with their own natural log equation to reach this conclusion. Democratic socialism's use of high taxation in order to fund transfer payments which they argue actually reduces macroeconomic growth (Helms and Jay 575, 578-579). This finding goes hand in hand with the idea that democratic socialism reduces macroeconomic growth.

Other studies have concluded that increasing government regulation leads to a decrease in economic growth. Dawson and Seater start by introducing regulation as one of the more recent topics in macroeconomists. In their approach, the authors gathered all regulations from the *Code of Federal Regulations* in order to have accurate data on all regulation guidelines during the period of study. The authors argue that their data set is much more relevant and reliable because they added to data sets already in existence. Their findings conclude that regulation has had a significantly negative impact on GDP growth by an average of around two percentage points during the period 1949-2005. This accounts to more than a 38.8 trillion dollar loss as of 2011. Furthermore, they argue that this could explain the period of stagflation in the 1970s. This study has demonstrated that the impacts of regulation on total output during this period were negative for growth (Dawson and Seater 138,140). Their mathematical model showed reduction in not only output, but also productivity.

Similarly, various studies suggest that democratic socialism reduces productivity. Dawson and Seater concluded this same point in their research. When analyzing the productivity slowdown of the 70s, they observed that total factor productivity (TFP) declined as regulation increased. Their work clearly shows the steady decline from just after the period of WWII to the mid 70s. They imply that regulation and taxes were the cause for this decline. They observed that the "productivity slowdown" (1965-1980) happened just as regulation increased. In addition, they also include that TFP started increasing just as regulation increased, accompanied by the marginal tax rate. The path to the peak, right before 1980, was almost always rising, which is attributed to American economic policies at the time. Before former president Ronald Reagen was elected, the US was in an era of heavy government intervention in the American economy. The TFP decrease, they suggest, is a direct result of the increase in intervention and regulation.

Furthermore, Nicoletti and Scarpetta also found a similar conclusion: regulatory policies limit productivity growth through various pathways including reducing technological innovation and technological spillovers. To test their hypothesis, Nicoletti and Scarpetta looked at the policies that impacted competition rather than try to measure competition itself. Their main aim, though, was to further expand and explain the "regulation-productivity link". They used three Multi Factor Productivity (MFP) regressions. It's important to note that MFP and TFP refer to the same economic concept. These regressions are included because of limitations in the data set concerning regulation. Their regressions results, which also suggest a negative correlation between regulation and productivity, suggest a few implications. First, regulation reduced competition and private governance negatively affecting productivity because it slowed down the technological catch-up, which is what Nicoletti and Scarpetta argued leads to growth via an increase in the MFP. Second, they found that there is a positive effect in privatizing the manufacturing and service industry. Finally they found a positive effect of privatization on other specific industries. Overall, Nicoletti and Scarpetta's findings came to the same conclusion as Dawson and Seater: increases in regulation lead to a decline in TFP/MFP. Together, these two research papers attack a vital aspect of democratic socialism (Dawson and Seater 164, 166, 168; Nicoletti and Scarpetta 7, 37, 42, 45, 48). Democratic socialism calls for heavier government regulation and control of enterprise which, despite being meant as something to protect the people, has shown to negatively affect industry and, hence, growth.

Other work finds that democratic socialism works due to its incorporation of free market traits and not socialism itself. In his work, Murphy analyzes cases of socialism that seem to be performing well economically. He performed case studies of China, Israel, and most importantly, Hungary, in the year 1985. He chose this year because the EFW index had a large number of countries being graded while there was still a very prevalent socialist influence in the world. Hungary, especially, is a very interesting case due to its location in the eastern pact while on the border of the iron curtain. Under the leadership and influence of communism and Janos Kadar respectively, Hungary was wealthier and more liberal compared to other states in the same situation. Despite having had a communist state for over 40 years, Hungary still had a very capitalist mentality in the sense that there was an emphasis on materialism as a whole despite it being a core evil of communism. Nevertheless, Hungary's goulash communism, while technically a socialist state, already had that western, capitalist mentality which saw it greatly increase most of its socioeconomic factors in a 30 year span. Murphy is able to conclude, all in all, that goulash communism in Hungary already had a capitalist mentality which contributed to an increased pace of liberalization after leaving the Eastern Bloc and becoming a state with more democratic and free-market traits (Murphy 286, 291).

Positive Economic Effects

Despite various findings that seem to portray democratic socialism as an unhealthy partner of macroeconomic growth. Multiple findings suggest that the same things that the aforementioned findings concluded were negative are actually positive. Some research suggests

that higher taxation and welfare spending leads to positive economic growth. Others have argued that having socialist foundations can be positive even if socialism isn't prevalent. Some pieces point to the idea that democratic socialism helps protect the economy during periods of recession through numerous mechanisms that include strong unionization throughout a nation. Democratic socialism is often talked about as a system that limits the economy relative to the free-market system. Hence, it's important to observe how that might not necessarily be the case.

Another body of literature suggests that higher taxation may lead to higher economic growth. High taxation is a vital element of democratic socialism and is used to fund projects, industry, infrastructure, and many other elements. Helms, whose piece on the effect of taxes on growth rejects the positive effect of taxation when funding transfer payments, had an interesting finding: taxation does lead to economic growth when used to fund government services such as education, highways, and health care. These three services all had a net positive effect on short-term growth. Furthermore, Katz et al concluded that there isn't a tradeoff between having an active government with high taxes, intervention, and regulation, and a growing economy (Helms 579; Katz et al 884). Overall, these two pieces found that democratic socialism's high taxation strategy does work. High taxation was found to, at the very least, not retard growth, as critics of democratic socialism claim.

Thinkers also conclude that high government welfare spending leads to higher macroeconomic growth. This school of thought based on ideas such as universalism, strong labor laws, and protective labor laws, concludes that increased welfare spending leads to higher economic growth. Madni's and Chaudhary's case study of Pakistan from the period 1984-2015 analyzed the relationship between growth and fiscal spending paired with effectiveness of institutions. They observed increases in economic growth when both government spending and institutional effectiveness were higher. They concluded that government spending and institutional development are necessary in order to foster economic growth. This finding goes hand in hand with the pillars held by the Nordic Model of democratic socialism. This concept is explained further in Iacono's research. Iacono provides a framework with three features of which the third, public welfare spending, is of particular relevance. His point theorizes that a narrow income gap between all workers leads to a majority wanting and needing the same things from the state. This, then, leads to more people wanting high public welfare spending because a large number of citizens would benefit, instead of a rich minority having to fund welfare benefits for the lower classes, as is the case in most societies. As a result, while not explicitly stating that high public welfare spending leads to growth, Iacono creates a logical link between the immediate effect of the welfare spending and how that in turn, leads to a higher demand in welfare benefits. Iacono also suggests that this high degree of wage differential compression leads to "creative destruction". This destruction is meant for effective and efficient companies to survive while having low paying, ineffective companies die out, logically leading to economic growth. Together with Madni's and Chaudhary's work, Iacono emphasizes that high welfare spending leads to growth (Madni and Chaudhary 120, 122; Iacono 185, 187).

Other research suggests that strong labor unions with high participation lead to higher productivity. The Nordic Model of democratic socialism emphasizes this idea of strong unionization which is implied in Iacono's work. Iacono explains that strong unionization and high participation rates lead to relatively high degrees of wage differential compression. In other words, the pay gap between high skilled workers and low skilled workers isn't large at all (Iacono 185). Joshi and Navlakha expand on this point and confirm it by first explaining that labor unity is strongly instrumental in expanding freedom. While economic growth isn't the aim of their research, they include data on labor unions for Scandinavian countries with social democracies. The average percent of union workers was 73 percent for countries with the Nordic Model (Joshi and Navlakha 75). Moreover, Hicks expands on past research by Lange and Garrett. He mentions wage restraint in unions can lead to economic growth through likely increases in profits, investments and growth. To simplify this, Hicks implies a high number of union members being controlled by large workforce unions that can collectively ask for higher wages. With this comes a caveat, however: if most workers are in a union, they can ask for exceptionally high wage increases. This has a very negative effect however, as the economy would decline quickly. Yet, by practicing wage restraint, Hicks finds that the economy can grow because of high union participation rates (Hicks 684).

Some research suggests that democratic socialism creates a welfare state that over time transforms the economy positively. Joshi and Navlakha's case study of Sweden's freedom in a social democracy included the "Universal welfare state" as one of the main ways that Sweden enhances freedom and explains that, in the end, it ends up continuing to grow the economy because of how beneficial it is to the people of Sweden. Their work incorporates various living standard statistics that rank Sweden among the top in the OECD, UN, and EU. They link these astonishing statistics to Sweden's determination to invest heavily into infrastructure, connectivity, and education. As a result of Sweden's heavy investment into material freedom, they find that Sweden's universal welfare state has led to increases in productivity and trust in the government, overall boosting the economy (Joshi and Navlakha).

Some work shows that having a socialist central planning policy (SCPP), even as a foundation, has long term positive economic development. This argument is especially prevalent in cities that were under Soviet influences during the Cold War, as was the case of the German Democratic Republic (GDR), more commonly known as East Germany. Rosenfeld and Heider conducted a case study on the cities of Chemnitz, Rostock, Suhl, and Weißwasser. Due to limitations in data, they conducted various interviews with locals and experts in a wide array of industries to find the most notable changes before and after German unification. Chemnitz and Rostock were two mid-major cities in the GDR that held important industrial, trade, and administrative functions for the east. Compared to Suhl and Weißwasser, Chemnitz and Rostock are historically more important and have more industry and a larger workforce to support respective industries. As a result, the GDR named both of these cities regional capitals, investing heavily into their industries. Chemnitz had three major industries and Rostock had a large industrial maritime presence. Following the return of the market economy, both these cities lost

most of their administrative functions, but revitalized their former SCPP industries. While Chemnitz and Rostock are now thriving cities, Suhl and Weißwasser haven't experienced this regeneration despite having some of the same characteristics and investment by the GDR (Rosenfeld and Heider 318-320). All in all, SCPP has shown that socialism, even if not present, can still have a great impact in the long term development of cities, but it doesn't mean it will always work out in this fashion. Yet, it shows how pairing market economies with socialist industry, which democratic socialism does to an extent, can have success in economically developing cities in the long run.

Research supports the idea that democratic socialism protects the economy in times of economic crisis relatively better than free-market economies. Democratic socialism inherently has more control of industry and hence can more or less protect the industries from recessionary risk. This is opposite to free market economies that are more privatized resulting in crises hitting them harder relative to democratic socialist nations. Joshi and Navlakha made this exact conclusion measuring unemployment rates in 2009. In this year, Sweden's unemployment rate (8.3 percent) was lower than both the USA (9.3 percent) and the EU average (8.9 percent). Paired with the widespread, aforementioned Swedish welfare state that offers around 70 to 80 percent of one's previous salary for up to a year if unemployed, Sweden and other countries, especially Scandinavian countries, were shown to greatly reduce the impact of recession compared to capitalist countries (Joshi and Navlakha 77). As a result, democratic socialism presents a system that can still offer great economic freedom to its citizens while protecting them from the volatility of capitalism.

Political Implications

While numerous studies concentrate on the economics of democratic socialism, other works examine the political effects. There is debate about whether or not freedom is increased or decreased in this system. Some argue the benefits and others that it can lead to dictatorship. Analyzing these views is interesting because it is a field of contradiction and connects the political to the economic.

Some view that democratic socialism as a system for the people. Democratic socialism has the word "democratic" in its name for a reason. Wright describes socialism as a movement that fights for things that now seem more relevant than ever. More technically, he describes socialism as an economic policy that allocates and employs resources for different social purposes through social power. He then goes on to explain social power as the power to gather people to mobilize for voluntary collective actions in civil society. Furthermore, he explains that the state power is directly subservient to the power of social power (Wright 154). This logical connection between state power and social power is used to demonstrate why democratic socialism is a system for the people. The link between state power and the people is further emphasized in Pena's work on sustainable socialism. When describing the political system of sustainable socialism, he argues that it must be a people's democracy (Pena 293). The people's democracy is progressive and is there for the long run goals of society. This is the simple

definition of democracy. However, it differs from just a democracy in that the state actually has to allocate resources, choose where to spend them, and also be directly responsible for production. Overall, Wright and Pena's findings conclude that socialism has to be democratic as that is the political system that the people can have power over while still achieving the long term goals of society as a whole.

Others think that democratic socialism reduces personal economic freedom. In his work Pena analyzes what he believes is needed to have sustainable, successful, and long term socialism. He postulates four pillars of sustainability which are needed in his view: political, economic, national cultural characteristics, and environmental. The economic system is particularly interesting as he believes it requires sacrifice in time, freedom, and money from the people it benefits. The economic system is there to serve the people's agenda and promotes long-term sustainability instead of short-term profit maximization. Pena suggests that an economy that does not work for the people will eventually be put to an end by the same people for not being able to recognize its worth. As a result of this economic system that Pena argues is fundamental for long term sustainable socialism, the people in this system sacrifice some of their wants for the greater good of the state (Pena 293, 294). While arguably beneficial for the people, it does mean that their economic freedom is limited *in the short term*. For example, high taxes may in the short run reduce personal spending, but in the long run provide a larger net benefit.

Other literature also suggests that democratic socialism actually goes hand in hand with increasing and protecting freedom instead of inhibiting it. Socialism is often portrayed as having higher state control over its people. As a result, socialism is associated with reducing freedom as a whole. While that is the case with previous socialist states, such as the USSR, modern democratic socialism argues otherwise. Love asserts that, while the above is true, there needs to be new definitions of both freedom and socialism. He states that the most common definition of freedom is negative liberty which Isaiah Berlin describes as, "I am normally said to be free to the degree to which no man or body of men interferes with my activity. Political liberty in this sense is simply the area within which a man can act unobstructed by others". On the contrary, Love proposes the Kantian concept of freedom. Rather than just having unobstructed freedom as in negative liberty, this concept of freedom grants the right to act as you want in your own will without interfering in others' wills. Furthermore, Love proposed that the right to freedom encompasses a right of self-governance. Here, everybody would be inherently equal with the government being the glue that holds everything together. The government protects the rights of freedom that everybody shares equally. Despite previous socialist states not promoting freedom, these modern democratic socialist systems of governance, Love emphasizes, can promote and secure freedom while still having socialist values deeply engrained (Love 132, 134, 149).

Conclusion

This literature review has analyzed two opposing views of democratic socialist traits on economic performance whilst also analyzing the political consequences that it brings. The research is varied surrounding the effects of democratic socialism on macroeconomic growth

through both an economic and political lens and there isn't a clear answer on which is better. However, it does offer perspective, which is always needed when analyzing certain economic situations and political changes. For example, while a country looking to adopt higher taxation might receive backlash, if one pairs it with high welfare spending, then it might boost economic performance while still increasing freedom and socioeconomic benefits for citizens.

Avenues for further research include the impact of long-term leadership in helping the economy. Other research can focus on the steps for welfare implementation and improvement. On the political side, research can focus on democratic backsliding in democratic socialism. Overall, this literature review analyzes democratic socialism's effect on the economy and politics.

Works Cited

Dawson, John W., and John J. Seater. "Federal Regulation and Aggregate Economic Growth." Journal of Economic Growth, vol. 18, no. 2, 2013, pp. 137–177. JSTOR, https://www.jstor.org/stable/42635321.

Helms, L. Jay. "The Effect of State and Local Taxes on Economic Growth: A Time Series–Cross Section Approach." The Review of Economics and Statistics, vol. 67, no. 4, 1985, pp. 574–582, https://www.jstor.org/stable/1924801

Hicks, Alexander. "Social Democratic Corporatism and Economic Growth." The Journal of Politics, vol. 50, no. 3, 1988, pp. 677–704. JSTOR, https://www.jstor.org/stable/2131463.

Iacono, Roberto. "The Nordic Model of Economic Development and Welfare: Recent Developments and Future Prospects." Intereconomics, vol. 53, no. 4, 2018, pp. 185–190. https://www.intereconomics.eu/contents/year/2018/number/4/article/the-nordic-model-of-economic-development-and-welfare-recent-developments-and-future-prospects.html

Joshi, Devin, and Neha Navlakha. "Social Democracy in Sweden." Economic and Political Weekly, vol. 45, no. 47, 2010, pp. 73–78. JSTOR, https://www.jstor.org/stable/25764156.

Katz, Claudio J., Vincent A. Mahler, and Michael G. Franz. "The Impact of Taxes on Growth and Distribution in Developed Capitalist Countries: A Cross-National Study." The American Political Science Review, vol. 77, no. 4, 1983, pp. 871–886. JSTOR, https://www.jstor.org/stable/1957563.

Love, S. M. "Socialism and Freedom." Philosophical Topics, vol. 48, no. 2, 2020, pp. 131–157. https://www.pdcnet.org/philtopics/content/philtopics_2020_0048_0002_0131_0157

Madni, Ghulam Rasool, and Muhammad Aslam Chaudhary. "Economic Growth in Context of Institutions and Fiscal Policy." Pakistan Economic and Social Review, vol. 55, no. 1, 2017, pp. 79–98. JSTOR, https://www.jstor.org/stable/26730215.

Murphy, Ryan H. "The Best Cases of 'Actually Existing Socialism.'" The Independent Review, vol. 23, no. 2, 2018, pp. 283–295. JSTOR, https://www.jstor.org/stable/26591784.

Nicoletti, Giuseppe, and Stefano Scarpetta. "Regulation, Productivity and Growth: OECD Evidence." Economic Policy, vol. 18, no. 36, 2003, pp. 9–72. https://academic.oup.com/economicpolicy/article-abstract/18/36/9/402823

Pena, David S. "21st-Century Socialism and the Four Components of Sustainability." World Review of Political Economy, vol. 1, no. 2, 2010, pp. 290–304. JSTOR, https://www.jstor.org/stable/41942921.

World Bank Group. “World Bank Open Data.” World Bank Open Data, https://data.worldbank.org. Accessed 12 Aug. 2026.

World Happiness Report. https://data.worldhappiness.report/table. Accessed 12 Aug. 2026.

Wright, Erik Olin. “Socialism as Social Empowerment.” Berkeley Journal of Sociology, vol. 50, 2006, pp. 147–67. JSTOR, https://www.jstor.org/stable/41035615

An Agent-Based Simulation Framework for Modeling Household Debt Accumulation and Structural Wealth Inequality: A Case Study of Thailand

By Jenny Loungjornpibool [1] and Komsan Suriya [2]

1. Abstract

Household debt accumulation and its compounding effects on wealth distribution present highly complex structural challenges in developing economic landscapes. Rather than relying on traditiona retrospective empirical data, this study introduces a MATLAB-based agent-based computational framework to simulate the long-term interaction between micro-level income volatility and macro-level wealth inequality. The model constructs a simulated economy with 500 heterogeneous agents divided across five distinct socioeconomic categories parameterized with baseline metadata from the World Bank and the National Statistical Office of Thailand. Through a 100-month computational timeline, the system models five distinct stress-test configurations, including localized income disruptions and targeted safety-net policy mechanisms. This framework shows a clear pattern. When cash shortages happen again and again, the gap in wealth grows larger. Over time, this shifts how debt spreads across the entire system.In the end, this paper offers a reliable test model. Researchers can use it to study and predict big economic trends under strict rules of human behavior.

Keywords Household debt; income inequality; agent-based modelling; Thailand; computational framework.

2. Introduction

Household debt has increased rapidly in Thailand and has become a significant concern for both policymakers and economists, reaching nearly 95.5% of gross domestic product (GDP), making it one of the highest household debt levels in Southeast Asia (Bank of Thailand, 2022). While borrowing can help households manage temporary financial difficulties, it becomes problematic when they rely on debt for extended periods or use credit to finance routine consumption. Over time, this reliance can reduce households' ability to save, build wealth, and recover from future economic shocks, raising concerns that debt may widen inequality among lower-income households.

Thailand is a useful calibration case for examining the link between household debt and inequality because both problems remain structurally persistent. Household debt reached 95.5% of GDP in 2021Q1 (Xu, 2025), while Thailand's income Gini coefficient, which measures income inequality on a scale from 0 for perfect equality to 1 for perfect inequality, remained at approximately 0.433 in 2019 (World Bank, 2022). During the pandemic, severe income losses forced many Thai households to rely more heavily on loans and credit to cover basic living expenses (Bank of Thailand, 2023). This raises a core computational question: *to what extent can*

[1] International Community School
[2]Faculty of Economics, Chiang Mai University

an algorithmic agent-based model replicate the cascading financial vulnerabilities and structural debt accumulation observed under simplified, model-generated income shocks that represent differences in income volatility across household types?

Previous literature establishes that income shocks exacerbate household leverage and systemic financial fragility (Iacoviello, 2008; Kumhof et al., 2015; Cardaci & Saraceno, 2018). However, contemporary research predominantly relies on aggregated national accounts and abstract theoretical frameworks, obscuring the high-frequency, monthly dynamics of debt accumulation within heterogeneous populations experiencing repeated economic shocks. This creates a methodological gap: existing evidence identifies the relationship between debt and inequality, but it less clearly shows the step-by-step mechanisms through which income instability produces different debt and wealth trajectories across household types. Households do not respond to financial shocks uniformly. A household with a stable salary may recover relatively quickly from an unexpected expense, whereas a household with irregular income may need to borrow repeatedly to cover basic living costs. Understanding these differences is important for explaining why debt accumulates more rapidly for some households than for others. This study addresses that gap by using an agent-based simulation to model how household-level income volatility and expenditure shocks compound over time. This leads to the central research question:

To what extent can an agent-based simulation effectively map how individual income instability drives long-term household debt and wealth inequality, and how do economic shocks or targeted policy interventions change these patterns over time?

3. Literature Review

Existing research shows that income volatility and unequal access to financial resources shape household debt and inequality. Borrowing can help households maintain consumption in the short term, but repeated borrowing increases long-term financial vulnerability (Cardaci)..

One of the most influential studies on this topic is Iacoviello's heterogeneous agent model. His research suggests that households with limited financial resources often borrow to maintain consumption during income shocks. Over time, this can create a cycle in which debt and inequality reinforce one another (Iacoviello, 2008).

Agent-based models support similar conclusions: Russo et al. (2016) found that relaxing credit constraints can increase consumption and support economic activity in the short run because households can continue spending despite income shortfalls. However, over time, savings begin to decline, and debt accumulates among vulnerable households, increasing financial instability (Russo et al., 2016). Likewise, Cardaci and Saraceno simulated an inequality shock and observed a debt-financed consumption increase among lower-income households. Although borrowing helped smooth consumption temporarily, it also increased financial fragility and widened wealth gaps (Cardaci & Saraceno, 2018).

Together, these studies show that debt becomes most harmful when borrowing shifts from a temporary response to a repeated strategy for managing unstable income (Iacoviello, 2008; Russo et al., 2016; Cardaci & Saraceno, 2018).

Thailand's experience clearly reflects many of these theoretical findings, with household debt increasing significantly over the past decade. According to the Bank of Thailand, household debt as a percentage of GDP rose from approximately 60% in 2010 to 90 percent in 2022Q1 (Bank of Thailand, 2022). More recent data from the Bank for International Settlements indicates that household debt remained high at approximately 88% of GDP in 2025 (Bank for International Settlements, n.d.). These figures are among the highest in Southeast Asia and have raised serious concerns about long-term financial stability.

This debt growth has not affected everyone equally; between 2011 and 2021, average debt among low-income households increased by about 36%, compared with 17% for middle-income households and just 1.5% for high-income households (Bank of Thailand, 2022). This massive gap suggests that lower-income households have become increasingly dependent on borrowing just to cover routine consumption rather than building long-term investments.

The COVID-19 pandemic further highlighted these structural challenges. Many Thai households experienced sharp income losses, particularly those working in informal sectors or unstable occupations. Although emergency government cash transfers provided short-term relief, gaps in formal social protection excluded many informal workers from sustained safety nets. As a result, many households relied more heavily on borrowing to maintain basic consumption during prolonged periods of reduced income (International Labour Organization [ILO], 2020).

Labour market inequality is another important factor behind persistent debt vulnerability in Thailand. More than half of Thailand's workforce is employed in the informal sector, where workers generally have less stable incomes and more limited access to social protection than formal workers (National Statistical Office, 2024; OECD, 2025). These conditions make informal-sector households more exposed to income shocks and more likely to borrow during financial hardship. In addition, unequal access to education and economic opportunities limits upward mobility, making it more difficult for disadvantaged households to accumulate wealth over time (Boonlert & Ariyaarpakamol, 2023; Paweenawat & Liao, 2018).

These debt patterns are further reinforced by stark income disparities, highlighted by Thailand's high income Gini coefficient of approximately 0.433 (World Bank, 2023). Because wealth ownership is hyper-concentrated, the vast majority of households lack the asset cushions necessary to absorb unexpected economic shocks (OECD, 2025; World Inequality Lab, n.d.). Without assets to provide long-term financial security, these disadvantaged populations are systematically forced into borrowing cycles just to maintain baseline consumption.

In addition, access to affordable credit remains unequal. Households with limited access to formal financial institutions may rely on short-term or informal loans that often carry higher borrowing costs (Charoenratana et al., 2016; Hemvanich et al., 2025). Higher interest payments can make debt more difficult to repay, increasing the likelihood that financially vulnerable households remain indebted for extended periods. This issue is particularly important among

lower-income households, who often have higher debt burdens and greater difficulty managing repayments because of limited financial resources and restricted access to affordable credit (Chantarat et al., 2017; Hemvanich et al., 2025).

As shown in Figure 1, household debt increased across all income groups between 2011 and 2021, with particularly strong growth among low- and middle-income households. Consumption-related debt also grew substantially during this period, suggesting that many households are increasingly borrowing for everyday expenses rather than productive investment, which may reduce their ability to build wealth over time (World Bank, 2023).

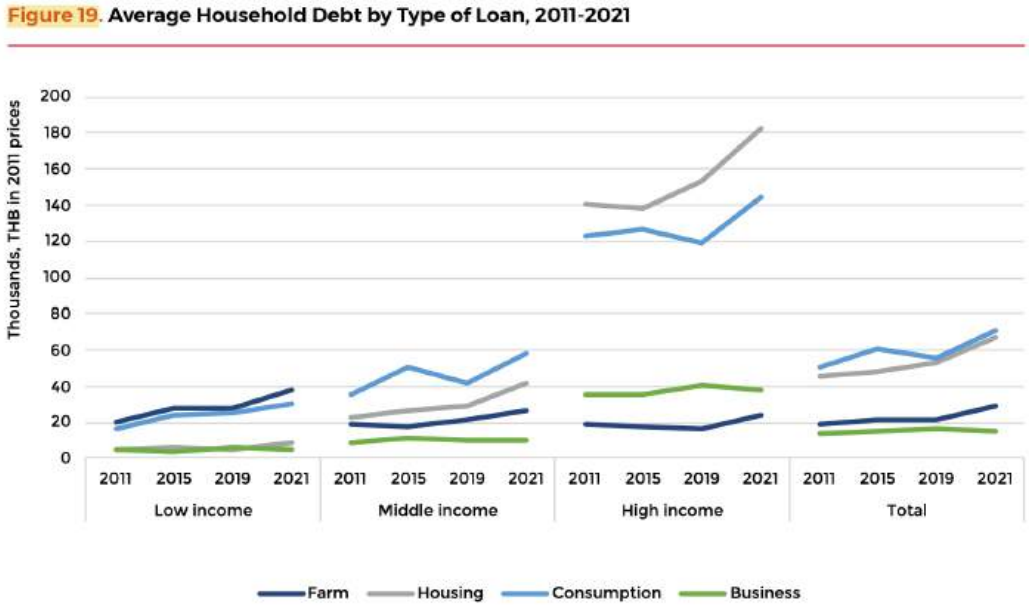


Figure 1: Average Household Debt by Type of Loan and Income Group in Thailand (2011-2021)

Debt among low-income groups grew the fastest (panels B–C), particularly in consumption-related borrowing. In contrast, high-income groups accumulated assets (panel A). Source: World Bank, based on Thailand Socio-Economic Survey (2011-2021).

Overall, previous research shows that income instability, unequal access to financial resources, and labour-market differences contribute to rising household debt. However, many studies rely on national surveys, macroeconomic indicators, or theoretical models that identify broad relationships without fully showing how financial outcomes evolve across households with different employment conditions. This limitation motivates the simulation-based approach used in this study.

In addition, relatively few studies examine how households with different socioeconomic characteristics respond differently to the same economic environment. This is particularly relevant in Thailand, where formal employees, informal workers, welfare recipients, farmers, and small business owners experience substantially different levels of income security and financial risk (Bank of Thailand, 2022; World Bank, 2023).

Simulation models can complement traditional statistical analysis by showing how debt accumulates over time under different economic conditions (Epstein, 2006; Gilbert & Troitzsch, 2005). Rather than replacing empirical evidence, simulation provides a controlled setting for testing assumptions and observing long-term financial outcomes.

Building on these studies, the present research uses Thai socioeconomic data to examine how employment status, income stability, and expenditure volatility shape household debt and wealth accumulation over time. By combining findings from previous literature with a Thailand-specific simulation framework, this study seeks to provide a clearer understanding of the mechanisms contributing to economic inequality and to evaluate potential policy interventions that could reduce long-term financial vulnerability.

4. Methodology

This study utilizes a MATLAB-based agent-based model (ABM) to analyze how income stability influences household debt accumulation and wealth inequality. An ABM approach was selected because it captures how heterogeneous households respond uniquely to income shocks, borrowing constraints, and policy interventions. Unlike aggregate macroeconomic frameworks, static spreadsheet models, or standard regressions, an ABM simulates evolving household financial trajectories over time while keeping underlying parameters controlled. This methodology is uniquely suited for this study, as it isolates the long-term, compounding interactions between repeated income shocks, behavioral borrowing choices, and stochastic expenditure patterns.

The model is calibrated using representative Thai socioeconomic data to investigate how debt inequality develops over time. It includes 500 representative households, with 100 households assigned to each socioeconomic group. This scale balances computational simplicity with the ability to compare behavioral differences across groups. Because the objective is to examine the mechanisms of debt accumulation rather than estimate national population statistics, a small population of representative agents is sufficient to demonstrate the model's dynamics. Similar agent-based models frequently use stylized populations to analyze economic behavior rather than replicate the exact size of a real population (Epstein, 2006). The simulation runs for 100 monthly periods, or approximately 8.3 years, because debt accumulation and changes in wealth occur gradually. A longer timeline allows repeated income and expenditure shocks to compound, making long-term patterns easier to observe while remaining computationally manageable. These specific household groups were selected based on Thai labor market characteristics reported by the National Statistical Office (NSO) and previous studies on household debt and employment.

4.1 Household Groups in the Model

The 500 households were divided into five equal groups of 100 households each. These groups were selected to represent major socioeconomic and employment categories in Thailand based on labour force statistics and household economic data from the Thailand National

Statistical Office (NSO), together with information on informal employment from the International Labour Organization (ILO) and SME classifications from the Office of Small and Medium Enterprises Promotion (OSMEP) (National Statistical Office [NSO], 2025; International Labour Organization [ILO], 2023; Office of Small and Medium Enterprises Promotion [OSMEP], 2024).

1. **Welfare-Dependent/Ultra-Poor:** Households without stable employment that rely primarily on government welfare support, such as state welfare cards.
2. **Formal Sector - Stable:** Households with regular employment, such as office workers, factory workers, or government employees, who receive a stable monthly salary.
3. **Informal Sector - Volatile:** Households working in informal occupations, such as street food vending, motorcycle taxi driving, or delivery work. These workers may earn reasonable income during periods of high demand, but they face precarious employment conditions and limited access to social protection during illness or business slowdowns.
4. **SME - Commercial:** Small business owners or shopkeepers who have a stable income from selling goods and services.
5. **Agricultural - Volatile:** Agricultural households, particularly those in rural regions such as the Northeast and North of Thailand, whose income fluctuates substantially due to weather conditions, droughts, and changing crop prices.

These categories represent major sources of income variation within Thailand's economy. Although simplified, they reflect important differences in employment stability, earnings volatility, and financial vulnerability across Thai households.

4.2 Income and Expenditure in the Simulation

In each simulated month, household agents receive income and incur expenditures based on socioeconomic group, with parameters calibrated from representative Thai socioeconomic statistics to reflect differences in employment stability and financial vulnerability.

- **Welfare group:** Households rely on government assistance rather than a regular salary. Based on the Ministry of Finance (n.d.), qualified recipients receive allowances for essential groceries (300 THB), transportation (750 THB), and a cooking gas discount credit every three months (80 THB) (Money & Banking, 2025).
- **Formal-sector households:** This group receives a fixed monthly income of 28,300 THB, representing the average monthly household income of 28,308 THB and serving as a benchmark for stable employment (National Statistical Office [NSO], 2025).
- **Informal Group:** Monthly income depends on p_unpaid, a random probability variable representing months with zero income due to illness, reduced working hours, or temporary job loss. In positive-earning months, households receive approximately 70% of formal-sector income, or 19,810 THB, as a modelling choice for lower and less stable informal earnings.

- **SME-Commercial Group:** Monthly income is fixed at 40,000 THB to represent relatively successful small-business owners whose earnings exceed Thailand's national average. This is a calibration parameter, informed by NSO household income statistics, rather than an exact national statistic (National Statistical Office).
- **Agricultural-Volatile Group:** Agricultural household income is centered around 27,000 THB, or 60% of the SME benchmark, and varies randomly with a standard deviation of 30,000 THB. This modelling parameter represents high income volatility from weather, seasonal production, and commodity price fluctuations (National Statistical Office, 2025).

4.3 Model Assumptions

To keep the simulation manageable, the model makes several simplifying assumptions. Households are assumed to have continuous access to credit whenever expenditure exceeds income, and all borrowing is modeled as interest-free. This assumption acts as a baseline control, ensuring that debt accumulation is driven by cash deficits rather than interest compounding. The model also applies a standardized repayment rule of 10% of monthly income across all five household groups. In reality, debt is shaped by interest rates, repayment schedules, collateral requirements, credit limits, and differences between formal and informal lenders. These factors were excluded so the simulation could focus specifically on how income stability and expenditure shocks influence debt accumulation.

Monthly household expenditure is based on the National Statistical Office's reported average household expenditure of 22,420 THB. Random expenditure shocks are then introduced to simulate unexpected financial events, such as medical expenses, increases in living costs, or fluctuations in energy prices. Different socioeconomic groups experience different levels of expenditure volatility, reflecting the varying degrees of financial stability observed across Thai households (National Statistical Office [NSO], 2025; Bank of Thailand, 2022).

4.4 Calibration of Parameters

Not all numerical values in this model represent official national statistics. To keep things clear, the model divides its inputs into three separate categories: observed empirical values, calibrated modeling assumptions, and behavioral rules. Observed empirical values, like household expenses and formal-sector income, come directly from data published by Thailand's National Statistical Office. Calibrated modeling assumptions—including SME income levels, agricultural income variance, and informal-sector unpaid-month probabilities—are set using stylized bounds from Thai economic metadata. Behavioral rules include the fixed 10% repayment rate, interest-free borrowing, and the rule that households borrow immediately when expenses exceed income. These distinctions matter because the simulation is designed to show the mechanisms of debt accumulation under realistic conditions rather than make exact economic forecasts. By separating actual data from calibrated assumptions, the model stays transparent

about which inputs are empirical and which ones are just setting up the computational experiment.

4.5 Borrowing and Capital Accumulation Logic

Each month, the simulation compares household income with household expenditure:

- **Surplus Capital:** If income is higher than or equal to expenditure, the household uses 10% of its monthly income to repay outstanding debt. Any remaining income after debt servicing is added directly to long-term cash savings.
- **Deficit and Debt Accumulation:** If expenditure exceeds income, the household faces a financial deficit. To maintain consumption during the shortfall, the household borrows immediately. This deficit amount is added directly to their total outstanding debt stock (Cardaci and Saraceno, 2018).

At the end of each month, net wealth is calculated for every agent by subtracting their outstanding debt liabilities from their cumulative saved earnings.

4.6 Parametric Stress-Test Configurations

The framework evaluates systemic resilience by systematically altering baseline parameter sets across five distinct synthetic shock and policy configurations, each created by altering specific model variables.

- **Configuration 1 (Baseline Control Matrix):** Models standard parametric bounds in Thailand to examine how inequalities may widen over time.
- **Configuration 2 (Labour Shock):** Raises the p_unpaid factor from 0.4 to 0.6 to simulate a major economic contraction or pandemic-related disruption affecting informal workers (Chote et al., 2025).
- **Configuration 3 (SME Income Shock)**: Reduces business income to 30,000 THB to model a global export decline or commercial market contraction.
- **Configuration 4 (Targeted Relief):** Reduces the expenditure variance vector for selected household groups, simulating targeted government price controls on fuel, cooking gas, and basic food items (OECD, 2025).
- **Configuration 5 (Sector-Specific Expenditure Stabilization):** Reduces spending shocks for SME-Commercial and Agricultural-Volatile households, simulating policies that reduce unexpected costs for economically active households.

4.7 Quantifying Debt Inequality (The Gini Index)

To track whether the debt burden becomes more evenly distributed or concentrated, the simulation calculates the Gini coefficient for outstanding debt at each time step. A score of 0 means every household holds identical debt, while a score of 1 means the entire economy's debt is concentrated within a single group.

4.8 Mathematical Logic of the Simulation

The MATLAB script tracks financial changes over the 100-month timeline using three basic accounting equations for each household in every simulated month.

A. The Monthly Cash Deficit

Each month, the code evaluates whether a household's income (I) is sufficient to cover monthly expenditure and expenditure shocks (E). If a household experiences an income shortfall or an unexpected expenditure increase ($I_{i,t} < E_{i,t}$), it faces a cash deficit ($D_{i,t}$), which represents the same amount of money it is short:

$$D_{i,t} = E_{i,t} - I_{i,t} \quad \text{for } I_{i,t} < E_{i,t}$$

In the MATLAB script, this corresponds to the equation line d
$d = d + (expenditure - income)$; inside the loop, it tracks the immediate financing requirement that the household must borrow.

B. The Debt Accumulation and Repayment Loop

Next, the simulation updates the household's total outstanding debt stock ($d_{i,t}$) from one month to the next. If the household faces a cash shortfall, the deficit is added directly to its total debt. If the household has a surplus ($I_{i,t} \geq E_{i,t}$), it uses 10% of monthly income to repay outstanding debt:

$$d_{i,t} = \begin{cases} d_{i,t-1} + D_{i,t} & \text{if } I_{i,t} < E_{i,t} \quad \text{(Borrowing Step)} \\ d_{i,t-1} - \min(0.10 \times I_{i,t}, d_{i,t-1}) & \text{if } I_{i,t} \geq E_{i,t} \quad \text{(Repayment Step)} \end{cases}$$

When household income exceeds expenditure, the model assumes households repay up to 10% of their monthly income toward outstanding debt. This percentage is not intended to represent a national repayment standard but serves as a behavioural assumption that allows debt to decline gradually while households continue to meet current living expenses. Using a fixed repayment proportion also allows repayment behaviour to remain comparable across all socioeconomic groups.The code uses the min function (repay = min(0.1 * income, d);) to ensure that if a household's 10% repayment amount exceeds its remaining debt, the repayment is limited to the outstanding balance.

C. Cumulative Net Wealth Tracking

Finally, at the end of each month, the model calculates the absolute net wealth ($W_{i,t}$) for every individual household. Net wealth is calculated as cumulative income minus cumulative expenditure minus outstanding debt. In this model, outstanding debt is treated as a remaining financial liability at the end of each month. This definition follows the implementation of the MATLAB simulation and allows households' long-term financial positions to be compared across different socioeconomic groups.

$$W_{i,t} = \sum_{\tau=1}^{t} I_{i,t} - \sum_{\tau=1}^{t} E_{i,t} - d_{i,t}$$

The mathematical summation symbol represents a running total. In the code, this corresponds to the variables cum_income(i) and cum_expenditure(i), which accumulate lifetime financial cash flows and track the development of the wealth gap over time.

Overall, the simulation illustrates mechanisms of debt accumulation and wealth inequality rather than precise forecasts of household financial outcomes.

5. Simulated Output Dynamics and Model Behavior

5.1 Compiled Simulation Metrics Matrix

Table 1 summarises the MATLAB simulation outputs at Month 100 (T = 100) for all five configurations. These outputs include each household group's final debt stock and net wealth, as well as the overall debt Gini index, which measures how evenly debt is distributed across the simulated population. The results therefore compare which groups became more indebted and whether each shock or policy reduced debt inequality.

5.2 Baseline Control Matrix Behavior (Configuration 1)

Configuration 1 serves as the baseline scenario, modelling normal economic conditions without additional shocks or policy interventions. It provides a reference point for the later configurations by showing how existing differences in income stability affect debt accumulation over time.

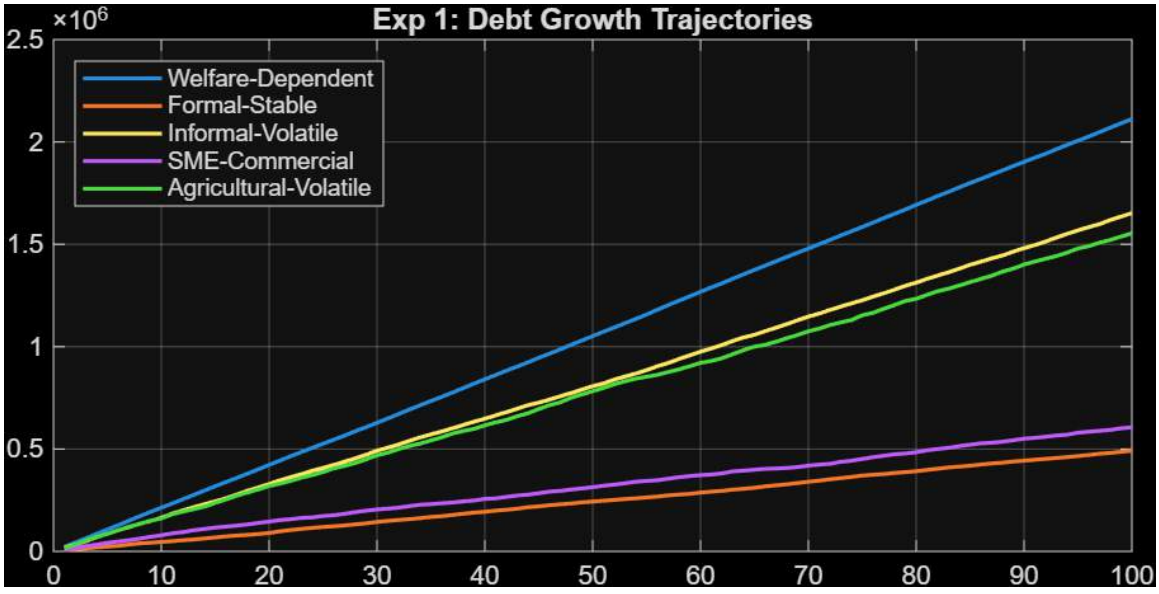


Figure 2. Debt accumulation under baseline economic conditions

Figure 2 shows that debt increased across all household groups throughout the 100-month simulation, although the rate of accumulation varied considerably. The Welfare-Dependent group

accumulated the highest level of debt, ending the simulation with approximately 2.11 million THB in outstanding debt and a net wealth of –4.22 million THB. The Informal-Volatile group also accumulated substantial debt, reaching 1.65 million THB, as periods of interrupted income led to repeated borrowing. By comparison, the Formal-Stable and SME-Commercial groups ended the simulation with much lower debt levels, at approximately 492,000 THB and 605,000 THB, respectively. The SME-Commercial group was the only group that maintained positive net wealth at the end of the baseline simulation.

These results show that income instability can drive debt accumulation and widen wealth inequality even without an external economic crisis. Households with the least stable or insufficient income borrowed repeatedly to maintain consumption, while stable-income households accumulated debt more slowly and preserved higher net wealth. This suggests that policy responses should not only reduce existing debt, but also improve income security through stronger labour protections, more reliable welfare support, and programmes that help vulnerable households maintain income during disruptions.

5.3 Response to Synthetic Income Disruption Vectors (Configuration 2)

Configuration 2 increases the probability that informal workers experience months without income, representing temporary unemployment, illness, or business interruptions that reduce household earnings.

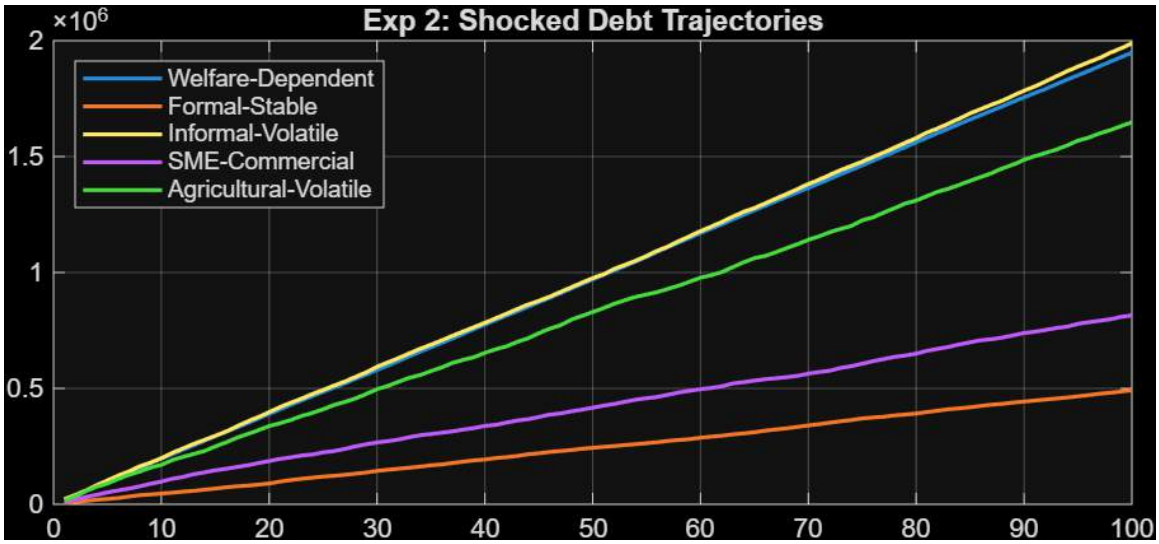


Figure 3 compares debt accumulation after the income shock was introduced.

Compared with the baseline control matrix, the most important intended shift appears in the Informal-Volatile group. Final debt increased from approximately 1.65 million THB to 1.99 million THB, while net wealth declined from –3.14 million THB to –3.87 million THB. The Welfare-Dependent group remained heavily indebted, although its results changed only slightly because its income was already consistently below household expenditure levels. The SME-Commercial and Agricultural-Volatile groups also show some changes in the final table;

because Configuration 2 directly modifies the informal-sector unpaid-month parameter, these additional differences should be interpreted as variation from the model's random income and expenditure draws rather than as a direct policy or shock effect on those groups.

These findings show that income shocks magnify the effect of income instability on debt accumulation. Even temporary income interruptions substantially increased debt among informal workers, demonstrating that unstable employment creates greater financial risk than regular monthly earnings. From a policy perspective, informal workers may need more than emergency loans or temporary cash transfers; stronger interventions could include unemployment protection, sick-leave support, crisis income replacement, and broader access to social insurance.

5.4 Response to Constrained Commercial Asset Floors (Configuration 3)

Configuration 3 examines the effect of reducing SME household income, representing an economic slowdown affecting small businesses. This configuration investigates how lower business earnings influence debt accumulation across different socioeconomic groups.

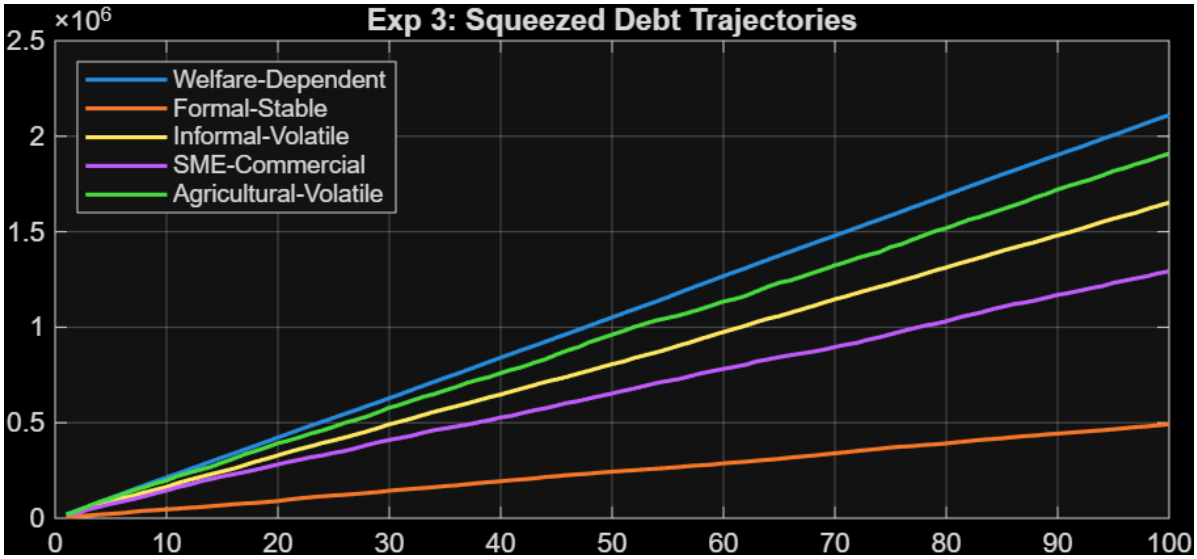


Figure 4: Reducing SME income accelerated debt accumulation most clearly for the SME-Commercial group, while other observed changes should be interpreted cautiously unless they are directly generated by the model's altered parameters.

Figure 4 shows that reducing SME income substantially worsened outcomes for the SME-Commercial group. The SME-Commercial group experienced the largest direct change, with final debt increasing from approximately 606,000 THB in the baseline simulation to 1.29 million THB, while net wealth declined from approximately 383,000 THB to –1.80 million THB. The Agricultural-Volatile group also accumulated more debt in this run, increasing from approximately 1.55 million THB to 1.91 million THB, with a corresponding decline in net wealth. However, because Configuration 3 directly modifies SME income rather than agricultural income, the agricultural result should be interpreted as a sign of the group's

continuing sensitivity to volatility rather than as evidence of a direct SME-to-agriculture spillover.

Debt trajectories for the Welfare-Dependent, Formal-Stable, and Informal-Volatile groups remained largely unchanged because their income conditions were not directly modified in this configuration. Overall, the results suggest that reductions in SME income can substantially worsen financial outcomes for business owners, while agricultural households remain vulnerable to their own income and expenditure volatility.

5.5 Targeted Policy Intervention (Configuration 4)

Configuration 4 investigates whether reducing unexpected expenditure shocks for selected household groups can slow debt accumulation. Rather than increasing household income directly, this configuration represents policies that reduce financial pressure, such as targeted subsidies or cost-of-living support.

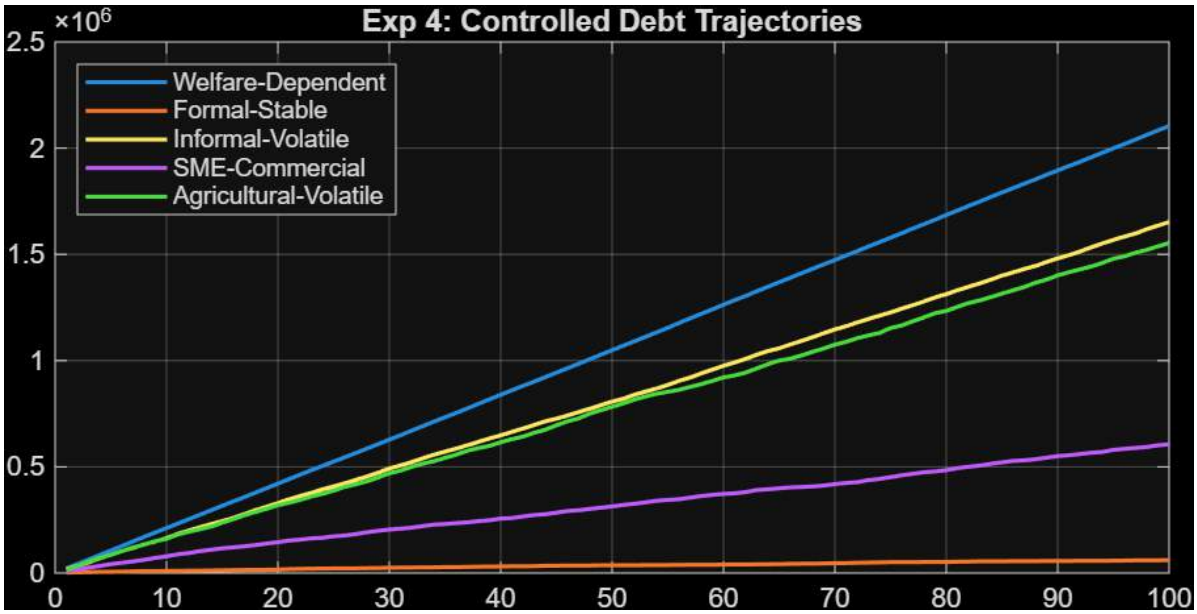


Figure 5 presents the debt trajectories after this policy was introduced.

The policy produced measurable improvements among some household groups. The Formal-Stable group experienced the largest improvement, with final debt falling from approximately 492,000 THB in the baseline simulation to 61,000 THB. Its final net wealth also increased from –350,000 THB to approximately 428,000 THB, making this the only configuration in which the group finished with positive wealth. The Welfare-Dependent group improved only slightly compared with the baseline and continued to record the highest debt level overall. These trajectories demonstrate that dampening expenditure variance vectors improves the net capital position of agents independently of directly increasing income. However, the findings also show that policy effectiveness depends on which households benefit most. Targeted cost-of-living support helped some groups—especially the Formal-Stable group in this simulation—reduce debt and improve net wealth, but it did not fully solve the debt burden of

households whose income remained far below basic expenditure needs. This means that policies aimed at reducing debt inequality should combine expenditure support with income-stabilising measures, especially for welfare-dependent and informal-sector households.

5.6 Broad Expenditure Stabilization Policy (configuration 5)

Configuration 5 investigates the effect of a sector-specific expenditure stabilization policy by reducing expenditure volatility among SME-Commercial and Agricultural-Volatile households. Rather than increasing household income directly, this configuration represents policies that reduce unexpected household expenses and improve financial stability for economically active households.

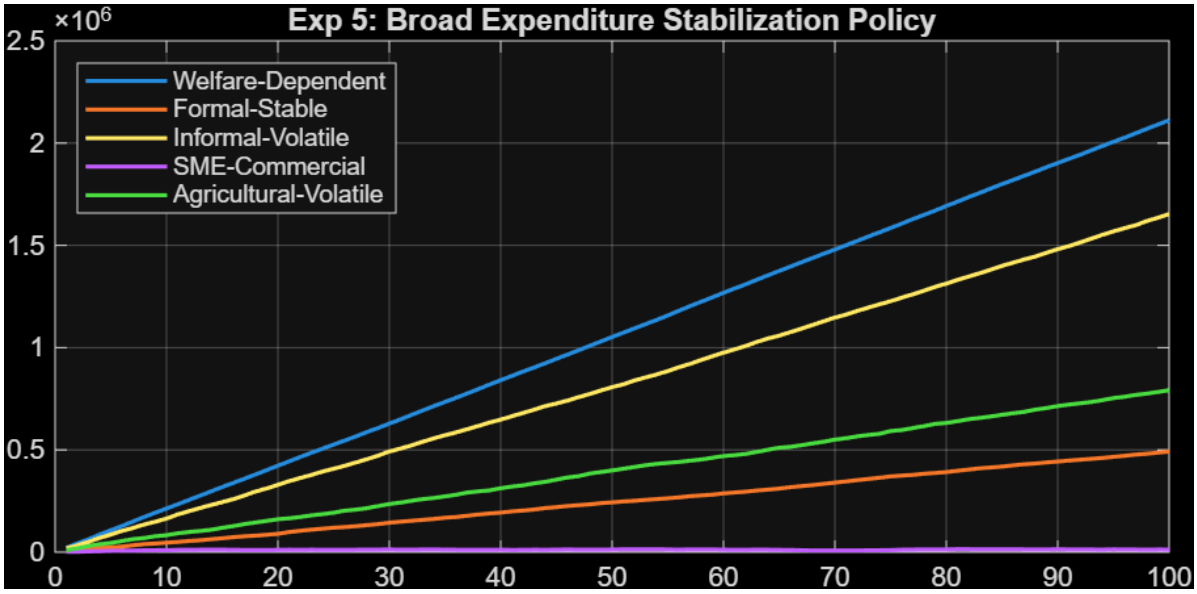


Figure 6: Debt accumulation under a broad expenditure stabilization policy.

Figure 6 shows that the policy substantially improved financial outcomes for the SME-Commercial and Agricultural-Volatile groups. Final debt for SME-Commercial households decreased from approximately 606,000 THB in the baseline simulation to approximately 13,000 THB, while net wealth increased from approximately 383,000 THB to approximately 1.82 million THB. Agricultural-Volatile households also experienced substantial improvements, with final debt decreasing from approximately 1.55 million THB to approximately 792,000 THB, accompanied by a considerable increase in net wealth.

In contrast, debt trajectories for the Welfare-Dependent, Formal-Stable, and Informal-Volatile groups remained broadly similar to those observed in the baseline simulation because the policy did not alter their underlying expenditure conditions.

5.7 Overall Comparison of the Five Simulation Configurations

Socioeconomic	Exp 1:	Exp 2:	Exp 3:	Exp 4:	Exp 5:

Group Matrix	Baseline	Labor Shock	SME Squeeze	Targeted Relief	Sector Stabilization
Welfare-Dependent - Final Debt Stock - Final Net Wealth	2111578.15 -4223156.31	1948078.15 -3896156.31	2111578.15 -4223156.31	2104133.22 -4208266.45	2111578.15 -4223156.31
Formal-Stable - Final Debt Stock - Final Net Wealth	492349.01 -350408.74	492349.01 -350408.74	492349.01 -350408.74	61107.49 428050.90	492349.01 -350408.74
Informal-Volatile - Final Debt Stock - Final Net Wealth	1652638.82 -3139822.42	1988368.72 -3866997.92	1652638.82 -3139822.42	1652638.82 -3139822.42	1652638.82 -3139822.42
SME-Commercial - Final Debt Stock - Final Net Wealth	605981.55 382888.78	815123.99 -326253.66	1293171.80 -1804301.48	605981.55 382888.78	12738.08 1819905.26
Agricultural-Volatile - Final Debt Stock - Final Net Wealth	1553637.02 -2000529.45	1674167.31 -2355005.47	1909077.48 -3029909.68	1553637.02 -2000529.45	792115.79 -411590.21
System Debt Gini Index	0.2911	0.2607	0.2271	0.3637	0.4366

(All units in THB)

Table 1 compares the outcomes across all five simulation configurations and demonstrates the internal consistency of the model when household groups respond to baseline conditions, synthetic income shocks, and parameterized policy interventions. This comparison matters because the research question examines the mathematical mechanisms through which income instability compounds absolute debt and skews the distribution of wealth across agent populations.

Across the simulated configurations, the Welfare-Dependent and Informal-Volatile groups mechanically recorded the highest debt levels and lowest net wealth due to the continuous algorithmic activation of cash deficits. However, the Agricultural-Volatile group also became highly vulnerable in configuration 3, proving that the model's structural rules can produce severe debt outcomes across multiple agent types when income variance is high. In contrast, agents with stable income patterns generated consistent cash surpluses, supporting the theoretical baseline that income stability is a primary structural factor in long-term financial resilience. This pattern directly addresses the central research question by illustrating that income instability drives systemic debt accumulation by leaving vulnerable groups mechanically exposed to recurring monthly deficits.

The five configurations show this mechanism from different angles. In configuration 1, baseline conditions already produced large debt gaps between stable- and unstable-income

households. Configuration 2 intensified this pattern: when informal workers faced more unpaid months, their final debt increased by approximately 336,000 THB compared with the baseline, while their net wealth fell by approximately 727,000 THB. Configuration 3 showed that income shocks can also weaken higher-income groups, as SME net wealth fell from approximately 382,000 THB to –1.80 million THB after business income declined. Configurations 4 and 5 showed that policy interventions can improve some financial outcomes, but their effects were pretty uneven. Targeted relief helped Formal-Stable households most, while the sector-specific expenditure stabilization policy produced the greatest improvement for SME-Commercial households. Therefore, the results suggest that shocks increase vulnerability most when they affect households with unstable income, while policies reduce inequality only when they directly improve the position of the most indebted groups.

One unexpected finding was the change in the system-wide debt Gini index. Although the policy configurations improved financial outcomes for several household groups, the debt Gini increased from 0.2911 in the baseline simulation to 0.3637 in configuration 4 and 0.4366 in configuration 5. This indicates that debt became more unevenly distributed across households, even though some groups accumulated less debt overall.

Another unexpected result occurred in configuration 3, which produced the lowest debt Gini coefficient (0.2271). Rather than indicating an improvement in household welfare, this result happened because debt became more evenly distributed after higher-income households accumulated additional debt. In other words, inequality narrowed because financially stronger households became worse off, not because lower-income households became more financially secure. This finding is important for policy interpretation because it shows that a lower debt Gini index does not always mean that households are better off. Policymakers should therefore evaluate debt inequality together with absolute debt levels, net wealth, and the financial condition of the most vulnerable groups.

6. Discussion

6.1 Interpretation of the Baseline Findings

The baseline simulation suggests that income stability plays a more important role in long-term financial outcomes than income level alone. Although all household groups faced similar expenditure pressures, those with unstable or insufficient income accumulated debt more rapidly than households with regular earnings. This pattern supports previous research showing that households often borrow to maintain consumption during income shocks, even when this increases long-term financial vulnerability (Iacoviello, 2008; Russo et al., 2016). It also reflects Thailand's labour market structure, where many workers are employed in occupations with unstable earnings and limited financial protection.

6.2 The Effect of Income Instability

When the frequency of unpaid months increased, informal workers accumulated debt more rapidly because expenditure continued even during periods without income. This finding is consistent with previous studies showing that income volatility is an important driver of household indebtedness, particularly among informal-sector workers. Similar patterns were observed during the COVID-19 pandemic, when income shocks forced many Thai households to rely more heavily on borrowing to cover daily consumption (Deléchat et al., 2025).

6.3 The Effect of Economic Shocks on Higher-Income Households

The SME income shock produced a significant result. Although SME-Commercial households generally performed well under normal conditions, a reduction in their monthly income quickly eliminated their financial advantage. This suggests that relatively high-income households are not necessarily protected from financial vulnerability if their earnings fall for an extended period.

The results also show that agricultural households stayed financially vulnerable in the same simulation run. However, because the model does not include a direct link between SME businesses and agricultural households, this outcome should be read with caution. The clearest direct finding from Configuration 3 is simply that a steady drop in SME income can quickly hurt the financial position of business-owning households.

6.4 Policy Implications

The policy configurations indicate that reducing household expenditure can improve financial outcomes, even without directly increasing income. Policies such as targeted subsidies, price controls, or cost-of-living support may reduce debt accumulation for several household groups, especially those with relatively stable incomes.

However, these policies were less effective for Welfare-Dependent households. Because their baseline income remained far below average household expenditure, reducing occasional spending shocks was not enough to prevent long-term debt accumulation. This suggests that expense-reduction policies may have limited effects unless they are combined with measures that address inadequate income.

6.5 Understanding the Debt Gini Coefficient

One of the most unexpected findings was the behaviour of the debt Gini coefficient. Although the policy interventions improved financial outcomes for several household groups, debt inequality increased rather than decreased. This result appears contradictory because lower debt levels are generally expected to be associated with lower inequality.

However, the simulation suggests that the increase in the debt Gini coefficient occurred because households with stable employment and business income experienced significant improvements, while the Welfare-Dependent group remained heavily indebted throughout the simulation. As a result, the financial gap between these groups widened, even though overall debt levels declined for many households. This finding demonstrates that improvements in

average financial wellbeing do not necessarily translate into lower inequality if the most vulnerable households receive relatively little benefit.

Similarly, the lowest debt Gini coefficient occurred during the SME income shock rather than under the policy interventions. This did not represent an improvement in household welfare. Instead, inequality declined because previously stronger households accumulated more debt, making debt levels more similar across the population. This highlights the importance of interpreting inequality measures alongside changes in overall household wellbeing rather than considering the Gini coefficient alone.

6.6 Comparison with Stylized Economic Indicators

Overall, the simulation generated trends that are qualitatively consistent with Thailand's broader macroeconomic baseline. National indicators show that real-world household debt remained persistently high at 95.5% of GDP in 2021Q1, while lower-income households experienced faster debt accumulation than higher-income groups over the last decade. The simulation reflects this pattern by showing that agents with more volatile income accumulate debt faster than stable-salaried agents. Although the model does not forecast exact real-world trajectories, it captures the core structural mechanism: income volatility prevents financially vulnerable agents from smoothing consumption during localized economic shocks.

7. Model Evaluation

7.1 Model Validation

The model worked as expected and showed logical results when different economic changes were introduced. Tweaking the income shocks in configuration 2 clearly increased debt for informal workers, while cutting business revenue in configuration 3 hurt the savings of SME owners. On top of that, the policy ideas tested in configurations 4 and 5 successfully slowed down debt growth for the specific groups they were meant to help. These patterns match classic economic studies using similar models, proving that this setup gives a clear and consistent look at how income changes drive debt inequality. Since this simulation purposefully leaves out complex factors like changing interest rates or credit limits to keep things straightforward, the results should be viewed as a basic look at how these economic rules work rather than an exact financial forecast.

7.2 Limitations

This study has several limitations that should be considered when interpreting the findings. First, the simulation simplifies household borrowing behaviour. Households are assumed to have continuous access to credit whenever expenditure exceeds income, and all borrowing is treated as interest-free. In reality, Thai households may face borrowing limits, interest payments, differences in loan conditions, collateral requirements, and unequal access to formal credit, causing some households to rely on higher-cost informal loans. Consequently, the

model establishes a conservative lower-bound framework for structural debt growth. By demonstrating that severe debt inequality manifests even under simplified, non-compounding financial friction, the model proves that structural inequality is an intrinsic mechanical consequence of income variance. If interest rates and credit constraints were included, debt could accumulate more rapidly for unstable-income groups, while some households might be unable to borrow enough to maintain consumption at all. Therefore, the simulation results should be interpreted as a conservative illustration of the mechanisms linking income instability, debt accumulation, and wealth inequality, rather than a full representation of Thailand's financial system.

Second, several parameters are modeling assumptions rather than exact empirical estimates. Although the model uses Thai socioeconomic data from sources like the National Statistical Office, the Bank of Thailand, and the International Labour Organization, variables like spending volatility and farm income swings are just meant to show the relative differences between groups. These assumptions help explain how the economy works but shouldn't be taken as exact measurements of real household outcomes.

The model also assumes that households pay back up to 10% of their monthly income whenever they make more than they spend. This fixed rule keeps things consistent when comparing the five simulation setups, but real-life repayment behavior varies depending on income, loan rules, and financial priorities. If the model used a higher repayment rate, debt would probably drop faster, while a lower rate would lead to way more debt over time. Future research could do a sensitivity analysis to see exactly how different repayment rates change the simulation results.

Fourthly, the model uses a simplified definition of household wealth. Net wealth is calculated as cumulative income minus cumulative expenditure and outstanding debt. This approach makes comparisons across groups possible, but it may overestimate financial vulnerability if some debt is already reflected in accumulated expenditure. Therefore, the wealth estimates should be interpreted as relative indicators of financial condition rather than precise measures of household net worth.

Finally, the simulation includes 500 representative households observed over one hundred monthly periods. This design makes the mechanisms of debt accumulation clear, but it cannot capture the full diversity of Thailand’s population or all possible external economic shocks. Future research could improve realism by incorporating larger populations, interest rates, inflation, demographic differences, credit constraints, and more detailed household financial behaviour.

8. Conclusion

Household debt and economic inequality are closely connected in Thailand, particularly for households with unstable or insufficient income. This study examined how employment status, income stability, and expenditure volatility influence debt accumulation by developing an

agent-based simulation model calibrated using Thai socioeconomic data. The simulation made it possible to observe how debt patterns develop under different economic conditions and policy scenarios.

The findings suggest that income stability is a major factor in long-term financial resilience. In the baseline simulation, households with unstable earnings accumulated substantially more debt than households with stable employment. The labour shock increased debt among informal workers, while the SME income shock showed that even higher-income households can become vulnerable when earnings decline for an extended period. The targeted relief and universal safety net configurations improved outcomes for some groups, but they did not always reduce overall debt inequality. In some cases, financially stronger households benefited more from policy interventions, while the most vulnerable households remained heavily indebted.

Overall, the simulation supports previous research showing that debt accumulation is shaped not only by income level, but also by a household's ability to withstand financial shocks. Although the model simplifies real household borrowing and should not be interpreted as a precise forecast, it provides a useful framework for understanding how debt and inequality may reinforce one another over time.

The findings suggest that policies aimed only at reducing short-term financial pressure may be insufficient to reduce long-term inequality. Addressing structural differences in income stability and improving the financial resilience of vulnerable households are also important for promoting sustainable household finances. Future research could extend the model by incorporating interest rates, credit constraints, inflation, and household behavioural responses to better represent the complexity of Thailand’s economy.

Works Cited

Bank for International Settlements. Total Credit Statistics: Household Debt to GDP (Thailand). BIS, https://data.bis.org/topics/TOTAL_CREDIT. Accessed 2 Aug. 2026.

Bank of Thailand. Financial Stability Report 2022. 2022, https://www.bot.or.th/content/dam/bot/documents/en/research-and-publications/reports/financial-stability-report/FSR2022e.pdf.

Bank of Thailand. Directional Paper on Sustainable Solutions to Thailand's Structural Debt Overhang Problems. 2023, https://www.bot.or.th/content/dam/bot/financial-innovation/sustainable-finance/household-and-sme/2023_DP_Household%20Debt_ENG_final.pdf.

Boonlert, R., and N. Ariyaarpakamol. "Inequality of Opportunity in Wealth: Unpacking the Wealth Myth of Thailand." Journal of Population and Social Studies, vol. 31, 2023, pp. 708–23. https://doi.org/10.25133/JPSSv312023.039.

Cardaci, A. "Household Debt and Income Inequality: An Agent-Based Perspective." Journal of Economic Behavior & Organization, vol. 149, 2018, pp. 434–58. https://doi.org/10.1016/j.jebo.2018.01.010.

Chantarat, S., et al. Thailand's Household Debt through the Lens of Credit Bureau Data: Debt and Delinquency. Puey Ungphakorn Institute for Economic Research, 2017. PIER Discussion Paper, no. 061, https://www.pier.or.th/files/dp/pier_dp_061.pdf.

Charoenratana, S., et al. "Inequality and Vulnerability among Informal Debt in Thai Society." Journal of MCU Peace Studies, vol. 4, no. 1, 2016, pp. 188–208. https://so03.tci-thaijo.org/index.php/journal-peace/article/view/76599.

Deléchat, C., et al. Inequality, Household Debt, and the Role of Social Protection in Thailand. International Monetary Fund, 2025. IMF Working Paper, no. 2025/248, https://doi.org/10.5089/9798229021524.001.

Hemvanich, S., et al. "Credit Segmentation and Household Vulnerability in Thailand: Formal versus Informal Debt Risks." Journal of Risk and Financial Management, vol. 18, no. 11, 2025, article 632. https://doi.org/10.3390/jrfm18110632.

Iacoviello, M. "Household Debt and Income Inequality, 1963–2003." Journal of Money, Credit and Banking, vol. 40, no. 5, 2008, pp. 929–65. https://doi.org/10.1111/j.1538-4616.2008.00142.x.

International Labour Organization. Thailand: Informal Employment Survey 2023. 2023, https://webapps.ilo.org/surveyLib/index.php/catalog/8820.

Kumhof, M., et al. "Inequality, Leverage, and Crises." American Economic Review, vol. 105, no. 3, 2015, pp. 1217–45. https://doi.org/10.1257/aer.20110683.

Ministry of Finance. State Welfare Registration Project. https://welfare.mof.go.th/. Accessed 2 Aug. 2026.

Money & Banking. "อัปเดต สวัสดิการแห่งรัฐ ม.ค. 69 ปีใหม่นี้ ว่าจะอะไรบ้าง [Update on State Welfare Benefits for January 2026: What Is Available This New Year?]." LINE TODAY, 26 Dec. 2025, https://today.line.me/th/v3/article/pe19Xk7.

National Statistical Office. The Informal Employment Survey 2024. 2024, https://www.nso.go.th/nsoweb/storage/survey_detail/2024/20241125143625_37256.pdf.
National Statistical Office. Statistical Yearbook Thailand 2025. 2025, https://www.nso.go.th/public/e-book/Statistical-Yearbook/SYB-2025/.
OECD. OECD Capital Market Review of Thailand 2025. OECD Publishing, 2025, https://doi.org/10.1787/0a975590-en.
OECD. OECD Economic Surveys: Thailand 2025. OECD Publishing, 2025, https://doi.org/10.1787/426b9bc0-en.
Office of Small and Medium Enterprises Promotion. Definition of MSMEs. 2024, https://en.sme.go.th/en/page.php?modulekey=363.
Paweenawat, S. W., and L. Liao. "Educational Assortative Mating and Income Inequality in Thailand." Journal of Demographic Economics, vol. 89, no. 2, 2023, pp. 283–98. https://doi.org/10.1017/dem.2022.9.
Russo, A., et al. "Increasing Inequality, Consumer Credit, and Financial Fragility in an Agent-Based Macroeconomic Model." Journal of Evolutionary Economics, vol. 26, no. 1, 2016, pp. 25–47. https://doi.org/10.1007/s00191-015-0410-z.
World Bank. Bridging the Gap: Inequality and Jobs in Thailand. 2023, https://documents1.worldbank.org/curated/en/099112823133018003/pdf/P17759905901d70ed0a968052455d5252f0.pdf.
World Bank Group. Thailand Rural Income Diagnostic: Challenges and Opportunities for Rural Farmers. 2022, https://www.worldbank.org/en/country/thailand/publication/thailand-rural-income-diagnostic-challenges-and-opportunities-for-rural-farmers.
World Inequality Lab. World Inequality Database: Thailand. https://wid.world/country/thailand/. Accessed 2 Aug. 2026.
Xu, Y. Household Deleveraging: International Practices—Thailand. International Monetary Fund, 2025. IMF eLibrary, https://doi.org/10.5089/9798229010801.018.A001

The Global Market for the "Caviar of the East"Ethics, Regulation, and Sustainability
By Ilinca Harja-Iliescu

Abstract

This study examines the global market for the Edible Bird Nest (EBN), commonly known as the "Caviar of the East," focusing on the intersection of trade regulations, bioethics, public health, and environmental sustainability. Produced by specific species of swiftlets using a glutinous, high-protein salivary secretion, EBN represents a multi-billion dollar luxury commodity driven primarily by demand within East Asian markets. The rapid expansion of "swiftlet farming"—the practice of luring wild birds into specialized urban structures—has generated significant economic opportunities, particularly within Southeast Asia's "golden triangle" of production (Indonesia, Malaysia, and Thailand). However, this expansion has simultaneously introduced complex ethical and regulatory challenges.

Through a systematic literature review spanning publications from 1970 to 2026, this paper evaluates the compositional profile of EBN, current harvesting and cleaning methodologies, regulatory frameworks across major producing and consuming nations, and the socio-ecological impacts of urban swiftlet farming. The analysis highlights severe negative externalities associated with inner-city farming, including sensory pollution, structural degradation, biosecurity threats (e.g., Avian Influenza, Cryptococcosis, and vector-borne diseases like Dengue fever), and acute food safety concerns such as excessive nitrate/nitrite contamination and product adulteration. From an animal welfare perspective, premature nest harvesting and acoustic manipulation exploit the biological instincts of non-domesticated swiftlets, causing metabolic exhaustion and reproductive failure.

To reconcile economic growth with ethical and public health considerations, this study examines sustainable farming models, including low-cost purpose-built master houses located outside urban centers. Finally, the paper proposes a modernized regulatory and technological roadmap—incorporating AI computer vision for post-fledging harvest verification, blockchain-backed Digital Product Passports (DPPs) for supply chain transparency, closed-loop processing, and dedicated agro-forestry zoning laws. This research concludes that the long-term viability of the EBN industry depends on transitioning from an unregulated, extractive model to a transparent, technology-enabled stewardship framework that aligns producer profitability with avian welfare and public safety.

1. Introduction

One of the most highly valued food products, commonly referred to as the „Caviar of the East", fetching a premium price, is the edible bird nest ("**EBN**") created by certain species of swiftlets. Swiftles are birds from the four genera *Aerodramus*, *Collocalia*, *Hydrochous* and *Schoutedenapus*, which form the tribe **Collocaliini** within the swift family **Apodidae** **(**Brooke R. K., 1970). They are in many respects typical members of the Apodidae, having narrow wings for fast flight, with a wide gape and small reduced beak surrounded by bristles for catching

insects in flight. What distinguishes many but not all species from other swifts and indeed almost all other birds is their ability to use a simple but effective form of echolocation to navigate in total darkness through the chasms and shafts of the caves where they roost at night and breed.

Swiftlets build their home by generating a high-protein glutinous secretion produced by their salivary glands. This secretion hardens on exposure after the bird has shaped it into a cup-shaped nest which can usually be found glued to the ceiling or corner of a cave or building. There are more than 24 species of swiftlets, but only a few produce nests that are deemed edible. Two heavily exploited species, the **White-nest swiftlet** (Aerodramus fuciphagus) and the **Black-nest swiftlet** (Aerodramus maximus) compose the majority of the EBN trade worldwide (Babji, A.S., Nurfatin, M.H., Etty Syarmila, I.K. & Masitah, M., 2015, "Agriculture Science").

Their natural habitats range from the **Nicobar Islands** in the Indian Ocean to twisted sea caves in the coastal regions of **Thailand, Vietnam, Indonesia, Borneo** and the **Palawan Islands** in the Philippines, forming a "golden triangle" of edible bird nest production. Traditionally, swiftlet nests are double boiled together with rock sugar to prepare a delicacy, a soup known as "yan". Its Chinese name translates literally as „swallow's nest". This dish has a rich history; the nests were first traded in China during the Tang Dynasty (618-907 A.D.). There are historical sources stating that between the years 1368 and 1644, Admiral Cheng He introduced the bird nest soup to the imperial court of the Ming Dynasty and the nests quickly became a most prized possession, traded throughout China for stoneware and porcelain, iron, brass, gold, glass beads and textiles (Lee, T. H., Wani, W. A., Lee, C. H., and Cheng, K. K., 2021, "Critical Reviews in Food Science and Nutrition"). Swiftlet nest soup is not considered just a luxury food item to be consumed; it is also traditionally believed to provide health benefits, such as aiding digestion, raising libido, improving the voice, alleviating asthma and improving concentration.

The production of swiftlet nests is skyrocketing; Malaysia is situated right at the heart of the "golden triangle" of swiftlet nest production, making it a strong producer in this industry. According to Kuan and Lee (2005), Malaysia is the world's third largest supplier of swiftlet nests after Thailand and Indonesia, which contributes to only 10% of the 2,800 tons produced annually and consumed by top buyers in China and Taiwan. The edible bird nest global sector is now worth up to 5 billion US dollars. (Koperasi Penternak Burung Walit Malaysia Berhad) and according to the Malaysian Federation of Bird's Nest Merchants Association, Malaysia's annual production of bird nests has reached 1 billion ringgit (290 million U.S. dollars) in value. The primary demanding entity is constituted by the international Chinese community, and while Taiwan, Singapore and North America are the main consumers, there are many new emerging markets such as the Middle East, Japan and Korea.

The swiftlet farming industry has significantly transformed the physical form of cities and even the nesting behaviours of swiftlets, which poses significant implications for the use of urban space and the sustainability of urban swiftlet farming as a whole. "Swiftlet farming" is a colloquial term that refers to the cultivation system of edible birds' nests by preparing specially designed buildings for swiftlets to roost and nest in. As Lim and Cranbrook (Creighton Connolly, 2016, "'A place for everything': Moral landscapes of 'swiftlet farming' in George Town,

Malaysia") put it, "scientists and laymen alike are deeply divided into two schools of thought. One promotes the advantages of swiftlet farming while the other strongly opposes it". Although cities have long been hybrid spaces of human, animal and insect inhabitation, attempts are frequently made to expel the "other" and particularly the "nonhuman" from "civilised" urban spaces (Bingham and Hinchliffe, 2008, Gaynor, 1999, Philo, 1995).

The recent increasing presence of domesticated swiftlets in Malaysian cities has intruded upon ideas about the "proper" place of these birds in urban areas. In contrast to other animals like dogs, cats, chickens and other livestock animals which have been domesticated for centuries, the domestication of swiftlets is currently ongoing and open to study. Swiftlet farming is therefore an important case to analyse because of the semi-domesticated nature of the birds and their liminal status in cities, disrupting the categories of "wild" and "domesticated". More broadly, swiftlet farming can be seen as a form of urban livestock keeping, which, one one hand "has received increasing support as a strategy for food security and urban sustainability" (Kathryn J. A. Colasanti, Michael W. Hamm,Charlotte M. Litjens, 2012, "The City as an 'Agricultural Powerhouse'? Perspectives on Expanding Urban Agriculture from Detroit, Michigan"), but discouragement on the other because of the mass-population's rooted preconceptions towards wildlife. Consequently, the Malaysian government has been faced with the contradictory task of trying to mitigate the industry's impacts on affected city dwellers, whilst simultaneously promoting the industry and protecting the interests of swiftlet farmers, all while a proper study on the ethics of swiftlet farming is yet to be made.

The scope of this paper is to determine whether the EBN industry and the current methodologies of swiftlet farming are ethical towards both the consumers and the birds themselves, and whether this industry presents itself with the quality of sustainability in the future. Through the use of empirical evidence, this paper will explore the methodologies used in farming swiftlet nests, the regulations implemented plus their actual impact and the possibility of disease outbreaks and the imposed risks and dangers, alongside the extent of the cultural bond between this product and its consumers and the positive and negative ways in which the industry has been perceived so far by the mass audience. Furthermore, the discussion will evaluate both the benefits and drawbacks of the current swiftlet farming activities and their foreseeable expansion into a broader market, reaching other countries and converting them into additional EBN producer states. By synthesizing research and theoretical perspectives, this paper aims to provide a balanced and evidence-based understanding of the impact the EBN industry has on both the consumers and the swiftlets, analysing the ethical and sustainable aspects of the current farming environments and procedures.

1.1 Methodology

This chapter aims to summarize a comprehensive body of resources used to support the methodological framework necessary to identify, select and analyse the scientific and economical literature that depicts the risks and ethics regarding the swiftlet farming industry. The purpose of this review is to analyse the ethical values of the EBN production along with its risk of disease

outbreaks, by taking into account both the economical interests of the involved nations and the salient position in the animal product market which this industry holds. The findings support the primary goal of the thesis: the EBN industry and the current methodologies of swiftlet farming are not ethical towards both the consumers and the birds themselves and present great possibility of disease outbreaks. To identify the most relevant and up-to-date evidence, searches were conducted using databases such as Google Scholar, Research Gate, Food Research, W117: Journal for the Advancement of Performance Information and Value, Semantic Scholar, Sci-Hub and the National Library of Medicine. The research period began February 2026 and extended to March 2026. The keywords used during the process, as well as their various combinations, included "EBN", "swiftlet farming", "bird nest soup", "golden triangle of swiftlet nest production" and "EBN producer states". Boolean operators (AND/OR) were used to narrow or expand search results as the need for more thorough search parameters arose.
In the context of this thesis, the studies spanned between 1970 to 2026 to ensure a comprehensive analysis of the topic and provide credible political and economical evidence on the importance of chokepoints on trade routes. Moreover, only articles published in English were considered as it offered a broader library of resources to formulate compelling findings.
The initial search yielded a library of approximately 30 articles. After a thorough review of their titles and abstracts, only 17 articles were selected for in-depth reviews, of which 11 met the date, methodology, and accuracy criteria established. The following variables were decisive in the process of data extraction from the final pool of studies: type of study, publication year, sample size, focus on ethics regarding swiftlet farming and EBN production and volume of studies made on the risk of disease outbreak. The scientific and economical literature reviewed revealed trends consistent with the findings of this thesis. The industry of swiftlet nest production and the current methodologies of swiftlet farming utilised are not ethical towards neither the consumers nor the birds themselves and present great possibility of disease outbreaks, with real-world examples to sustain this claim

2. Chapter 1

2.1 Methodologies and Procedures used in farming swiftlet nests

In the last decade, the traditional housing of the swiftlets and the farming and cleaning process of their nests was regarded as a trade secret which the owner of the business would not reveal to any outsider other than their immediate family members. However, the raised awareness of the consumers towards quality control and their demand for increased EBN quality has forced the traditional business owners to abandon their well-kept secret in search of a better approach. (CIB Working Commission W117, 2018). Bird nests were formerly harvested from caves, mainly from the enormous limestone caves at Gomantong and Niah in Borneo. With the escalation in demand, these sources have been overtaken since the late 1990s by purpose-built nesting houses. At first, disused shop houses and cinemas were converted into artificial cave habitats, but knowledge of what the birds' primary and basic needs are has improved the

facilities to the point that the birds are now enticed to nest in specially designed reinforced concrete housing provided with the appropriate light, humidity and other features. These nesting 'blocks' are normally found in urban areas adjoining the sea, since the birds have a propensity to flock in such places. Many people have invested heavily in hopes of attracting the swiftlets to nest in these newly emerged facilities. Before 1998, there were about 900 swiftlet farms throughout Malaysia. Five years later, the number of swiftlet farms throughout the country was close to 36,000 units, with an average annualized growth rate of 35% (Hameed, 2007). Collecting and harvesting the swiftlet nests should follow a strict set of rules: only complete nests should be considered for harvesting, and should only be harvested with a polished stainless-steel trowel. The harvesters should select only nests with no eggs or chicks. If the nest is harvested or destroyed before laying, the birds will build a new one to replace it, which is called a "re-nest", an essential cycle to maintain in swiftlet farming (Jandam, 2017). If the nest is harvested or destroyed after eggs are laid, the birds will not build a substitute, rendering the associated pair as "unproductive", which is to be avoided. However, proof of disregard towards these strict ethical guidelines can be found and extreme over-harvesting often occurs.

The traditional cleaning process has been implemented for more than half a century. It has been passed on from generation to generation without much improvement. This process is time consuming and without scientific proof or analysis. The raw bird nests harvested from the swiftlet farms or limestone caves go through the process of sorting according to its grade. This is to facilitate the value-added diversity of EBN products. The fewer feathers or impurities, the higher the grade of the bird nest. Premium grade bird nests are followed by normal grade and then by low-grade products, named "biscuits" or "bird nest strips". After the sorting process, raw bird nests will be soaked in clean water for further cleaning. During the process, most of the dust or surface impurities will float up; the bird nest starts to absorb the water and expands in size by 20% to 30%, with an increase of up to five times its original weight. The condition of the nest turns soft and sticky, therefore, it is easier for the workers to pick out the feathers with forceps or tweezers. This process may take 20 to 30 minutes for a skilled worker to do and it depends on the cleanliness of the raw bird nest. After cleaning, the process of molding the nest takes place. The molding process requires a set of plaster mold or stainless-steel netting, like half cups. By placing moist bird nests into the plastic mold, the cleaned bird nest is pruned into its original shape upon drying up. The plastic mold also requires some clips or tools to hold the bird nest in place during drying. Most of the bird nests will be dried under a low heat cabinet with assistance of a direct fan blowing.

2.2 Grading of swiftlet nests

The traditional grading system of swiftlet nests is based on specific physical determinants such as **colour, cleanliness, shape, and size**. However, the grading system is subjective and hard to implement due to the different seasonal harvesting periods; various grading and trading mainly involve lump-sum traders. As the problem in grading the high-priced commodity arises, Malaysia has established MS 2334:2011 and SIRIM/DVS 3:2015 to standardize the grading of

swiftlet nests based on the **length and contact planes of the nests** (Badan Karantina Pertanian Republik Indonesia, No 374/Kpts/KH.210/L/5/2010, Petunjuk Teknis Penanganan Dan Pemeriksaan Sarang Burung Walet Dan Sriti, 2010). These standards are developed to ensure the quality and the food safety of EBN for human consumption. Moreover, the grading system establishes a uniform method for evaluating quality and price; the higher the grade is, the more expensive the farmed nest is. Unfortunately, the adulteration of EBN products by unethical suppliers for greater financial gain results in only 44.19% of producers using the MS grading, 22.66% of traders trading in bulk buying, and 33.14% using the traditional grading system (Malaysia's edible bird nest (EBN) industry | FFTC agricultural policy platform (FFTC-AP)). Over the years, more sophisticated techniques to determine grades, authenticity, and identification of the nests have been developed with the advancement and availability of technology. This includes microscopic inspection, automation (radio frequency identification), and analytical methods (genetic and observational analysis).

2.2.1 Colour

Food with visually appealing **colours** is important in offering high-quality products with pleasant appearances to attract customers. A colour measurement system (colourimetry) has been used in the food industry to measure a broad range of food products. This includes the swiftlet nests since it represents a consumer's first quality assessment and influences their preference and choices. In general, there are three distinct colour variations of swiftlet nest: **white EBN** derived from the white-nest swiftlet (A. fuciphagus), **black EBN** sourced from the black-nest swiftlet (A. maximus), and **red EBN**, which is essentially white EBN that has undergone a colour transformation ('The edible nest swiftlet industry in Southeast Asia: Capitalism meets Commensalism, Human Ecology, 2015). Red EBN has the highest market price (five times the cost of white EBN) as it is relatively rare. The colour of the nests is highly variable based on the origin (cave or house) due to different nitrite and nitrate concentrations present in the environment and types of minerals contained in the swiftlet's diet. EBN colour can be detected using the ***International Commission on Illumination (CIE) L*a*b* system***, where L* is for lightness, a* for redness or greenness, and b* for yellowness or blueness (International Journal of Food Properties, 2018).

As for **a* (redness or greenness) and b* (yellowness or blueness)**, swiftlet nests farmed from caves tend to have dull, brownish, or orange colour. According to the previous works of literature, there are many factors affecting the formation of red EBN, such as the presence of blood during bird nest construction, the swiftlets' diet, iron oxidation caused by limestone dripping, the formation of aryl-C-N and NO2 side group on aromatic amino acid and presence of albumin ('Edible Bird's Nest: Physicochemical Properties, Production, and Application of Bioactive Extracts and Glycopeptides', 2021). The red EBN results from the nitrifying bacteria that are present in the nest's structure reacting with ammonia vapours from the bird soil containing guano dropping ('Edible Bird's Nest: Physicochemical Properties, Production, and Application of Bioactive Extracts and Glycopeptides', 2021). However, red EBN has a lower

nitrite level than black EBN and lower protein extraction than white EBN. Besides, red EBN has a higher total lipid content, nine times greater than white EBN (‘Journal of Agricultural and Food Chemistry’, 2018). As the red EBN has a higher market price than white EBN, some swiftlet nest traders began treating white EBN with ammonia, sodium nitrate, or redding agents to produce more expensive red EBN (‘Journal of Ethnopharmacology’, 2013).

2.2.2 Shape and Size

Even though all swiftlet nests provide a potent nutritional value, each type of nest comes in different shapes and sizes. These two factors are specifically emphasised when grading in the new system. Related to both attributes, there are four different types of shapes and sizes of swiftlet nests: 1. cup, whole nest from large to extra-large size, 2. triangle, medium size, 3. strips, broken pieces from a former whole nest, and 4. cakes, small sized, shaped as a biscuit. Based on MS 2334:2011 and SIRIM/DVS 3:2015, EBN can be categorized into Grades A, B, C, and others (Journal of Food Measurement and Characterization, 2020). Others include EBN in filament, fragment, fine grain, and horn types, with each grade of EBN having a different market value. Grade A offers astounding quality, followed by grades B and C. Thus, higher-grade EBN products have better quality attributes and greater profit. Therefore, compliance with the national bird nest industry standards demonstrates the best practices to create the ultimate quality EBN in tandem with market demand. Based on the reference angle for EBN grading, the most sought-after shape is the perfect half cup that is large and lacks holes. However, the occurrence of a half cup (Grade A, α 180) is considerably rare as swiftlets prefer to construct their nests in the corners of swiftlet farming houses due to less material required (Journal of Life Science and Biomedicine, 2014). Similarly, the shape of grade B EBN is the same as grade A EBN, but the nest inclines 135° when placed on a flat surface, and grade C EBN has a 90° inclination when put on a stable, level surface. Cleanliness is also considered one of the main factors determining the value of a swiftlet nest. In the market, swiftlet nests with lesser content of impurities and feathers can be considered as a high-value, premium grade. Conversely, EBN that contains more impurities is classified as low quality. A lower grade EBN requires a prolonged, diligent cleaning process that affects its overall quality.

2.3 Adulteration of Swiftlet Nests

Swiftlet nests have become susceptible to adulteration over the years due to their high market price and limited supply, thereby threatening consumers’ interests and market competition. They are often adulterated with inexpensive ingredients to boost their weight and market value or deliberately mislabelled with false information regarding their origin and nutritional contents. These unscrupulous acts of fraudulent suppliers contribute to the deterioration of EBN quality and undermine genuine business as the boundaries of trade expanded globally. For instance, this dishonest practice is reflected by the addition of natural red colourants such as ***karaya gum, red seaweed, or Tremella fungus*** to White EBN to mimic the premium-grade Red EBN, which commands a higher market price. Therefore, prevalent issues

have become a growing concern as consumption of unnoticed adulterated EBN may pose serious health risks and allergies to consumers.

Adulterants utilized in swiftlet nests can be categorized into two groups (Journal of Food Science & Technology, 2016). Type I adulterants consist of solid materials externally added to EBN, such as karaya gum, Tremella fungus, ***coralline seaweed, agar strips, fish bladder,*** and ***pork skin***. In contrast, Type II adulterants are water-soluble substances added internally to EBN, including ***glucose, sucrose, hydrolysed collagen,*** and ***monosodium glutamate***. In addition, bleaching and staining of swiftlet nests are also routinely incorporated to simplify the tedious and labour-intensive cleaning process (Journal of Food Science, 2013). Currently, no standard methods are fully and truly used for EBN authentication and quality assurance.\

3. Case Study

In August 2011, the EBN industry faced a hurdle as a safety food scandal broke out when China received reports of contaminated swiftlet nests imported from overseas. The abrupt decline in the prices of EBN products occurred following China's imposition of a ban on swiftlet nest imports from Malaysia and Indonesia. This ban was prompted by detecting excessively high nitrite content, measuring 4,400 mg/kg from these regions, significantly surpassing the national limit. (Chu Zhang, 2024, "Journal of Food Quality")

There was a notable reduction in demand, resulting in a 60% drop in the selling price of high graded EBN, plummeting from RM 4,000/kg to around RM 1,200/kg. The nitrate scandal did not exclusively impact EBN exporters from Malaysia and Indonesia; it also affected Thailand. In response, each country took proactive measures by amending various standard operating procedures (SOPs) to ensure the highest EBN quality and safety standards for human consumption. Notably, in September 2012, Malaysia and China signed a protocol governing EBN exports to China, focusing on the examination, quarantine, and hygiene of EBN. Lastly, in December 2013, China lifted its trade ban on EBN imports from Indonesia and Malaysia after both governments signed memoranda of understanding to reinstate this lucrative trade relationship.

Now, this event serves as a reminder of the potential health risks the current EBN industry poses and how unregulated the exporters' activities still are. Despite all quality regulations and efforts to minimize risks, harvesters and traders still find ways to evade said reglementations in order to maximize profits at the cost of the industry's trustworthiness in the eyes of international markets and the consumers' health. Despite being the most notable, this specific scandal is not the only one which has occurred since 2011, raising the following questions: what diseases can swiftlet nests carry and what are the risks of a health crisis breakout?

Regarding the nitrite content risks, the daily intake of nitrite acceptable to the World Health Organization is 0–3.7 mg/kg body weight per day or 222 mg/day for 60 kg adults. In the body, nitrite can be converted into nitric oxide (signaling molecule) that can cause blood vessels to dilate and lower blood pressure. When nitrite and amino acids coexist, carcinogenic

compounds called nitrosamines form during high-temperature cooking, explaining why swiftlet nests pose such a great risk of nitrite contamination (Gunnars, 2020).

The current controversy regarding swiftlet farming is it being “a time bomb for a serious epidemic outbreak” (Ho, 2009: N49). Many researchers claim that swiftlet farms should be moved to rural areas as a regulatory solution, as currently almost all swiftlet farmers prefer the use of inner-city heritage buildings as opposed to purpose-build buildings in rural areas because they are not only much cheaper and easier for them to maintain, but they also devoid of potential predators which would be more prevalent in rural areas. However, this widespread practice is detrimental towards the quality of the industry as it represents a “shortcut” for proper swiftlet farming, and as the industry has developed, swiftlet farms proliferated, transforming the skyline and ambiance of towns across Malaysia and Indonesia. Severe health impacts are a potential effect of the proliferation of swiftlet farms within inner city, residential areas. This threat has been mobilized by researchers as further justification to relocate swiftlet farms to rural, agricultural, or industrial areas.

What is deeply debated today is how the perceived impact of swiftlets and swiftlet farms on human health and quality of life in Malaysian cities is framed by different stakeholders. As a focus point, the state Penang will prove most useful - UNESCO World Heritage City of George Town (Penang) -, which has experienced the most significant controversy over swiftlet farms in the city due to their perceived incompatibility with the UNESCO status. Swiftlet farms were thus declared illegal in the inner city area as of 2010, which resulted in the announcement of a three year ‘grace period’ for existing swiftlet farms to relocate elsewhere. Penang is a small state situated on the Strait of Malacca on the Northwest coast of Peninsular Malaysia, which consists of Penang Island and a sliver of adjacent land on Peninsular Malaysia.

For this case study, data from ethnographic research conducted by Creighton Connolly will be used as evidence. To trace the shifting discourses and politics of swiftlet farming in Malaysia, said multi-sited research was conducted for a total of six months from September 2013 to May 2014. In order to capture as wide of a range of voices as possible, discourses were examined in news articles (digital and print media), web blogs (and associated online comments), and in formal interviews and informal conversations with government officials, swiftlet farmers, heritage activists and city residents. Newspaper articles were collected from the major English dailies in Malaysia, namely The New Straits Times, The Sun, and The Star. At the peak of the birds' nest controversies these newspapers often contained one article on swiftlet farming per week. User comments on online news articles were also analyzed to get a wider sample of narratives.

‘Like any reared animal, house-swifts are messy, smelly potential sources of infection transmissible to humans. Anyone who cares to visit the vicinity of a swift hotel can attest to the unhygienic conditions. In fact, I know of no owners who choose to stay nearby.’ (Tan 2007) As this quotation suggests, health concerns have been one of the most frequently occurring reasons cited in newspapers, blogs, and personal interviews supporting the removal of swiftlet farms from urban areas. As Nash (2006: 9) puts, "…the material qualities of a specific landscape are

critical to the production of certain diseases: local habitats that support anopheles mosquitoes, warm temperatures that allow the survival of parasites and bacteria, [and] the material and biological conditions of human communities." These conditions have all been linked to urban swiftlet farms in Malaysia because of the insalubrious features of most swiftlet farms. In order to illustrate the interior condition of a typical swiftlet house, the following excerpt from Creighton Connolly's field notes will be used: "The first thing that hit me was the stench from the bird droppings layering the floor, as well as the lack of ventilation in the building. The only light coming in was from the entrance hole on the 'jack roof' at the far end of the building. Furthermore, on the roofs of the buildings (typically not visible to most passersby) are often pools of stagnant water from which the swiftlets can drink. Such pools are also typically placed on the interior of the buildings to keep the humidity levels up, or, alternatively, humidifiers are used." Stagnant pools of water present an ideal environment for mosquito breeding and have been targeted as potential sources of dengue. Lung diseases have been attributed to the effluent that is often emitted from buildings, including guano, mildew, mould and associated forms of bacteria. Moreover, the noise emitted from the buildings and the birds inhabiting them negatively affect mental health, it has been argued, and are seen as general sources of nuisance. These characteristics of swiftlet farms have accounted for much of the negative externalities and potential health impacts. "Swiftlet farming in Malaysia has, over the past twelve years, been embedded in an episode of 'acute collective anxiety' set off by perceived risks of dengue fever, avian flu, and other disease outbreaks that could be triggered by the activity in urban areas", says Creighton Connolly.

Four specific areas of concern have been recently raised in regards to the potential impacts of swiftlet farms on public health: ***avian flu, dengue fever, lung and respiratory disease*** and ***quality of life***.

3.1.1 Avian Influenza

Avian influenza, or bird flu, is a common enough condition in wild and domestic birds alike. The popularized spread of diseases between species barriers, such as the H5N1 strain, which crossed to humans in 1997 (in Hong Kong, with 18 cases) has caused considerable concern that the same could happen in swiftlets as well. As one Penang doctor cautioned in a newspaper editorial collected by Creighton Connolly: "We should ask ourselves why these new diseases are cropping up more regularly than ever in the past decade. Is it because human populations are being brought closer to wildlife due to our own selfish needs to satiate our hunger for exotic animals and their products?" (2004). This editorial commended the economic benefits brought about by the swiftlet farming trade, but also cautioned about the 'unknown diseases' that the public could be exposed to as a result. Similarly, in an open letter written by Rebecca Duckett-Wilkinson - a former member of the PHT, 'The Penang Heritage Trust' - to the Penang State Government on June 20, 2010, she cited a claim from the World Health Organization (WHO), that: "All birds are thought to be susceptible to infection with influenza viruses." She also cited a report indicating that "The avian flu virus could be introduced to

swiftlets [...] from migratory waterfowl which are thought to carry the H5N1 virus in its highly pathogenic form, sometimes over long distances.". Living in such close proximity to one another, contaminated swiftlets would quickly transfer the disease between themselves, increasing the chances for human infection either via direct contact with contaminated birds or contaminated feces, the two most common sources of infection. These accounts recognize that swiftlets are not presently carriers of avian flu, but draws attention to the possibility that they could contract the disease if the proliferation of swiftlet farms is to continue unchecked.

The mutability and adaptability of viruses, along with the intensity and density of animal-human cohabitation in Malaysian cities has accounted for considerable public concern. In response to a question about whether swiftlets could potentially spread avian flu, a PHT council member, Clement Liang, shared some of the same concerns, emphasizing the close proximity of the birds to humans and the mutability of viruses between different species. In cementing this point, he asked rhetorically: "…do you really think that swiftlets can't possibly carry bird flu? They still share the same water and food sources of other birds, like dirty insects, and live in a human environment which is polluted with all sorts of things. Then you think about what the insects eat, the dirty leftover garbage and other things from the streets, the city is really an unclean environment." (Liang interview, 2013). This statement also contains a conflation of ideas about the dirt and impurities often associated with the urban environment, with ideas about the proper place of animals within cities. In one of Creighton Connolly's interviews with Liang, he recalled a case in 2013 where there was an alleged outbreak of the bird flu virus H5N1 in Vietnam that was blamed for the death of a five year old boy and nearly 5,000 swiftlets in Southern Vietnam (the Nation, 2013). This raised concerns from the PHT who later announced in an open letter on April 15, 2013 that the event proves that swiftlets are indeed susceptible to H5N1. The event also seemingly disproved the former assumption that swiftlets were somehow immune to avian flu, as one prominent swiftlet farming expert, Chin, once boasted: "No health official anywhere in the world has ever found a single strain of avian flu virus amongst swiftlets" (Chin 2009). Given the 104 recorded swiftlet houses in the George Town World Heritage Site at the time, the PHT argued that this event should serve as a warning to the State Government and Malaysian authorities to take the health threats posed by swiftlet farming in cities seriously and renew their actions to eliminate swiftlet farms from residential areas. Yet, the PHT charged that the Malaysian authorities have been negligent regarding these potential dangers, as the deadly H7N9 strain of bird flu which also emerged in 2013 was being "ignored by authorities." However, it should be noted that human outbreaks of the H5N1 virus to date have been limited in number, and there have been few, if any, clear cases of human-human transmission, which suggests that the current H5N1 virus is poorly adapted to human hosts.

3.1.2 Lung and respiratory disease

A second health concern identified through Creighton Connolly's research is the 'extraordinarily high' incidence of lung disease amongst inner-city residents in the heritage zone, which Duckett-Wilkinson believes is directly correlated to the clustering of swiftlet farms

(Duckett-Wilkinson, interview, 22 October 2013). The reason for this is that dried bird droppings have been known to harbor the yeast spore Cryptococcus, which forms colonies at 20-37 degrees Celsius (making Malaysia's tropical environment ideal) and is known to cause meningitis and lung infections in human beings through inhalation of the spores. Duckett-Wilkinson has spoken to several local doctors in George Town about this issue who have verbally confirmed that cases of lung disease are 'disproportionate' in Georgetown (Duckett-Wilkinson, interview, 22 October 2013). This point reflects insights about how the environmental conditions in a particular place can influence the incidence of disease, and how uneven health landscapes can be produced. Yet, considerable debate emerged around the question of whether or not these were from swiftlets or other birds. For instance, Carole Loh, President of the Association of Swiftlet Nest Industries (ASNI), claimed that these 'are actually that of the pigeons, as swiftlets are very particular and only release their droppings in the house' (in PLGCF 2010: 14). Furthermore, Loh made the (possibly exaggerated) claim that, "5000 tests have been conducted over the course of 10 years, and JPV can confirm that no dangerous pathogens have ever been detected in swiftlet droppings", rather, she attested, "pigeons and crows are the culprits, producing acidic and pathogenic droppings". Moreover, she argued that "the [swiftlet] houses are cleaned weekly" (Tan 2010: M4). Such claims were encountered repeatedly during Creighton Connolly's recorded fieldwork, yet they were also refuted by critics and by proponents of the industry. For instance, Duckett-Wilkinson challenged in response: "I do not see any of the houses around me being cleaned on a weekly basis and maybe [Loh] would like to explain the green slime oozing from the back of 21 China Street - a very large swiftlet house.". Responding to Loh's claims about the droppings, she argued: "I can assure you that the swifts do defecate whilst they fly about and you can see their droppings, which I need to scrub off with disinfectant from all surfaces of my home. I can assure you also that I certainly do not have hundreds of pigeons defecating on my property everyday.". In a similar fashion, many other residents raised concerns about the effluent from swiftlet houses, which is often dumped into largely uncovered city drains along the sides of streets. One respondent elaborated on this danger by noting that: "Bird hotels are well ventilated. Some have exhaust fans on rooftops. Dried bird droppings, dried skin and muck are disposed of through these fans. The naked eye may not be able to see these fine airborne particles, but those living nearby breathe and eat this muck. These particles are also virus carriers" (New Straits Times, 9 September 2009). Such conditions clearly have broader health implications, such as the potential for lung infections or other bacteria-related diseases. These deleterious aspects of the swiftlet farming industry were even acknowledged by the Malaysia Bird's Nest Association President, who cautioned that: "Some members have no experience in the industry and their operations might cause pollution that would harm the environment and health of residents living near their farms" (in Nathan 2003: 13). Yet, local governments and health authorities in Malaysia have not taken these personal accounts of urban health seriously, which has resulted in the proliferation of such unhealthy urban environments.

A further health hazard are swiftlet farms which operate on the upper story of another business, such as restaurants, cafes, and even hotels. This is something that many stakeholders

expressed concerns about, including Chow Kon Yeow, a recurrent customer of such a business, who pointed out that many customers of these buildings are not even aware of the swiftlet premise above. As he argued, "this is totally unacceptable I think...if you put a sign at the top, that says the top is a swiftlet farm, then the customer can choose to go, knowing full well, but deceiving people is not ethical, and people are more than often deceived in such a way." (interview, 22 October 2013). Nonetheless, such business patterns are not uncommon in George Town, and there were two restaurants in George Town's UNESCO zone which had very active swiftlet farms operating on the upper levels during the time of Creighton Connolly's fieldwork. Furthermore, there was also a hotel in Taiping, Perak, which operated a swiftlet farm on the top story, while the bottom three stories were rented out to human occupants. The health danger of such premises was made clear by Duckett-Wilkinson, who cautioned that the businesses underneath can potentially bring in dangerous fungus spores released from the feces through their air conditioning systems. However, the 2007 Malaysian Swiftlet Farming Industry Report claimed that studies criticizing the cleanliness of swiftlet farmers focus on "errant swiftlet farmers" and are inspired by "jealousy and envy of the potential financial returns that such farms may bring" (Merican, 2007), choosing this way to avoid taking accountability for this issue.

3.1.3 Dengue Fever

Dengue fever is a tropical, mosquito borne disease that is strongly associated with urbanization. Cities provide ideal habitats for the Aedes Aegypti mosquito, its primary vector, which depends on anthropogenic water sources for breeding. There are three types of dengue fever in humans, including dengue fever and its more severe forms of dengue hemorrhagic fever, and dengue shock syndrome. There is currently no vaccine or cure for the disease, which means that the primary method for controlling it is the reduction of breeding sites for Aedes mosquitoes. According to numerous sources, one of the primary causes of dengue is rapid urbanization combined with poor or non-existent urban planning (Dickin, 2014; Mulligan, 2012; Shankar, 2014). From Creighton Connolly's notes, it can be extracted that Duckett-Wilkinson and her family have been personally affected by the industry over the years, and that she has been quite active in trying to lobby against the (at the time) pervasive swiftlet farms in George Town. As a result of this experience, Duckett believes that there is a direct link between swiftlet farming and dengue fever, as explained in the following statement: "Just on China Street [her street] we had 5 cases of dengue. On Stewart Lane, Love Lane and Muntri Street, which, according to 2005 figures, have 23 swiftlet houses, there were way too many cases of dengue. I personally know and work with 5 of the people who contracted dengue in this area. This is appalling! Studies have to be done to see if there is a correlation between dengue cases and location of swiftlet farms." (Duckett, letter, May 5, 2011). However, there has been no research done to link cases of dengue to the location of swiftlet farms, despite repeated attempts to get this information from the State Health Department, who seemed unwilling to divulge figures. Therefore, in the absence of official evidence to support or deny the relationship between swiftlet farming and dengue fever, many swiftlet farmers continue to debunk these concerns by arguing that, "a well-run swift farm

has no mosquitoes, as swiftlets consume insects." (Chin 2009). The question of whether or not swiftlets are in fact a way of controlling dengue has been much debated, as indicated in the foregoing paragraph. For instance, Lord Cranbrook, a renowned swiftlet ecologist has demonstrated that their preferred foods include winged termites, flying ants, small hymenopterans including parasitic wasps, and flies. These are found mainly in forests and around plantation fields (Cranbrook 2010). Similarly, Duckett-Wilkinson has argued that: "Stagnant pools of water do attract mosquitoes, but the birds obviously do not eat these, as swiftlets are fully aerial birds and fly out to agricultural or forested areas to feed during the day, not over pools of water". (Duckett-Wilkinson comment on NSFiGT 2010). Likewise, Liang also expressed doubt in Creighton Connolly's interview with him that mosquitoes constitute the main diet of swiftlets: "Simply because they're so small...they would have to eat thousands of them to sustain themselves." (Liang, interview, 12 November, 2009). Their diet has also been studied by several swiftlet farmers who have been interested in the possibility of farming swiftlets in a more traditional (captive) manner, but have not found mosquitoes in their diets. (Pak, interview, 1 October 2013). There is a clear dilemma between the perceived need for insect generation inside swiftlet farms - to provide food for the birds - and, on the other hand, the potential health threats created by potential mosquito breeding areas within urban areas. Such divergent perspectives speak to the proliferation of uncertainties with regards to 'proper' swiftlet farming methods, and their impacts on human health and wellbeing. This speaks to the need for more scientific research on the industry and also proper regulation of farms.

3.1.4 Intangible forms of disease - Quality of life

Many of the externalities posed by swiftlet farming discussed in the previous sections do not only have a detrimental impact on human health, but also on the general quality of life and mental health. This section focuses more on those aspects of swiftlet farming which do not directly cause tangible forms of disease. It is important to note that these should not be viewed as mere nuisance complaints, but have in fact been linked by some individuals to increased stress levels or other impacts on mental health and well-being, which often are not accounted for in standard definitions of health and, particularly, disease. This chapter discusses the sensory impact of swiftlet farming on urban residents and the perceived impacts upon quality of life in the city. For the majority of people, the presence of swiftlets may go unnoticed, depending on their place of residence or work. However, for most people living or working in the vicinity of a swiftlet farm, the presence of such large numbers of swiftlets was 'offensive to the senses.' This resonates with Philo's (1995) study of slaughterhouses in Anglo-American cities, in which he described how the "dreadful smells and sounds" intrinsic to slaughterhouses, in part, resulted in their ultimate removal from the city. The number one complaint from people regarding swiftlet farms is noise, which is actually more prevalent than health concerns. As one resident put it: "Having the swifts for neighbors is like living in hell. No human being should be forced to live this way. I can't remember the last time I was able to open my windows for some fresh air. I just can't take this anymore." (Resident, Kuala Terengganu, quoted in Muhammad 2007). This was a

common experience of critics of the industry, who did not open their windows in fear that swiftlets would fly in. Moreover, as Rebecca Duckett-Wilkinson wrote in a letter to UNESCO: "On a daily basis we have to put up with these birds flying over us and their droppings smudging all over our walls, courtyards and outdoor furniture. The noise of the birds is appalling; terrible while they are flying around at daybreak and sunset, and a constant nuisance at night." (Duckett-Wilkinson letter to UNESCO, May 5, 2011). However, such noise complaints do not only refer to the noise of the birds themselves, but also to the recordings that many operators use to attract swiftlets to their premises. These broadcast swiftlet calls are emitted on specially designed sound systems and played in the morning and the evenings, which are the most active times for the birds. As one city resident complained: "All over these breeding places, noisy recorders are installed to lure the swallows from early dawn till the night. We are, in fact, living with the swallows" (in The Star, 3 March 2005). Another George Town resident explained that: "This has severely affected our quality of life. On weekends, when we look forward to just relaxing at home, we have to put up with the aural assault." (Lee 2010). One question which could arise is whether such noise pollution is actually linked to mental health and quality of life, or if it is just a nuisance.

4. Ethical Debates regarding Bird Welfare versus Economic Gains

The evaluation of swiftlet farming sits at an unusual intersection in conservation bioethics: it is neither traditional factory farming, as the animals remain free-roaming, nor purely wild harvest, as humans construct specialized microclimates and use acoustic playback to manage nesting and so on. Debating whether this industry serves as a sustainable, non-degradable rural ranching activity which has the benefit of incentivizing forest canopy protection, or if its base premise of exploiting un-domesticated avian species is ethically problematic requires weighing two competing paradigms. Critics furthermore question whether relying on wild birds for intensive commercial gain constitutes ethical animal husbandry, as the birds remain free-roaming foragers dependent on local insect populations and forest habitats. Two major key points are worth debating in this situation: the economical benefits versus the ethical risks.

From an economic standpoint, there can be observed several bonuses and gains which make the current state of the swiftlet farming industry more than adequate for the participants who do not intend to conform to bioethical norms. The high market demand for swiftlet nests is a prime motive for maximizing output regardless of the consequences at all costs. Moreover, some participants enthusiastically claim that the process of utilizing ex-warehouses or constructing specially designed spaces inter-city for the raising of swiftlets contributes to the rural urbanization of towns and provides a boost in regional construction and the pool of local job opportunities, furthermore increasing property values. One cannot argue that this industry is not a successful and flourishing one and it constitutes a much needed market for the producer countries. Removing swiftlet farming activities altogether is not a viable option, as the lives of tens of thousands of independent farmers have been built on harvesting swiftlet nests, with an

approximate of at least 100,000 specialized nesting structures as of 2026 (Vietnam News Agency), with the industry generating a minimum of 5 billion USD per year, driven by soaring demand in Asian markets and other alike (Cognitive Market Research, 2025). From an ethical point of view, significant academic debates have emerged along the years, raising several main concerns: ***avian welfare, harvest timing, habitat and behavioral manipulation.***

4.1.1 Avian Welfare

The foundational ethical conflict in swiftlet farming stems from the biological nature of nest construction. Unlike standard avian species that weave nests from twigs or mud, swiftlets construct their nests entirely from hardened saliva. The nest is not a redundant byproduct; it is a vital parental investment. Sustainable management protocols dictate that nests should only be harvested after young swiftlets have completed their breeding cycle and fledged. However, market pressure and theft risks frequently lead to premature harvesting. Removing nests while eggs or nestling chicks are inside leads directly to chick mortality and reproductive disruption. When said nests are repeatedly harvested, swiftlets are forced into a state of chronic physiological exhaustion. Driven by reproductive instincts to protect their clutches, the birds expend metabolic reserves to reconstruct lost nests. Constant nest reconstruction diverts crucial metabolic energy away from the birds' immune function and adult maintenance, increasing mortality during stressful foraging or seasonal transitions. In addition, female swiftlets under severe physiological stress may lay eggs on bare concrete ledges when nest removal occurs faster than their rebuilding speed, leading to total reproductive failure. By viewing the bird's internal physical secretions purely as an extractable commodity, the industry subordinates basic physiological health to economic yield.

4.1.2 Harvest Timing

The timing of nest collection represents the most direct conflict between financial profit and animal welfare as of now. Harvesting the nests *after* the eggs have hatched and nestlings have fully fledged and abandoned the nest does not provide the same quality as that of an “unused” nest, ultimately reducing its market price. These nests contain feathers, eggshell fragments, and feces, making them labor-intensive to clean. This economic structure incentivizes premature harvesting and ethical tradeoffs. When operators harvest nests containing active clutches or young nestlings, the eggs are discarded, and un-fledged chicks fall to the concrete floor to die of starvation or trauma. Even when nests are taken prior to egg-laying, the birds are forced into an immediate, exhausting re-nesting loop. While eco-certification bodies advocate strictly for post-fledging harvests, widespread theft, lack of regulatory oversight, and price disparities mean premature harvesting remains a common issue across decentralized rural swiftlet houses.

4.1.3 Habitat and Behavioral Manipulation

Environmental manipulation is used to entice wild birds to inhabit human-made structures. Swiftlets are wild, non-domesticated migratory organisms. Swiftlet farming does not domesticate the species in a genetic sense. Farmers install high-decibel external speakers running continuous recordings of swiftlet colony distress and mating calls ("tweeters") to lure wild birds passing overhead. Internally, localized calls trick birds into believing a colony is safe, established, and thriving. This continuous acoustic bombardment alters natural flocking and navigation behaviors, creating artificial population densities in urban and semi-urban landscapes. Ethicists debate whether utilizing behavioral trickery to exploit a wild species' reproductive instinct constitutes a form of involuntary captivity. This creates an asymmetric relationship: humans exert total environmental control over wild animals without offering nutritional support, medical care, or domestic safeguards (Taylor & Francis, “Swiftlets Becoming Neighbors: Sensory Place-making in the Edible Birds’ Nest Trade in Kapuas Hulu, Indonesia”, 2024).

4.2 Forecast

Creating a sustainable economic model requires identifying the core motivations and systemic risks faced by each actor within the supply chain matrix. For the swiftlet farming industry to transition toward long-term sustainability, it must abandon the traditional, extractive model in favor of a synchronized framework where the success of every stakeholder, from smallholder farmers and consumers to regulatory bodies and avian populations, is mutually reinforcing. Smallholder farmers and commercial house owners face heavy upfront capital investments in building structures, controlling microclimates, and running acoustic lures. Their primary drivers are financial return, debt mitigation, and yield security. Under un-certified systems, producers are frequently driven to harvest nests before eggs are laid. To align producer interests with ethical practices, sustainable farming must mitigate these operational risks and provide ***clear financial advantages*** for post-fledging harvests. As for the consumers of EBN which face significant market risks, when they lack a way to verify ethical harvesting practices, the price of said harvested nest should be lowered accordingly. Government bodies, municipal planners, and local communities are tasked with maintaining public health and environmental balance. Unregulated urban swiftlet houses raise concerns for the locals. Regulations which impose audit-ready data, enforceable biosecurity standards, and proper zoning to ensure the industry should be created and enforced in order to support local economic growth without compromising public health.

Discussion

The empirical data and case studies analyzed across this study strongly validate the primary thesis: the contemporary swiftlet farming industry operates on an extractive, economically driven model that systematically compromises animal welfare, consumer safety, and urban biosecurity.

A critical insight emerging from this research is the liminal status of the semi-domesticated *Aerodramus fuciphagus*. As highlighted by Thorburn, swiftlet farming

escapes traditional industrial livestock critiques because the birds remain free-roaming foragers. However, this study demonstrates that "free-roaming" is decoupled from behavioral freedom. The deliberate deployment of high-decibel acoustic looping devices and artificial microclimates constitutes an asymmetric form of behavioral manipulation. This creates an ethical vacuum: humans exert total environmental control over wild animals to harvest their biological secretions without providing any domestic safeguards, veterinary care, or nutritional provisioning.

The data regarding harvest timing reveals a systemic, market-driven conflict between financial yield and conservation bioethics. Because "clean," unused nests fetch premium market prices, operators are heavily incentivized to harvest nests prematurely, directly violating post-fledging sustainability protocols. When nests are removed prior to egg-laying or while active clutches are present, the extreme reproductive drive of the birds forces them into an exhausting re-nesting loop. The metabolic cost of continuously generating hardened saliva shifts vital physiological energy away from the birds' immune and adult maintenance functions.

The proliferation of swiftlet farming inside inner-city historical zones, most notably observed in the George Town case study, presents an immediate public health hazard that challenges standard paradigms of urban livestock keeping. High-density roosting spaces act as ideal incubation chambers for viral recombination. Because swiftlets forage over expansive regions, they serve as geographic vectors capable of acquiring highly pathogenic avian influenza strains from migratory wild fowl and bringing them directly into high-density urban residential centers. The intentional use of internal water pools and humidifiers to mimic cave microclimates creates artificial breeding zones for *Aedes aegypti* mosquitoes, directly correlating with local spikes in Dengue fever. The massive accumulation of unmanaged avian feces generates severe respiratory risks. Warm, damp guano allows the yeast spore *Cryptococcus* to thrive.

While national quality specifications exist—such as Malaysia’s MS 2334:2011 and Thailand’s TAS 6705-2014—their efficacy is heavily undermined by low compliance rates and decentralized trading platforms. As demonstrated by Rabu and Nazmi, over half of domestic producers entirely bypass formal grading frameworks in favor of traditional visual systems or bulk selling. This resistance to serialization creates a fragmented supply chain where chemical adulteration and toxic nitrite contamination remain highly profitable. To resolve these compromises, the industry must transition from an extraction model to a verified, technology-backed stewardship model.

However, several limitations must be acknowledged within this research: A significant hurdle remains the lack of transparent, localized epidemiological data from state health departments. The ongoing reluctance of regulatory bodies to publish official cross-referenced figures matching Dengue or respiratory clusters to swiftlet house locations forces researchers to rely heavily on qualitative, self-reported ethnographic data from affected residents. Additionally, the architectural and social insights are heavily weighted toward Southeast Asian producer nations (Malaysia, Indonesia, Thailand). The rapidly emerging markets and production spaces within other equatorial nations may display distinct socioeconomic behaviors and regulatory compliance patterns not fully captured in this literature analysis.

Works Cited

Babji, A. S., Nurfatin, M. H., Etty Syarmila, I. K., & Masitah, M. (2015). Secrets of edible bird nest. UTAR Scientific Journal, 1(1), 7–14.

Brooke, R. K. (1970). Generic limits in old world Apodidae. Bulletin of the British Ornithologists' Club, 90, 116–124.

Cognitive Market Research. (2025). Global edible bird’s nest market report 2025: Industry analysis, market size, share, growth, trends, and forecast to 2032. Cognitive Market Research.

Connolly, C. (2016). "A place for everything": Moral landscapes of "swiftlet farming" in George Town, Malaysia. Geoforum, 77, 182–191.

Dai, Y., Cao, J., Wang, Y., Chen, Y., & Jiang, L. (2021). A comprehensive review of edible bird’s nest: Composition, functions, and industrial applications. Food Research International, 140, Article 109875.

Department of Standards Malaysia. (2011). Edible-birdnest (EBN) — Specification (MS 2334:2011). Ministry of Science, Technology and Innovation.

Lee, T. H., Wani, W. A., Lee, C. H., & Cheng, K. K. (2021). Edible bird’s nest: Recent updates and industry insights based on laboratory findings. Frontiers in Pharmacology, 12, Article 746656.

Rabu, M. R., & Nazmi, M. S. M. (2015). Malaysia’s edible bird nest (EBN) industry. Food and Fertilizer Technology Center for the Asian and Pacific Region (FFTC-AP) Agricultural Policy Platform.

Thorburn, C. C. (2015). The edible nest swiftlet industry in Southeast Asia: Capitalism meets commensalism. Human Ecology, 43(1), 179–192.

Zulkefle, N. N., Ibrahim, M. A., Azuan, N. F., Ch'ng, S. E., Ismail, N., Abu Bakar, M. Z., Chan, K. W., & Zhang, C. (2024). A review on the edible bird's nest quality and manufacturing standards of the three largest exporting countries in the world. Journal of Food Quality, 2024, Article 5608357.

Improving Option Pricing Accuracy Using a Hybrid Machine Learning and Black-Scholes Model By Vyshal Sreenivas

Abstract

The Black-Scholes model (BSM) is the dominant framework for pricing options and derivatives in quantitative finance, yet its fixed assumptions regarding constant volatility and lognormal price distributions limit its accuracy in volatile or illiquid markets. Standalone machine learning (ML) models have also failed to regularly outperform the BSM across all market conditions. We hypothesized that a hybrid model that combines a residual neural network (ResNet) architecture with the Black-Scholes framework would improve the validity of call option pricing compared to the standalone BSM. Real market option-chain data for Apple Inc. (AAPL) call options were obtained from marketdata.app, yielding 666 cleaned contracts spanning 24 expirations from a single quote date. To avoid target leakage, volatility was computed as the 30-day historical (realized) volatility of the underlying asset rather than implied volatility, which is back-solved from option prices. The hybrid model decreased the mean absolute percent error (MAPE) from 18.68% to 4.21%, a 77% reduction in error; improved the coefficient of determination (R^2) from 0.864 to 0.986; and decreased the mean absolute error (MAE) from $8.20 to $2.09. The hybrid model also tracked the observed volatility smile more closely than the standard BSM.

Keywords

Finance, Quantitative Finance, Black-Scholes Model, Residual Neural Network, Option Pricing

Introduction

The Black-Scholes model, often referred to as the “trillion-dollar equation,” serves as the basic framework for calculating options, derivatives, stocks, shorts, and betas in quantitative finance.[1] Financial professionals, including analysts, researchers, and quantitative analysts, rely on the BSM for risk management and objective valuation.[1] However, inaccuracies in the model can lead to misleading financial reporting, notably concerning employee stock options.[1] Resolving the limitations of the BSM and developing a more accurate model for valuing employee stock options could enhance corporate financial reporting and accountability.
The BSM demonstrates limited effectiveness in analyzing volatile or rapidly changing markets.[2] Consequently, there is a need to develop new option-pricing models, either by integrating artificial intelligence (AI) into existing algorithms or by fundamentally redesigning present methodologies. Previous attempts to innovate in trading models have often been unsuccessful, partly because standalone AI models have not consistently outperformed the BSM.[3] Although AI models offer greater flexibility by incorporating real-time market factors, they often yield inaccurate predictions of market sensitivity. In contrast, the BSM relies on fixed-input algorithms

engineered for stable markets, but these static parameters can result in significant volatility and unreliability when external factors disrupt market equilibrium.[2]

This study considers the extent to which hybrid models that combine machine learning techniques with the Black-Scholes framework can improve option pricing accuracy in volatile and illiquid markets. The objective is to develop a model that achieves higher precision in option price prediction while requiring less data and fewer computing resources than alternative methods, thereby enabling more efficient fair value calculations for stock analysts. The research generates risk and price prediction data, which are compared with observed market records to assess model effectiveness. Ultimately, this work aims to provide quantitative analysts, stock analysts, and risk managers with a more accurate, data-efficient tool for risk assessment and price prediction.

Current debates in the literature focus on whether standalone AI models or alternative pricing methods, such as the binomial pricing model and Monte Carlo simulation, can achieve greater forecasting performance.[4,5] The present study merges three key themes, mathematical modeling, option trading, and artificial intelligence, to develop an improved option pricing model.

There is a considerable gap in the literature regarding the simultaneous application of AI with the standard BSM or other pricing equations. Developing combined models may address inefficiencies present in both standalone AI and BSM approaches. Amilon demonstrates that, despite the flexibility of multilayer perceptron (MLP) networks, these models frequently overfit, misinterpreting temporary market oscillations as persistent trends.[3] While such overfitting highlights the limitations of certain AI architectures for price prediction, it also offers opportunities to evaluate alternative AI models and identify stronger solutions. D'Uggento and colleagues challenge this pessimistic view by showing that machine learning algorithms can identify nonlinear patterns in data and outperform the BSM across moneyness, the relationship between an option's strike price and the current price of the underlying asset, and across maturities.[6] They also suggest that machine learning techniques such as random forests, used in a hybrid fashion with the BSM, can improve pricing performance. Building on that concept, Qiao and Wan identify deep learning approaches, specifically a residual learning network, that can improve BSM delta-hedging performance with less data.[7] Residual neural networks can reduce data scarcity stemming from financial volatility, a problem that affects other AI models. Paialunga further shows that physics-informed neural networks can outperform the standard BSM by training a feed-forward neural network on the BSM's partial differential equation (PDE), thereby decreasing error and producing a more accurate model.[8]

Many models are used for option pricing and risk hedging. Chakrabarti and Santra argue that, compared to the standard BSM, the Heston model, which adds stochastic volatility to Black-Scholes, tends to perform better, even though specific combinations of moneyness and time to maturity exist where the BSM holds its own.[9] This suggests that a hybrid AI model could dynamically switch between the standard BSM and the Heston specification depending on market conditions. Ahmad Dar and Anuradha argue that other competitive models, such as the

binomial model, do not significantly outperform the BSM according to the Tukey test, and that the binomial model also requires more data and greater operating power.[4] This supports the view that the BSM is better suited to low-computing-power environments and is a stronger foundation for building an AI model, since a BSM-based AI provides more predictable results than the binomial model, which models multiple possible outcomes. Jankova, however, notes that the assumptions of the BSM, in particular the assumption of constant volatility across different levels of moneyness, limit its suitability in contemporary markets.[2] An AI model can be trained to predict the difference between BSM values and actual values, thereby lessening the effects of these limiting assumptions.

Haug and Taleb stress that while the BSM is simple in nature, its assumptions, such as constant volatility and a lognormal price distribution, limit its usefulness in unstable market conditions.[10] They argue that most real-world pricing deviations stem from volatility smiles and heavy-tailed return behavior that the BSM cannot handle. Building on this idea, Derman and Kani argue that market participants should use local volatility and implied tree models to address BSM drawbacks by incorporating volatility that varies with strike price and time.[11] While their model offers a solution for the rigidity built into the BSM, it still relies heavily on calibration and historical volatility estimates, an issue that a hybrid AI model can address by training directly on market data. Ding and colleagues suggest that deep learning frameworks can outperform classical numerical methods, achieving higher out-of-sample accuracy in pricing American options at lower computational cost.[12]

Ruf and Wang conducted a critical literature review, identifying data leakage and overfitting as persistent obstacles to the realistic application of standalone neural network architectures.[13] Their findings informed the data partitioning strategy used here: 70% for training, 15% for validation, and 15% for testing. Sood *et al.* compared four machine learning models, support vector machine, extreme gradient boosting, multilayer perceptron, and long short-term memory (LSTM), against the BSM benchmark using S&P 500 Apple stock option data from 2018 and 2019.[14] LSTM achieved the best performance for both call and put options and remained consistently superior across various levels of moneyness and maturities. However, no standalone machine learning model could regularly outperform the BSM across all market conditions, further strengthening the case for a combined architecture.

Previous research has primarily concentrated on improving the BSM via mathematical extensions, such as the Heston model, or by applying standalone AI algorithms to option pricing. However, a significant gap remains in the development of hybrid mathematical-AI frameworks that preserve theoretical interpretability while adjusting to real-time market dynamics and volatility. We hypothesized that integrating the BSM with AI in a residual learning architecture would yield more accurate fair value predictions than the standalone BSM, especially in unstable market conditions.

Methods

Research design

To achieve higher accuracy and precision, a hybrid model was built that combines machine learning techniques with the BSM to produce more accurate fair value predictions. The project used real AAPL call-option market data obtained from marketdata.app, including traded option prices, underlying asset prices, strikes, and expirations. Because the goal was to compare predictions from the standalone BSM against the hybrid architecture, the study relies on experimental quantitative data generated from model forecasts and error metrics such as RMSE, MAE, and MAPE.

Materials

Real market option-chain data for Apple Inc. (AAPL) call options were obtained from marketdata.app, which provided underlying asset prices, strike prices, expirations, traded option prices (bid, ask, and last), and implied volatilities (Table 1). Because implied volatility is backsolved from option prices and would introduce target leakage, it was not used as a model input; instead, volatility was computed as the 30-day historical (realized) volatility of the underlying asset. All programming and development were conducted in Python 3.11.9 with the following libraries: NumPy 2.3.4 for numerical computation, pandas 2.3.3 for data cleaning and structuring, SciPy 1.16.2 for mathematical functions, scikit-learn 1.8.0 for feature preprocessing and data splitting, and PyTorch 2.10.0 for implementing and training the neural network components of the hybrid model. Option-chain retrieval, file handling, and timing used the Python standard library modules urllib.request, json, os, and time, which are versioned with the interpreter.

Table 1. Data collection sources and purposes. Spot price and traded market price data were obtained from marketdata.app and used as both model inputs and targets for residual analysis, while volatility was computed as the 30-day historical realized volatility of the underlying asset.

Data type	Source	Frequency	Purpose
Spot price (S)	marketdata.app	Daily	BSM input / residual target
Strike (K)	marketdata.app	Static	Contract specification
Market price (C)	marketdata.app	Snapshot	Ground truth for residual analysis
Risk-free rate (r)	Constant (5%)	Static	BSM input
Historical volatility (σ)	Computed	Calculated	30-day realized volatility of AAPL returns

Data collection

Option-chain data for AAPL call options were collected from marketdata.app for a single quote date, spanning 24 distinct expirations (Table 1). The raw chain was cleaned by removing rows with missing bid or ask prices, computing the traded market price as the bid-ask midpoint, removing options with non-positive time to expiration or non-positive volatility, removing contracts priced below $0.01, and restricting moneyness (S/K) to the range 0.75 to 1.25 to

remove deep in-the-money and out-of-the-money noise. After cleaning, the dataset comprised 666 call-option contracts.

Data cleaning and normalization

Using Pandas, rows with missing bid or ask prices were removed, contracts priced below $0.01 were filtered out, and deep in- and out-of-the-money options were removed by restricting moneyness to 0.75 -- 1.25. With NumPy and Pandas, all model inputs were computed: moneyness (S/K), log-moneyness, time to expiration, 30-day historical volatility, the underlying spot price, the strike, and the risk-free rate. The cleaned dataset of 666 contracts was split into 70% training (466 contracts), 15% validation (99 contracts), and 15% testing (101 contracts). Features were standardized using a scaler fitted only on the training set to prevent information leakage into the validation and testing sets.

Hybrid model architecture

The hybrid model begins with raw market inputs, spot price (S), strike price (K), risk-free rate (r), historical volatility (σ), and time to expiration (τ), which are passed through the standard BSM to generate a base price prediction (Figure 1). A residual neural network, consisting of a dense layer (32 units) followed by two residual blocks with identity shortcuts, then learns a systematic correction term (Δ) that accounts for volatility surface deviations and market frictions. The final hybrid price is the sum of the BSM base prediction and the learned correction: Hybrid Price = BS + Δ.

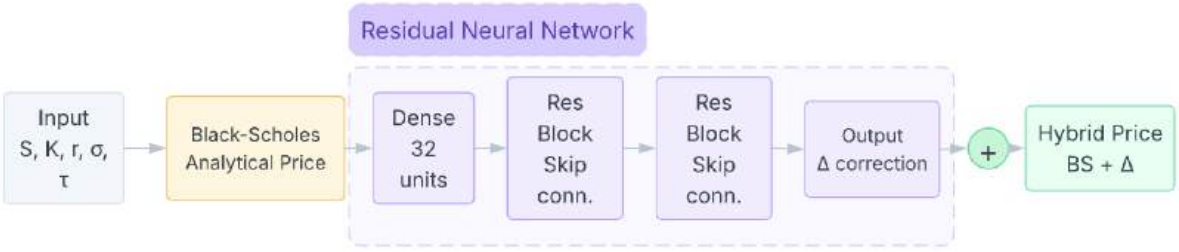


Figure 1. Hybrid model design combining the Black-Scholes model with a residual neural network. Raw market inputs (S, K, r, σ, τ) pass through the Black-Scholes analytical pricing model to generate a base prediction, and the residual network learns a systematic correction term (Δ) that is added to it. Diagram created by the author.

Training and evaluation

Using SciPy and NumPy, the standard BSM formula was computed for each option in the dataset to provide baseline predictions. The machine learning component was then constructed in PyTorch as a residual-learning neural network consisting of a dense layer (32 units) followed by two residual blocks with shortcut connections. The model was trained to learn the correction term Δ, the difference between the actual market price and the BSM price, by minimizing loss between hybrid outputs and actual market prices, allowing the machine learning layer to adjust for volatility shifts, non-normal price movements, and other deviations described in the

literature.[3,6,7] All machine learning inputs, such as moneyness, volatility, spot price, and days to maturity, were normalized using scikit-learn preprocessing functions. After training, the hybrid model was evaluated on the withheld test set. Performance measures, including the coefficient of determination (R^2), root mean square error (RMSE), mean absolute error (MAE), and mean absolute percent error (MAPE), were computed to compare the accuracy of the hybrid model relative to the standard BSM. The maximum error was also recorded to assess tail risk.

Software and data availability

All analysis code, including the Black-Scholes baseline, the residual network, and the evaluation routines, was written by the author in Python (https://www.python.org). Figure 1 was produced by the author as an original schematic diagram, and Figure 2 was generated by the author using Matplotlib 3.10.7 (https://matplotlib.org). Tables 1 and 2 were compiled by the author from his own analysis. Option-chain data are available from marketdata.app (https://www.marketdata.app).

Results and Discussion

Comparative performance

Model effectiveness improved substantially across all five error metrics when comparing the standard BSM to the hybrid model (Table 2).

Table 2. Comparative performance of the standard Black-Scholes model and the hybrid model on the withheld test set. Across every metric the hybrid model outperformed the standalone Black-Scholes model, most notably reducing mean absolute percent error by 77%.

Metric	**Standard Black-Scholes**	**Hybrid (BSM + residual ML)**	**Improvement**
R^2 (coefficient of determination)	0.864	0.986	+14.2%
MAPE (mean absolute percent error)	18.68%	4.21%	77% reduction in error
RMSE (root mean square error)	$8.33	$2.68	~68% better fit
MAE (mean absolute error)	$8.20	$2.09	~75% better fit
Maximum error (outlier)	$11.46	$7.65	Improved tail risk

Explanatory power

The coefficient of determination (R^2) increased from 0.864 to 0.986 (Table 2). The BSM's R^2 of 0.864 indicates that it captures core market features such as intrinsic value and time decay, but leaves approximately 14% of price variance unaccounted for, variance typically arising from the volatility smile and market liquidity premiums. The hybrid model's R^2 of 0.986 indicates that the residual network successfully learned the shape of the volatility surface.

Accuracy in tail events

MAPE decreased from 18.68% to 4.21% (Table 2). This metric is particularly important when evaluating out-of-the-money options, which the standard BSM often significantly underprices by underestimating the probability of extreme market movements. The 77% reduction in MAPE suggests that the hybrid model yields a more thorough estimate of fat-tail risks, a well-documented gap in theoretical pricing models.

Error magnitude and risk management

Both RMSE and MAE improved by more than 67% (Table 2). Because RMSE penalizes larger errors more heavily, the significant decrease indicates that the hybrid model is effective at eliminating extreme mispricings. By narrowing the mean absolute error to $2.09 per contract, the hybrid model transforms a theoretical estimate into a more precise pricing tool.

Volatility smile comparison

The implied volatility comparison across all three model conditions confirms that the BSM's flat-volatility assumption leads to systematic divergence from the true underlying volatility surface at different levels of moneyness, whereas the hybrid model closely tracks it (Figure 2).

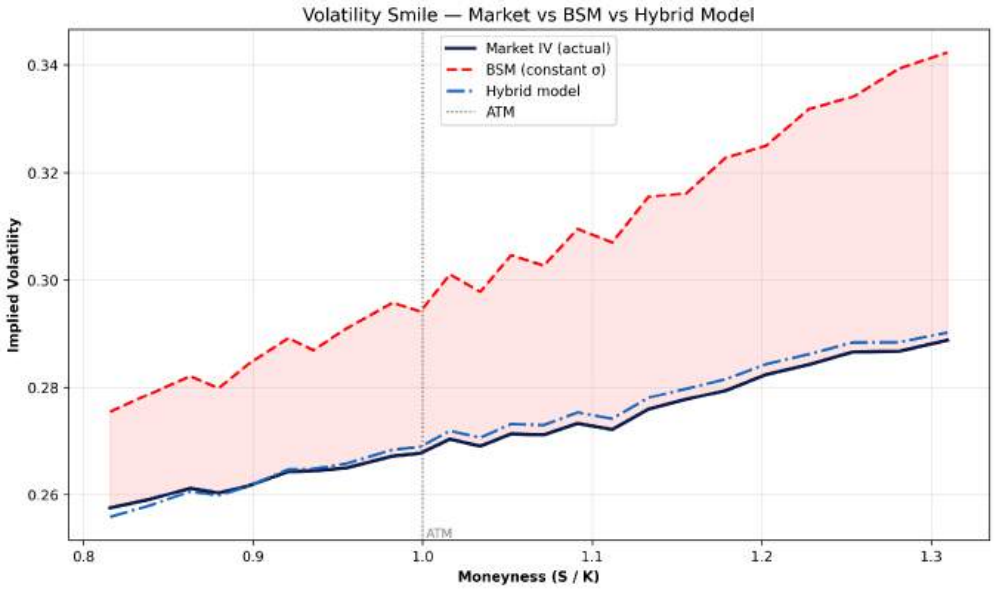


Figure 2. Implied volatility comparison across moneyness levels for 101 test-set AAPL call options, binned into 25 equal-width moneyness intervals. The hybrid model tracks the market implied volatility curve closely across all moneyness levels, while the Black-Scholes prices diverge systematically, illustrating the volatility smile the standard model fails to capture. Figure generated by the author using Matplotlib.

Interpretation

These results show that the residual neural network hybrid model outperforms the standalone Black-Scholes formula by capturing nonlinear market behavior. By reducing MAE to $2.09, the model demonstrates that machine learning can capture the shape of the real volatility

surface, a feature that the classical model holds constant. The mispricings produced by the standard BSM are therefore a learnable, correctable error, consistent with the argument that machine learning methods identify nonlinear patterns in option data,[6] and with the use of a residual learning architecture as a practical foundation.[7] Figure 2 empirically confirms Haug and Taleb's identification of the volatility smile as a systematic market bias.[10]

These results also address the divide between classical financial theory and empirical market microstructure. Conventional practice holds that correcting the BSM requires adopting more sophisticated stochastic volatility models, such as the Heston model.[9] However, the hybrid approach attains similar or better accuracy without the computational weight of solving complex partial differential equations, consistent with Nayak's conclusion that hybrid architectures offer a favorable balance between predictive strength and computational viability.[15]

Target leakage

A central methodological concern in this type of study is target leakage. Implied volatility, which marketdata.app and similar vendors report, is back-solved from the very option price the model attempts to predict. An initial run that used implied volatility as an input produced an artificially near-perfect fit, with the standalone Black-Scholes model alone reaching an R^2 of 0.998, confirming that implied volatility leaks the target. To address this, the final model used 30-day historical (realized) volatility, computed solely from the underlying asset's returns, which removed the leakage and yielded the genuine improvement reported here. This concern is consistent with Ruf and Wang's warnings about data leakage in neural network option pricing models,[13] and with Amilon's findings on overfitting.[3]

Limitations and future work

A further limitation is that the dataset represents a single quote date rather than a multi-year history, so it captures one cross-section of the volatility surface rather than behavior across changing market regimes. Subsequent studies should extend the analysis across many trading days and varied market regimes, such as the elevated volatility of the 2020 COVID-19 market crash, to assess robustness.

Reliance on a free data source also introduces limitations. The marketdata.app option chain represents a snapshot rather than continuous tick-level data, and the study used a single quote date, which limits the size and temporal breadth of the dataset to 666 contracts. Future researchers should consider procuring institutional-grade, multi-day or tick-level option data from vendors such as OptionMetrics or Bloomberg, which would expand the sample, span multiple market regimes, and enable near-real-time validation.

A further concern is the loss of decision interpretability. The standard BSM provides an explicit pricing mechanism, and the Greeks (delta, gamma, vega) can be calculated with mathematical certainty. The machine learning residual, by contrast, functions as a black box, an issue that carries consequences for regulatory compliance, since financial firms must justify specific trading decisions. Applying explanation frameworks such as SHAP (Shapley Additive

explanations) values, which decompose a neural network's output into contributions from individual input variables, could provide explainability without sacrificing model effectiveness.

Conclusion

The results show that the residual neural network hybrid model outperforms the standalone Black-Scholes formula by identifying nonlinear market movement. By reducing MAE to $2.09, the model shows that machine learning can capture the shape of the real financial volatility surface, a feature that conceptual models hold constant. This research finds that the future of quantitative finance does not lie in replacing standard models but in augmenting them with data-driven layers to account for a more complex modern market microstructure. The blended approach developed here represents a meaningful step toward more accurate, accessible, and accountable option pricing for analysts, risk managers, and financial institutions.

Authors

Vyshal Sreenivas is a student at Hamilton High School in Chandler, Arizona, with interests in quantitative finance, mathematical modeling, and machine learning. He designed the study, wrote the analysis code, and prepared the manuscript.

Works Cited

1. PwC. 8.4 The Black-Scholes Model. Viewpoint. https://viewpoint.pwc.com/dt/us/en/pwc/accounting_guides/stockbased_compensat/stockbased_compensat__3_US/chapter_8_estimating_US/84_the_blackscholes__US.html (accessed 2024-06-30).
2. Jankova, Z. Drawbacks and Limitations of Black-Scholes Model for Options Pricing. J. Financ. Stud. Res. 2018, 2018, 1–7. https://doi.org/10.5171/2018.79814.
3. Amilon, H. A Neural Network versus Black–Scholes: A Comparison of Pricing and Hedging Performances. J. Forecast. 2003, 22 (4), 317–335.
4. Ahmad Dar, A.; Anuradha, N. Comparison: Binomial Model and Black Scholes Model. Quant. Finance Econ. 2018, 2 (1), 230–245. https://doi.org/10.3934/QFE.2018.1.230.
5. Culkin, R.; Das, S. R. Machine Learning in Finance: The Case of Deep Learning for Option Pricing. J. Invest. Manage. 2017, 15 (4), 92–100.
6. D'Uggento, A. M.; Biancardi, M.; Ciriello, D. Predicting Option Prices: From the Black-Scholes Model to Machine Learning Methods. Big Data Res. 2025, 40, 100518. https://doi.org/10.1016/j.bdr.2025.100518.
7. Qiao, C.; Wan, X. Enhancing Black-Scholes Delta Hedging via Deep Learning. SSRN Electron. J. 2024. https://doi.org/10.2139/ssrn.4886055.
8. Paialunga, P. When Physics Meets Finance: Using AI to Solve Black-Scholes. Towards Data Science, April 18, 2025. https://towardsdatascience.com/when-physics-meets-finance-using-ai-to-solve-black-scholes/ (accessed 2025-06-15).
9. Chakrabarti, B.; Santra, A. Comparison of Black Scholes and Heston Models for Pricing Index Options. SSRN Electron. J. 2017. https://doi.org/10.2139/ssrn.2943608.
10. Haug, E. G.; Taleb, N. N. Option Traders Use (Very) Sophisticated Heuristics, Never the Black–Scholes–Merton Formula. J. Econ. Behav. Organ. 2011, 77 (2), 97–106. https://doi.org/10.1016/j.jebo.2010.09.013.
11. Derman, E.; Kani, I. Riding on a Smile. Risk 1994, 7 (2), 32–39.
12. Ding, L.; Lu, E.; Cheung, K. Deep Learning Option Pricing with Market Implied Volatility Surfaces. arXiv September 7, 2025. https://doi.org/10.48550/arXiv.2509.05911.
13. Ruf, J.; Wang, W. Neural Networks for Option Pricing and Hedging: A Literature Review. J. Comput. Finance 2020, 24 (1), 1–46. https://doi.org/10.21314/JCF.2020.390.
14. Sood, S.; Jain, T.; Batra, N.; Taneja, H. C. Black–Scholes Option Pricing Using Machine Learning. In Proceedings of International Conference on Data Science and Applications; Lecture Notes in Networks and Systems, Vol. 551; Springer: Singapore, 2023; pp 481–493. https://doi.org/10.1007/978-981-19-6631-6_34.
15. Nayak, N. S. Mathematical Modeling of Option Pricing with an Extended Black-Scholes Framework. arXiv April 4, 2025. https://doi.org/10.48550/arXiv.2504.03175.

Election Cycles and Short-Term Fiscal Policy: A Comparative Study Between Parliamentary and Presidential Systems By Ginevra Colapinto

Abstract

This empirical paper examines the relationship between election cycles and short-term fiscal policy decisions and whether these effects differ between presidential and parliamentary systems. The paper adopts a comparative approach by focusing on the United States and Germany, these two countries were strategically chosen as they both possess similar economic development and stability, which therefore ensures the analysis solely focuses on differences in political institutions rather than economic conditions and development levels. The study utilises secondary quantitative data from the World Bank's World Development Indicator which measures the general government consumption expenditure of both countries as a percentage of GDP between 2000 and 2023. To answer the main research question addressed within this analysis: What is the relationship between election cycles and short-term fiscal policy decisions in presidential versus parliamentary political systems?, the paper compares election years and non - election years to assess and compare short term fiscal behaviour. The findings ultimately suggest that government consumption expenditure increases modestly during election years in both systems, however, Germany consistently spends larger sums of money on government consumption than the United States.

Keyworks Fiscal policy; Election cycles; Government consumption expenditure; Presidential systems; Parliamentary systems; Political institutions

1. Introduction

Elections and democratic practices have taken place across societies for centuries in various and distinct forms. As time passes, the world evolves, likewise do the systems we exist in. Elections within the past may have only impacted political leadership and governance structures, although nowadays they have a much more prominent impact on various aspects of society, more specifically: the economy. How? This is the question that's targeted within this empirical paper. The paper focuses on the relationship held between election cycles and fiscal decisions, where it solely focuses upon Germany and the United States to incorporate a comparative element between parliamentary and presidential systems.

Before the paper delves into its central research question, it is crucial to establish the relevant context. Investopedia refers to fiscal policy as the "use of government spending and tax policies to influence economic conditions, especially macroeconomic conditions, which include aggregate demand for goods and services, employment, inflation, and economic growth" (Hayes, 2025). Various papers find that governments and political parties utilise fiscal policy strategies to win over elections and public support, especially as fiscal policy is more visible and immediate than monetary policy which are actions by a country's central bank to control the money supply and credit (Bank of England, 2024).

Fiscal policy is closely related to countries' elections due to a notion called political budget cycles. Political budget cycles refer to governments strategically and directly manipulating fiscal policy, often in the form of reducing taxes or increasing public expenditure by incumbent governments just before elections to potentially increase chances of re-election (Fatmawati et al., 2025). Governments may make these calculated changes to create favourable short term economic conditions - which could result into long term pain - to showcase to voters that their current government is competent and working in alignment with their personal preferences (de Haan et al., 2025). For instance this can be identified via targeted spending and revenue shifts which seek to increase public investments such as infrastructure, transportation, etc, to create the impression of an economic boom among citizens (Horton & El-Ganainy, 2025), or through revenue manipulation which lowers visible taxes while sometimes increasing less visible fees to ensure the government maintains a sustained revenue source (Richter, 2024), another notable procedure would be pension funds, where governments involve them in bond investments or off-budget financial support to gain additional funding before elections without increasing taxes or public acknowledgment (Wang & Bohn, 2019).

This research distinguishes itself from similar papers as it focuses on how the specific institutional design of governments, presidential vs parliamentary, affects its fiscal behaviour. Presidential systems are forms of government where a president serves as both the head of state and government and is directly elected by the people or a special body for a fixed term, a clear separation between powers is also present between executive, legislative and judicial branches (Horton & El-Ganainy, 2025). Similarly, parliamentary systems are also democratic systems, although they feature various branches who work collaboratively together, where voters have the ability to elect parliament members and the majority party of parliament elects the head of government which can be removed at any given time (EBSCO, 2025). These two government structure variations result in very diverse fiscal policy processes in periods leading up to elections. The paper utilises the United States as its presidential system and Germany as its parliamentary one.

The field of political economy contains many studies that examine how election cycles affect policy decisions. Many papers find that prior to elections government spending increases prior to elections alongside budget deficits especially when there are fewer fiscal restraints (Ebeke, 2026). Similarly, this paper investigates how governments spend money prior to elections; higher public spending to appeal to voters (Schuknecht, 2000). However, much of the existing literature focuses solely on single country analysis, relies on annual data, focuses on development levels or budget deficits as data. As a result, fewer studies isolate how differences in government structures shape fiscal policy behaviour during election periods, significantly limiting our understanding of the role political institutions play in election driven fiscal decisions. Given the lack of existing research, this study investigates the relationship between election cycles and short-term government consumption expenditure (as a percentage of GDP) in the four quarters (12 months) prior national elections in the United States and Germany.

This paper hypothesizes that election cycles are associated with larger increases in government consumption expenditure in parliamentary systems than in presidential systems as parliamentary governments are supported by the legislature and are able to adjust government consumption more easily than in presidential systems where the power is separate between the executive and legislature, which could ultimately constrain fiscal decisions.

The following sections of the paper will first review existing literature surrounding the political budget cycle, fiscal policy manipulation and institutional constraints before outlining the methodology and data used within the analysis. The paper will then present and compare the fiscal trends observed in the United States and Germany during election and non-election years, which is followed by a discussion on the findings and the role institutional differences may play in shaping fiscal behaviour in presidential and parliamentary systems.

2. Literature Review

Before the paper begins to analyse key metrics regarding the impact election cycles have upon fiscal policy decisions, it is essential to examine existing academic literature which focuses on political budget cycles and election related fiscal behaviour. This review is structured around key themes in the literature:

Political Budget Cycles

Shi and Svensson (2006) examine whether fiscal policy systematically changes during election years across various countries (85) and whether those changes differ between developed and currently developing countries. The paper finds that fiscal deficits are about 1% higher in election years than in other years, and that these effects are larger in developing countries than developed ones. Moreover, it also suggests that differences in political incentives and overall institutional environments help explain why some countries show stronger cycles than others. Klomp and de Haan (2012) questions whether political budget cycles really exist, the paper investigates this across a panel of 70 democratic countries. The findings highlight that fiscal policy does indeed change near election times in many countries, although it also suggests that its long-term effects are not consistent across all countries. Furthermore, it identifies that the short-term effects vary systematically amongst the different countries due to their diverse characteristics, such as political system, transparency, development, etc. This paper then led to the creation of another one by Klomp & de Haan (2012), which instead focuses on the impact political budget cycles have on election outcomes. Through a semi - pooled panel model the paper highlights that in countries where elections do occur, democracies, significant changes in fiscal policy can be identified. Moreover, these policy changes are also often linked to a positive effect on votes, which ultimately suggests that fiscal policy changes may provide incumbents with certain amounts of electoral advantage, although this effect is relatively small.

Fiscal Policy Manipulation

Murtinu et al. (2021) perform an empirical test to examine if adaptations made to fiscal policy are not solely based on economics, but also to signal competence to voters and increase their chances of winning. The dataset covers OECD countries from 1999 - 2008 and uses regression analysis to find that parties loosen fiscal policy to manipulate voters (bigger deficits) in systems where voters don't understand economics well and politics is polarized. Moldovan et al. (2013) analyse Romanian fiscal policy from 1999 - 2013 to test whether election timing is correlated with changes in fiscal policy. It focuses on various forms of data such as government revenue, government expenditures, voter turnout and many more. Their results suggest that there are significant correlations between election timing and fiscal policy variables, additionally the paper highlights that politicians manipulate fiscal policy likely to appear more to voters. Alt and Lassen (2006) investigate whether election cycles are associated with changes in fiscal balance, such as budget surplus or deficit, they also examine how institutional factors like fiscal transparency and political polarization affect these changes. The paper finds that electoral cycles do affect fiscal balance in advanced countries. Moreso, it finds that in countries with higher fiscal transparency, fiscal changes aren't significant, it also highlights that in more politically polarized countries (where parties are farther apart), election cycles tend to be larger.

Institutional Constraints

Veiga, Efthyvoulou and Morozumi (in *The Oxford Handbook of Time and Politics*) provide a comprehensive review of political budget cycle research, the chapter reviews past research and literature and focuses on what causes fiscal policy to change around elections, and what specific conditions, such as political, institutional, etc, strengthen or weaken those cycles. The key takeaways from this paper are that PBC's are not universal, and instead happen in certain contexts, such as the degree of political competition, how strong institutional checks and balances are, the level of voter information and whether elections are unpredictable, highlighting that a country's political institutions do change whether election cycles cause fiscal manipulation. Wyplosz (2005) discusses how fiscal policy outcomes are shaped by institutional structures and rules. It finds that fiscal institutions are necessary to help control budgets as rules alone are often too weak, too strong or even ignored. Persson (2002) focuses on studying how different institutional structures affect fiscal outcomes such as government spending and public debt. The paper finds that political institutions actively influence fiscal policy, it suggests that electoral rules affect how politicians respond to voters and regime types affect how much influence the executive branch of the institution has over budgets, which explains why fiscal policy decisions vary across different countries.

Germany vs US differences

Belkin and Ahearn (2009) provide institutional background on Germany and why its such a major global economic power, while also analysing the economic relationship between Germany and the United States. It finds that Germany has an export - driven economy, which made up around 47% of its economy in 2008, which is far higher than the US's. Additionally, it

also finds that Germany faces strong budget constraints due to social spending, it also highlights the new budget rules passed by Germany which significantly limit its fiscal decisions. Potrafke (2017) reviews research on how government ideology influences economic and fiscal policy in the United States. The review highlights that different partisan ideologies affect their fiscal and economic policies, where left - leaning governments favour higher social spending, whereas right - leaning governments focus on tax cuts. Additionally, the paper finds that polarization amplifies the differences, which ultimately shapes how these different systems compete against another during electoral periods. Ehrmann and Cour-Thimann (2021) examine whether EU fiscal rules support or hinder the government's ability to implement counter-cyclical policy, a government or central bank strategy that moves in the opposite direction of the economy to keep it stable (Worrall, 2025). Using a large cross country dataset, it finds that discretionary fiscal policy is often pro-cyclical, meaning that governments tend to tighten in downturns and loosen in large booms. It also highlights that compliance with EU rules is often linked to more counter cyclical behaviour. The paper ultimately suggests that not only implementing rules but following them improves fiscal outcomes. Dellinger (1996) explains the conditional rules and principles that govern how power is divided in the US between the president and the congress. It explains that this division is put in place to prevent one branch from expanding its authority into other functions and to protect liberty.

Overall, this literature review has analysed various papers which are essential to understand the background and framework of this paper. Generally, the literature suggests that election cycles can influence fiscal policy outcomes through political budget cycles, especially through changes in government consumption. Additionally, many studies suggest that the size and magnitude of the fiscal manipulation depends on various institutional factors such as transparency, fiscal constraints and political polarization. This indicates that the effects elections have on countries are not universal but instead vary across different systems. While all these papers provide valuable research, none of them focus on a direct comparison between the United States and Germany. Therefore, this study contributes to the political economy literature as it investigates the difference in fiscal policy incentives in the United States and Germany, two advanced democracies with diverse institutional systems.

3. Methodology

As briefly stated above, this paper utilises a comparative case study approach between Germany and the United States, these were deliberate choices as both countries are highly developed economies and key global actors, however they hold different forms of governments. Making them ideal countries to focus on for this paper as they help control for economic development and fiscal capacity, reducing any potential bias during the analysis. Germany possesses Europe's largest economy as it accounts for 24% of its GDP and third in the whole world where it operates in a social market economy which combines capitalism with social policies (International Trade Administration, 2023). The country currently faces economic stagnation due to various variables such as pandemic recovery, weak global demand, etc,

(Grömling, 2025) yet it holds a potential rebound of 1.3% in 2026 as it shifts to an expansionary stance (Yermolenko, 2025). Germany's government specifically operates under a federal parliamentary republic system which holds a chancellor as the head of the government and a president as the head of state (George Hall Kirby & Hamerow, 2019), where the legislative and executive branches of the government decide fiscal policy. Instead, the US holds the world's largest economy (Jones & Uteuova, 2025) which undergoes heavy influence by government regulation, following the decisions made by Congress and the President (Express, n.d.), although the US economy currently faces a high debt to GDP ratio (Trading Economics, 2019) which was expected to reach 125.40% by the end of 2025 due to decades of persistent budget deficits. Contrary to Germany, the US operates as a constitutional federal republic with three different branches to prevent imbalance in power, although fiscal policy is managed by the President and Congress (U.S. Mission New Zealand, 2022).

Moreover, secondary quantitative data will be utilised throughout the paper, as it enables a systematic comparison between different countries and time periods. Whereas, the primary fiscal indicator utilised will be government consumption expenditure (as a % of GDP), this indicator measures the value of goods and services produced and purchased by the government, excluding its military investment, which globally ranges between 15 - 20% of a country's GDP (BEA, 2024). This indicator is ideal for this specific analysis as it reflects direct government decisions as government expenditure can include public wages, spending on public services, etc, (Tutor2u, 2024) and as it responds quickly to election incentives unlike tax reforms or debt levels which take much longer. To measure the difference in government expenditure between the countries the paper will compare election years to non-election years to identify changes in government expenditure, which reflects short term fiscal policy changes. Additionally, as the government consumption will be measured as a percentage of GDP the comparison won't focus on the sheer amounts of money spent by the governments but rather by how big their spending is compared to their country's economy. As the analysis will focus on the 4 quarters leading up to elections as existing literature identifies significant changes in fiscal policy around these times the data will be presented using a line graph as it enables a clear visual display of pre-election trends and the differences among the two countries. The data will be sourced from the World Bank, which is a reputable international institution which provides standardised and comparable fiscal data across countries, which in turn increase the overall reliability and validity of the analysis, more specifically for this analysis the paper will utilise the World Bank’s World Development Indicator.

However, it is all too easy to ignore the limitations this methodology holds. As the paper only focuses on two wealthy and well-developed countries/democracies, the claims made may not be applicable to other democracies which exist around the world due to the conditions they exist in. Secondly, government consumption expenditure may not fully reflect electoral manipulation or incentives made by political institutions, thus the dataset chosen may not provide us with all the necessary data. Lastly, we’re focusing on the 4 terms occurring prior to elections, although some governments begin electoral incentives a year before elections,

therefore the paper is overlooking earlier policy adjustments. As a result, it is essential to highlight that the findings made from this paper are specific to the selected countries, data used and timeframe applied.

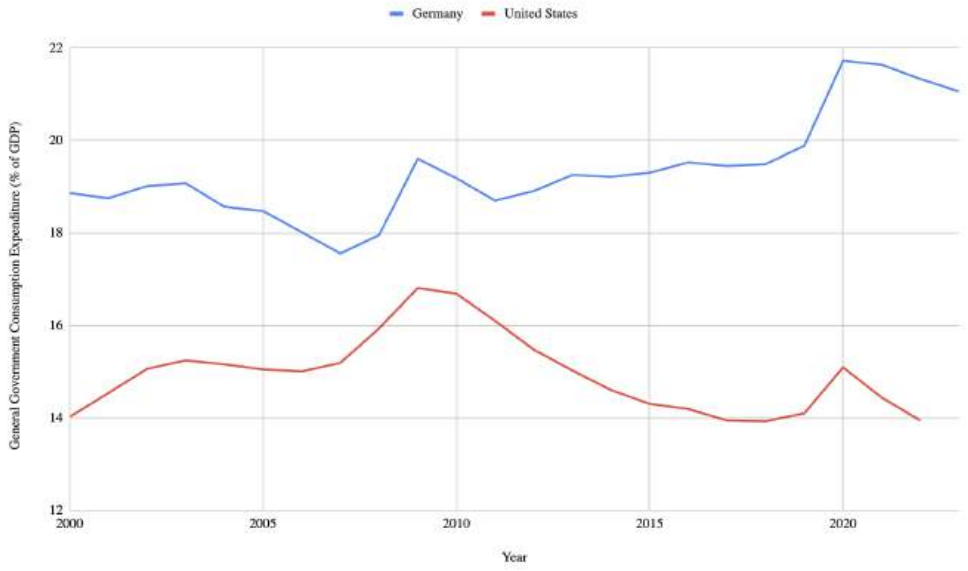
4. Data & Results

The study utilises secondary quantitative data collected from the World Bank, which is a reputable institution utilised in various research papers. The source was able to provide me with comparable data regarding government expenditure (as a % of GDP) for both Germany and the United States. To examine the relationship between elections and fiscal policy, the paper focuses on the 4 quarters (12 months) prior to federal elections in Germany and presidential elections in the United States. As Political Budget Cycle literature suggests fiscal activity is often the strongest in the immediate pre - election period. These election periods were then compared with non - election years in order to identify whether government expenditure tended to increase around election times.

Figure 1

Variable	Description
Dependent	Government consumption expenditure (as a % of GDP)
Independent	Proximity to elections (pre-election vs non-election period)
Control	Non-election quarters

Annual general government final consumption expenditure (% of GDP) in Germany and the United States (2000–2023).



(World Bank, 2021)

Figure 1 represents a line graph that showcases the data collected from the World Bank's World Development Indicator, the graph provides a clear visual comparison between general government consumption expenditure (as a % of GDP) in the US and Germany. The data indicates a stark difference between the amount of consumption expenditure in Germany compared to the US, where Germany possesses a much larger baseline. For instance, the US's highest expenditure sits at around 17%, whereas Germany's is 22%.

It is also crucial to highlight 2 key points in the graph where the percentage of government consumption greatly increased, around 2008 and 2021. This was due to the Global Financial Crisis of 2008, where it was necessary that governments stabilised collapsing economies, this was done by increasing spending on public administration, social support and emergency economic measures. Similarly, in between 2020 - 2021 Covid-19 quickly hit economies which required substantial government intervention through healthcare spending, emergency public services and crisis management and administration. These events highlight what a prominent role government expenditure holds not only to support citizens in times of distress but also to reconstruct economies piece by piece. However, these spikes also end up complicating the analysis, as the increases in expenditure during 2008 and 2020 were likely influenced by crises-related intervention, rather than electoral incentives alone. Therefore, the findings should be interpreted cautiously as some increases may not directly relate to the paper's central focus.

Figure 2
Comparison of Government Consumption (% of GDP) in Election vs Non-Election Years

Country	**Average non-election years**	**Average election years**	**Increase during election years**
Germany	19.30% of GDP	20.21% of GDP	0.91%
United States	15.46% of GDP	16.38% of GDP	0.92%

The paper has now identified that Germany - a parliamentary system - spends greater amounts of money on government expenditure than the United States - a presidential system - although it's yet to focus on the exact question the paper is asking: What is the relationship between election cycles and short-term fiscal policy decisions in presidential versus parliamentary political systems? The table above contains the result of various calculations used to find the average government consumption in both Germany and US during election and non - election years. The dataset extends between 2000 - 2023, which comprises a total of 6 federal and presidential elections in the US and Germany, allowing for a balanced comparison. The findings from the table suggest that in both countries, government consumption increases in election years by an estimated 0.9%. This suggests that modest fiscal expansions occur around election periods, regardless of the political system type, which remains consistent with the Political Budget Cycle theory. However, Germany consistently maintains a higher baseline level than the United States.

5. Discussion

Overall, the results of the paper suggest that there appears to be an increase in fiscal policy around election times, more precisely in government consumption expenditure. These findings align with the political budget cycle theory, which suggests that governments may increase visible spending prior to elections to improve re-election prospects. Increased fiscal spending could improve re-election prospects by signalling economic competence to voters through visible public services and government investment. Moreover, fiscal policy may also allow governments to target specific groups of voters easily by providing them with new benefits, such as public employee raises or social projects. Fiscal policy may also link to tax reductions prior to elections, which would increase voters' disposable income and improve voter perceptions of the current incumbent. It's also key to highlight *why* governments choose fiscal policy instead of other forms of policy such as monetary policy. Fiscal policy may be strategically preferred as it has the potential to directly and immediately impact citizens through visible legislative changes to taxes, public spending and transfer payments, allowing governments to influence voter perceptions more efficiently during election periods.

Moreover, the results from the data collected also suggest a greater baseline of government expenditure in parliamentary systems than in presidential ones. There are various reasons for this difference:

- Germany has a larger public sector role especially due to its social market economy, whereas the US relies much more on private provision, which denotes the delivery of goods or services by non-government entities, meaning that whole government involvement in the US is less visible to citizens.
- The executive and legislature branches of parliamentary systems tend to be more closely aligned in their beliefs and policies, thus budgets may pass more easily than in presidential systems, as the executive and legislative are separate, meaning that budgets can easily become blocked or delayed, which could end up reducing government activity.
- Many parliamentary systems operate through coalitions to pass budgets or policies as they need the majority of the government to agree with their notions, due to this it is key that all the parties involved compromise to fit both their own needs alongside the ones of others. As a result, parliamentary governments may be more likely to maintain or expand public spending to pass budgets and avoid political instability. The US works very differently to this as parties don't collapse if everyone doesn't agree, instead solely the president and congress have authority regarding budgets, which could increase the chances of gridlock and ultimately leads to lower economic productivity.
- Cultural ideals may also play into this division as parliamentary systems (especially in Europe) are often expected to possess stronger welfare and service traditions and larger public administration sectors, whereas presidential systems governments are often linked to limited government involvement and greater reliance on private provision of services.

Ultimately, the results of the data confirm the thesis statement that was made at the very beginning of the paper. The thesis remarked that 'election cycles are associated with larger increases in government consumption expenditure in parliamentary systems than in presidential systems'. The findings remain consistent with this claim, as government spending did increase during election periods and as this increase was much more evident and consistent in the parliamentary system analysed - Germany. This outcome highlights the importance of institutional structures in shaping fiscal behaviour and suggests that electoral incentives may operate differently in diverse political systems. Although it is also key that it is acknowledged that the 0.9% increase in government consumption expenditure in both institutional systems might be an upper bound estimate because of these coinciding global crises: 2008 economic crisis and Covid 19.

6. Limitations

It is crucial to highlight the limitations this paper possesses. Firstly, this study relies on aggregate government consumption, which denotes various strands of data that are all combined together into one number; thus the paper isn't able to truly capture the variations in spending across different policy areas. As a result, it's not possible for the study to determine whether fiscal policy increases refer to sectors such as public welfare or others such as infrastructure. Moreover, the period analysed includes major economic shocks such as the 2008 global financial crisis and the Covid 19 Pandemic, these events may have influenced government expenditure, regardless of elections, therefore they could've disrupted the overall data. In addition to that, the paper solely focuses on 2 countries, which are both advanced democracies, thus the findings made in this paper may not be generalised to other countries with different political systems, levels of development or financial institutions. Ultimately, even though the results highlight a clear association between election years and increased government spending, the analysis cannot establish a direct relationship between the two events, as other factors exist that may also influence fiscal decisions.

7. Conclusion

This paper set out to examine whether election cycles were associated with changes in fiscal policy (government consumption expenditure as a % of GDP) in two starkly different political systems: presidential and parliamentary. By comparing fiscal behaviour in the United States and Germany, the analysis sought to investigate whether different institutional structures and frameworks may respond differently to electoral incentives.

The findings of the paper suggest a clear association with increased government spending near election cycles in both presidential and parliamentary systems, where both systems averaged a 0.9% increase compared to their non-election year average. However, the analysis also revealed a substantial difference in baseline government expenditure between the two systems, as Germany had a much greater baseline number, around 4% higher than the United States. This supports the thesis that the parliamentary systems may be more susceptible to

election-driven fiscal expansion due to closer links between the executive and legislature branches of the institution and limits presidential systems hold, such as gridlock.

Furthermore, this paper contributes to the vast literature of political budget cycles, which refers to the tendency of incumbent governments to manipulate fiscal policy prior to elections. Firstly, the paper remains consistent with other conclusions made in pre-existing literature as it suggests that fiscal policy increases near election cycles. Although, it also contributes to the literature as it investigates the correlation between different political systems and fiscal policy, by highlighting the importance of political institutions in shaping fiscal outcomes, as different governments may interact differently due to their unique institutional constraints and governance structures which could ultimately influence their fiscal policy decisions and freedom.

Understanding these dynamics is crucial for policymakers, as increases in public spending during election periods can have long term effects on fiscal sustainability and overall economic stability. Especially as short-term electoral incentives may oppose long term economic goals in the future. Future research could build upon this study by analysing a wider range of countries or by examining clearer and more specific areas of government expenditure, such as welfare or infrastructure, as this would enable a clearer understanding on exactly how electoral incentives may influence fiscal decisions in different governments and countries.

Overall, this study concludes that political systems appear to play an important role in shaping the relationship between elections and fiscal policy. This appears to be largely influenced by the institutional context, limitations, and political pressures different governments face.

8. Acknowledgements

I would like to thank Ceyhun Elgin for his constant support and availability throughout my personal research journey and writing process, alongside Daryn Howland, who have both significantly shaped the outcome of this paper and my personal writing journey.

Works Cited

Alt, J. E., & Lassen, D. D. (2006). Transparency, political polarization, and political budget cycles in OECD countries. American Journal of Political Science, 50(3), 530–550. https://doi.org/10.1111/j.1540-5907.2006.00200.x

Bank of England. (2024, April 17). Monetary policy. https://www.bankofengland.co.uk/monetary-policy

de Haan, J., Ohnsorge, F., & Yu, S. (2025). Election-induced fiscal policy cycles in democratic and non-democratic emerging market and developing economies. Electoral Studies, 95, 102936. https://doi.org/10.1016/j.electstud.2025.102936

U.S. Bureau of Economic Analysis. (2024). Government consumption expenditures and gross investment. https://www.bea.gov/resources/methodologies/nipa-handbook/pdf/chapter-09.pdf

Ebeke, C. H., & Ölcer, D. (2013). Fiscal policy over the election cycle in low-income countries (IMF Working Paper No. 13/153). International Monetary Fund. https://doi.org/10.5089/9781475588514.001

Georgia Public Broadcasting. (n.d.). Econ Express fundamentals: Concept 12—Roles of government in the U.S. economy. https://www.gpb.org/education/econ-express/roles-government-us-economy

Fatmawati, S., Fitrady, A., & Widodo, T. (2025). Politically driven cycles in fiscal policy: Evidence from disaggregated budgets in middle-income countries. Economies, 13(6), 151. https://doi.org/10.3390/economies13060151

Gaddipati, S. (2019, June 7). The parliamentary system versus the presidential system. The Borgen Project. https://borgenproject.org/parliamentary-system-versus-presidential-system/

Kirby, G. H., & Hamerow, T. S. (2019). Germany: Government and society. In Encyclopaedia Britannica. https://www.britannica.com/place/Germany/Government-and-society

Grömling, M. (2025). The current economic situation in Germany in the context of previous crises. Intereconomics, 60(1), 55–59. https://www.intereconomics.eu/contents/year/2025/number/1/article/the-current-economic-situation-in-germany-in-the-context-of-previous-crises.html

Hayes, A. (2025, March 21). All about fiscal policy: What it is, why it matters, and examples. Investopedia. https://www.investopedia.com/terms/f/fiscalpolicy.asp

Horton, M., & El-Ganainy, A. (2025). Fiscal policy: Taking and giving away. International Monetary Fund. https://www.imf.org/en/Publications/fandd/issues/Series/Back-to-Basics/Fiscal-Policy

tutor2u. (2024). IB economics: Types of government expenditure. https://www.tutor2u.net/economics/reference/ib-economics-types-of-government-expenditures

Goodin, R. E. (2009). Institutional constraints on policy. Oxford Academic. https://academic.oup.com/edited-volume/28180/chapter-abstract/213073875

International Trade Administration. (2023, December 6). Germany—Market overview. https://www.trade.gov/country-commercial-guides/germany-market-overview

Jones, C., & Uteuova, A. (2025, December 28). The 2025 US economy—in charts: Rising prices, hiring slowdown, rollercoaster growth. The Guardian. https://www.theguardian.com/business/2025/dec/28/2025-us-economy-in-charts

Klomp, J., & de Haan, J. (2013). Do political budget cycles really exist? Applied Economics, 45(3), 329–341. https://doi.org/10.1080/00036846.2011.599787

Klomp, J., & de Haan, J. (2012). Political budget cycles and election outcomes. Public Choice, 157(1–2), 245–267. https://doi.org/10.1007/s11127-012-9943-y

Larch, M., Orseau, E., & van der Wielen, W. (2021). Do EU fiscal rules support or hinder counter-cyclical fiscal policy? Journal of International Money and Finance, 112, 102328. https://doi.org/10.1016/j.jimonfin.2020.102328

Dellinger, W. (2000). Appendix: The constitutional separation of powers between the president and Congress. Law and Contemporary Problems, 63(1–2), 513–571. https://scholarship.law.duke.edu/lcp/vol63/iss1/23/

Moldovan, N.-C., Hatmanu, M., & Lobonț, O.-R. (2014). Reflections on the Romanian fiscal policy-driven election manipulation phenomenon. Economic Computation and Economic Cybernetics Studies and Research, 48(2), 91–113. https://ecocyb.ase.ro/eng/Moldovan%20N.pdf

Murtinu, S., Piccirilli, G., & Sacchi, A. (2022). Rational inattention and politics: How parties use fiscal policies to manipulate voters. Public Choice, 190(3), 365–386. https://doi.org/10.1007/s11127-021-00940-8

EBSCO Information Services. (2025). Parliamentary system. https://www.ebsco.com/research-starters/history/parliamentary-system

Belkin, P., & Ahearn, R. J. (2009). [Insert the exact report title written on the first page of the PDF]. Congressional Research Service. https://apps.dtic.mil/sti/tr/pdf/ADA510975.pdf

Persson, T. (2002). Do political institutions shape economic policy? Econometrica, 70(3), 883–905. https://doi.org/10.1111/1468-0262.00313

Potrafke, N. (2017). Government ideology and economic policy-making in the United States—A survey. Public Choice, 174(1–2), 145–207. https://doi.org/10.1007/s11127-017-0491-3

Richter, J. (2024). The influence of fiscal policy on economic growth: A macroeconomic assessment. Journal of Economics and Economic Education Research, 25(4S), 1–3. https://www.abacademies.org/articles/the-influence-of-fiscal-policy-on-economic-growth-a-macroeconomic-assessment-17105.html

Saiegh, S. (n.d.). Policy differences among parliamentary and presidential systems. https://pages.ucsd.edu/~ssaiegh/oxford_pc_chapter.pdf

Schuknecht, L. (2000). Fiscal policy cycles and public expenditure in developing countries. Public Choice, 102(1–2), 113–128. https://doi.org/10.1023/A:1005026806998

Shi, M., & Svensson, J. (2006). Political budget cycles: Do they differ across countries and why? Journal of Public Economics, 90(8–9), 1367–1389. https://doi.org/10.1016/j.jpubeco.2005.09.009

Trading Economics. (n.d.). United States government debt to GDP. Retrieved January 7, 2026, from https://tradingeconomics.com/united-states/government-debt-to-gdp

U.S. Mission New Zealand. (2022, July 1). Branches of the U.S. government: How do they work? U.S. Embassy & Consulate in New Zealand, Cook Islands and Niue. https://nz.usembassy.gov/u-s-government/

Wang, X., & Bohn, F. (2019). Pension reserve fund, political budget cycles and fiscal illusion. European Journal of Political Economy, 56, 62–73. https://doi.org/10.1016/j.ejpoleco.2018.07.002

World Bank. (n.d.). General government final consumption expenditure (% of GDP)—Germany and United States. World Bank Open Data. Retrieved January 19, 2026, from https://data.worldbank.org/indicator/NE.CON.GOVT.ZS?end=2023&locations=DE-US&start=2000

Worrall, C. (2025, June 24). What is a "countercyclical" policy or program? How Medicaid protects people when the economy falters. SHADAC. https://www.shadac.org/news/countercyclical-fiscal-policy-program-medicaid

Wyplosz, C. (2005). Fiscal policy: Institutions versus rules. National Institute Economic Review, 191(1), 64–78. https://doi.org/10.1177/0027950105052661

Yermolenko, H. (2025, September 5). Economic institutes have lowered their GDP growth forecasts for Germany for 2025–2026. GMK Center. https://gmk.center/en/news/economic-institutes-have-lowered-their-gdp-growth-forecasts-for-germany-for-2025-2026/

Education Quality and Youth Unemployment: Evidence from a 30-Year Cross-Country Panel By Samit Mandhare

Introduction

Youth unemployment represents one of the most consequential challenges facing economies throughout the world *(ILO,2023)*. Beyond its direct economic costs, it carries detrimental social consequences including deteriorating mental health, eroding social cohesion, deeply rooted inequality, and the long-run decrease of individual productivity trajectory (*Bell & Blanchflower, 2011; Paul & Moser, 2009*). Even though macroeconomic fluctuations explain the short-run movements in unemployment, they leave lapses in cross-country variations. During consistent expansions, an array of countries maintain persistently high unemployment rates, pointing towards structural forces rather than cyclical: specifically towards the mismatch between the skills workers possess and skills that the economy demands(*Pissarides,2000; Blanchard & Wolfers,2000*). Structural unemployment is one type of natural unemployment that cannot be resolved when aggregate demand recovers, because its origin lies in the misalignment between the labour market requirements and the fundamental labour supply characteristics (*Lilien,1982; Pissarides,2000*). This misalignment has grown increasingly consequential as technological advancement and the rise of artificial intelligence continues to remodel the labour market. They automate routine tasks, displace low-skilled occupations, and generate demand for higher-order cognitive competencies that many educational systems have been slow to cultivate (*Autor,Levy & Murnane,2003; Acemoglu & Restrepo,2018*). Among all the age-groups, the youth bear the exposure to these structural changes, making the skill gap between what educational systems produce and what labour market require more pertinent to address than ever before.

The detrimental consequences of persisted youth unemployment extends far beyond the direct income loss. Early youth unemployment generates wage scars and it raises the probability of subsequent spells of unemployment, reinforcing the fact that the labour market exclusion at a younger stage tends to make itself worse over time rather than self-correcting. These costs lead into higher societal costs that include higher poverty rates, weaker institutional trust, and diminished productive capacity across various sectors (*Scarpetta et al.,2010*). From an efficiency perspective, the households are subject to market failure because they often cannot make investments in education on their own due to information failure about returns to different skills, and credit constraints (*Carneiro & Heckman,2002*). These market failures mean that without interventions, economies will underprovide the skills that the labour market requires, making educational policy a vital tool for addressing youth unemployment and skill mismatch.

Despite the empirical and theoretical case for this relationship, the evidence linking measured educational quality directly to youth unemployment outcomes remains quite underdeveloped. Existing studies have either looked at how years of schooling relate to youth unemployment or how individual skills affect each individual wages, but few studies have directly examined whether the quality of education affects the overall youth unemployment rate

across time. This paper addresses that gap. Using a dataset of over 30 years, it examines how educational quality measured through indicators of mathematics proficiency affects youth unemployment, alongside the role of vocational training, which is included to account for the practical and occupation-specific dimension of skill provision that general measures do not capture. A multivariate regression model is used alongside the controls for LogGDP per capita, urbanisation, service sector share, and net migration, so that among countries with similar demographics and macroeconomic features, we can isolate what a marginal improvement in education quality does to youth unemployment. The hypothesis states that countries with high educational quality are associated with lower youth unemployment, reflecting the role of education quality in reducing the labour market mismatch.

More broadly, the paper highlights a fundamental question in the economics of education and labour markets: does the quality of what students learn, rather than simply how long they spend in school translate into better labour market outcomes? If educational quality is in fact a structural determinant of youth unemployment, then the policy implications extends beyond the classroom, connoting that countries seeking to reduce youth unemployment must look towards whether their education systems are providing young people with skills that the labour market demands.

Methods

To analyse the empirical relationship between youth unemployment and education quality across countries, this paper employs Ordinary Least Squares (OLS) regression. The dependent variable is youth unemployment, defined as the share of the total labour force aged 15–24 that is unemployed, sourced from the World Bank's World Development Indicators. As shown in Figure 1, which maps youth unemployment rates across countries in 2024, there is striking variation in youth labour market outcomes globally, rates are very high across parts of the Middle East and North Africa, while East Asia and parts of Northern Europe exhibit very low rates. This raises an important question: why do some countries have persistently high youth unemployment while others maintain very low rates, even after accounting for differences in overall unemployment? This paper argues that education quality is a key part of the explanation.

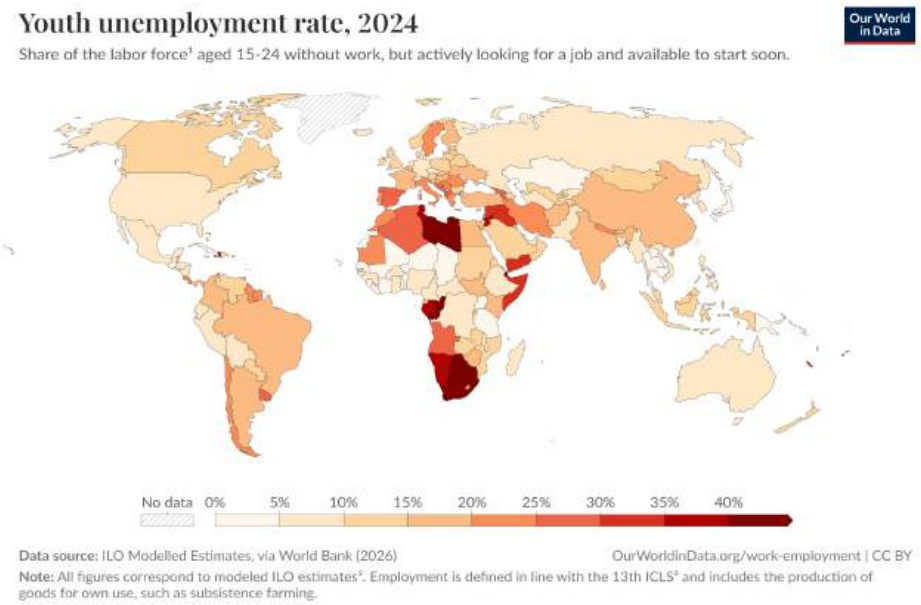


Figure 1.

The main independent variable is education quality, measured through three distinct dimensions sourced from the UNESCO Institute for Statistics. First, the share of lower secondary students achieving minimum mathematics proficiency, drawn from PISA examinations conducted across countries, captures the cognitive skill base of students at the lower secondary level. Mathematics proficiency at the lower secondary level is preferred over reading proficiency for two reasons: it has the greatest number of observations among the available UNESCO indicators at 528, and it is highly correlated with alternative measures including mathematics proficiency at the primary level, reading proficiency at the lower secondary level, and reading proficiency at the primary level, meaning it captures the same underlying dimension of education quality while providing the broadest country coverage. Second, youth literacy rate, defined as the percentage of the population aged 15–24 that can read and write, is included as a complementary measure of foundational cognitive skill attainment. Third, vocational training enrollment as a share of secondary students is included to capture the occupation-specific dimension of skill provision that general cognitive measures alone do not reflect. As shown in Table 1, there is substantial variation across all three education quality measures, mathematics proficiency ranges from 0.73% in Côte d'Ivoire in 2023 to 95.34% in Singapore in 1995, youth literacy from 19.41% to 100%, and vocational training from 0.01% in Barbados in 2005 to 41.69% in Vanuatu in 2023, with some countries having very high enrollment in vocational programs and others very low, underscoring the cross-country variation in education systems that this paper seeks to exploit.

Variable	Mean	SD	Min	Max	N
Total Unemployment	7.71	5.76	0.1	38.8	4368
Youth Unemployment	16.49	11.93	0.3	75	4368
NEET	20.51	9.76	2.97	55.62	3166
Vocational Training	9.67	8.42	0.01	41.69	2064
Youth Literacy Rate	90.69	14.6	19.41	100	1071
MATH LOWERSEC	59.04	25.27	0.73	95.34	528
MATH PRIMARY	50.08	28.38	1.29	97.8	373
READING LOWERSEC	67.46	19.69	5.04	95.19	482
READING PRIMARY	66.67	31.86	1.8	99.55	298
GDP per capita (constant 2015 USD)	13287.98	18406.04	211.27	112417.9	4307
Services (% of value added in GDP)	54.25	11.82	10.86	96.2	4139
Urban Population Share	57.89	23.21	7.21	100	4372
Net Migration (1000s)	-0.46	219.39	-5699.44	1889.13	4372

Table 1: Summary Statistics, Dependent Variable and Education Quality Measures

This data yields a panel dataset covering 183 countries over the period 1995–2024, with 1995chosen as the starting point as it marks the earliest year for which mathematics proficiency data are available. To isolate the education quality channel from broader macroeconomic and demographic influences, four control variables are included, all sourced from the World Bank's World Development Indicators: GDP per capita, urbanisation, service sector share, and net migration. GDP per capita is controlled because wealthier nations can afford higher quality education systems and simultaneously offer more diverse job opportunities, meaning any observed relationship between education quality and youth unemployment may otherwise simply reflect underlying differences in national wealth. Urbanisation is included because urban areas tend to concentrate higher quality educational institutions and denser labour markets, meaning higher urbanisation may independently drive both better education outcomes and lower youth unemployment. Service sector share is included as a proxy for the degree of economic formalisation, economies with larger formal service sectors generate stronger demand for certified skills. Finally, net migration is controlled because countries experiencing high net out-migration may see their educated workforce leave, distorting measured youth unemployment independently of domestic education quality. As shown in Table 1, these control variables exhibit substantial cross-country variation, with GDP per capita ranging from $211.27 to $112,417.90, urbanisation from 7.21% to 100%, service sector share from 10.86% to 96.2%, and net migration from −5,699.44 to 1,889.13 thousand. Each of the three education quality measures is examined across three progressively controlled specifications, without controls, with log GDP per capita and its squared term only, and with the full set of controls, yielding nine specifications in total, as set out below:

The nine regression equations estimated are as follows:

Mathematics Proficiency:

Equation (i): Mathematics Proficiency without Controls:

$$YU(i,\ t+k) = \beta 0 + \beta 1 \cdot MathLS(i,t) + \alpha t + \varepsilon(i,t)$$

Equation (ii) :Partial Controls:

$$YU(i,\ t+k) = \beta 0 + \beta 1 \cdot MathLS(i,t) + \beta 2 \cdot ln(GDPpc(i,t)) + \beta 3 \cdot ln(GDPpc(i,t))^2 + \alpha t + \varepsilon(i,t)$$

Equation (iii): Mathematics Proficiency with Controls:

$$YU(i,\ t+k) = \beta 0 + \beta 1 \cdot MathLS(i,t) + \beta 2 \cdot ln(GDPpc(i,t)) + \beta 3 \cdot ln(GDPpc(i,t))^2 + \beta 4 \cdot Urban(i,t) + \beta 5 \cdot Service(i,t) + \beta 6 \cdot Mig(i,t) + \alpha(t) + \varepsilon(i,t)$$

Youth Literacy Rate:

Equation (iv): Youth Literacy Rate without Controls:

$$YU(i,\ t+k) = \beta 0 + \beta 1 \cdot LitRate(i,t) + \alpha t + \varepsilon(i,t)$$

Equation (v): Partial Controls:

$$YU(i,\ t+k) = \beta 0 + \beta 1 \cdot LitRate(i,t) + \beta 2 \cdot ln(GDPpc(i,t)) + \beta 3 \cdot ln(GDPpc(i,t))^2 + \alpha t + \varepsilon(i,t)$$

Equation (vi): Youth Literacy Rate with Controls:

$$YU(i,\ t+k) = \beta 0 + \beta 1 \cdot LitRate(i,t) + \beta 2 \cdot ln(GDPpc(i,t)) + \beta 3 \cdot ln(GDPpc(i,t))^2 + \beta 4 \cdot Urban(i,t) + \beta 5 \cdot Service(i,t) + \beta 6 \cdot Mig(i,t) + \alpha t + \varepsilon(i,t)$$

Vocational Training:

Equation (vii): Vocational Training without Controls:

$$YU(i,\ t+k) = \beta 0 + \beta 1 \cdot VocTrain(i,t) + \alpha t + \varepsilon(i,t)$$

Equation (viii): Partial Controls:

$$YU(i,\ t+k) = \beta 0 + \beta 1 \cdot VocTrain(i,t) + \beta 2 \cdot ln(GDPpc(i,t)) + \beta 3 \cdot ln(GDPpc(i,t))^2 + \alpha t + \varepsilon(i,t)$$

Equation(ix): Vocational Training with Controls:

$$YU(i,\ t+k) = \beta 0 + \beta 1 \cdot VocTrain(i,t) + \beta 2 \cdot ln(GDPpc(i,t)) + \beta 3 \cdot ln(GDPpc(i,t))^2 \beta 4 \cdot Urban(i,t) + \beta 5 \cdot Service(i,t) + \beta 6 \cdot Mig(i,t) + \alpha t + \varepsilon(i,t)$$

Where YU(i, t+k) is the youth unemployment rate in country i at time t+k; MathLS(i,t), LitRate(i,t), and V ocTrain(i,t) refer to mathematics proficiency at the lower secondary level, youth literacy rate, and vocational training enrollment respectively, each measured in country at time t; ln(GDPpc(i,t)) is the natural logarithm of GDP per capita; ln(GDPpc(i,t))² is its squared term; Urban(i,t) is the urban population share; Service(i,t) is the service sector share of GDP; Mig(i,t) is net migration expressed in thousands; αt represents year fixed effects; and ε(i,t) is the error term. β1 is the coefficient of primary interest, capturing the estimated effect of a one-unit increase in the relevant education quality measure on youth unemployment k years later.

Literature Review

Youth unemployment rates run two to three times adult rates across every country and time period for which data exists, a ratio that has remained stable across business cycles, highlighting that excess youth unemployment cannot be attributed to cyclical fluctuation alone (*ILO, 2023*). Even during persistent expansions, countries maintain high unemployment rates, pointing towards structural forces, specifically between the skills workers possess and those the economy demands (*Pissarides,2000; Blanchard & Wolfers,2000*). In 2023, youth unemployment stood at a staggering 13% against a total unemployment rate of 5%, with particularly high concentrations in Sub-Saharan Africa, Southern Europe, and South Asia (*ILO, 2023; Eurostat, 2023*). These discrepancies continue through economic cycles and vary enormously across countries with similar macroeconomic conditions. Early unemployment compounds this issue by generating wage penalties that can last up to two decades and significantly increasing the probability of future unemployment spells, suggesting that labour market exclusion at a young age tends to worsen over time rather than resolve itself (*Bell & Blanchflower, 2011; Gregg, 2001*).

Structural unemployment arises when the labour forces lack the technical, cognitive, and occupational capabilities that employers require, and the skills that workers develop are rooted in the education and vocational training systems through which they pass (*Autor, Levy & Murname, 2003; Hanushek & Woessmann, 2015*). The dominant empirical tradition has used years of schooling as the primary measure of human capital, implicitly assuming a year of schooling produces the same cognitive gain regardless of where it is delivered (*Mincer, 1974*). This assumption has been directly challenged where cross-country evidence consistently showed that measured cognitive skills, rather than school attainment, are the key primary drivers of both individual market labour outcomes and aggregate economic growth (*Hanushek & Woessmann, 2008,2015*). Using the data from PISA and TIMSS across 50 countries over the period 1960-2000, a one-standard deviation increase in test scores is associated with approximately two percentage points higher annual GDP per capita growth, and once cognitive skills are accounted for, years of schooling becomes insignificant, suggesting that returns previously attributed to schooling were in fact returns to the skills that schooling produces (*Hanushek & Woessmann, 2015*).

Educational quality, moreover, differs across countries, far more than enrollment rate suggests, and this variation has been shown to map closely onto cross-country differences in structural unemployment outcomes (*Hanushek & Woessmann, 2015*). At least two-thirds of the world's youth fail to reach basic skill levels, and 62% of currently enrolled students fall below the PISA level 1 threshold (*Gust et al.,2022*). The economic costs of these deficits are substantial as achieving universal basic skills would generate a discounted gain of over $700 trillion in world GDP over the remainder of the century, with the gain from improving learning quality among currently enrolled students found to be more than twice as large as the gain from expanding enrollment alone (*Gust et al., 2022).*

At the Micro level, a one-standard deviation increase in numeracy skills is associated with an average 18% wage increase among prime-age workers across 23 countries, with cognitive skills also found to raise employment probability by approximately 8.3 percentage points (*Hanushek et al., 2015*). When the non-employed are incorporated into the analysis, estimated skill returns nearly double. Estimates based on early-career workers are further shown to understate the lifetime returns to skills by approximately one quarter, implying that the cost of skill deficits for youth accumulates over the career (*Hanushek et al., 2015*). Approximately 45% of aggregate skill returns are found to operate through occupational sorting, with lower-skilled workers sorted into lower-paying, more precarious occupations (*Hanushek et al., 2015*).

Results and Discussion

This section presents the results of the nine OLS regression specifications outlined in Section 4. For each specification, the coefficient on the education quality variable of interest, its t-statistics, year fixed effects, and the R-squared are reported, alongside the coefficients on the control variables where Included. Table 2 presents the results for youth literacy rates regressed on youth unemployment at a 15-year lag, both without and with controls.

Youth Literacy Rate:

Youth Literacy Rate and Youth Unemployment (15 Years Lag)

	Dependent variable:		
	Youth Unemployment (t+15)		
	(1)	(2)	(3)
Youth Literacy Rate (t)	0.245***	0.149***	0.151***
	(0.036)	(0.053)	(0.053)
Log GDP Per Capita (t)		19.058***	15.792**
		(6.505)	(7.569)
Log GDP Per Capita-Sq (t)		-1.089***	-0.800*
		(0.383)	(0.448)
Service Share (t)			0.082
			(0.079)
Urban Population (t)			-0.093*
			(0.055)
Net Migration 1000s (t)			0.010***
			(0.004)
Year FE	Yes	Yes	Yes
Observations	335	328	309
R^2	0.164	0.191	0.237
Note:	Standard errors in parentheses. *p<0.1; **p<0.05; ***p<0.01		

Table 2.

Across all three specifications, the coefficient on youth literacy rate is positive and statistically significant, a finding that runs counter to the hypothesis that higher education quality

reduces youth unemployment. In the baseline specification (Column 1), the coefficient is 0.245, significant at the 1% level, indicating that a one percentage point increase in youth literacy rate is associated with a 0.245 percentage point increase in youth unemployment 15 years later, with an R-square of 0.164. In Column 2, the addition of log GDP per capita and its squared term reduces the coefficient to 0.149, remaining significant at the 1% level, with log GDP per capita entering positively and significantly and its squared term negatively and significantly, consistent with the inverse U-shaped relationship between income and youth unemployment documented in Appendix Figure A3. The R-square rises to 0.191.

In the full controls specification (Column 3), the coefficient on youth literacy rate rises slightly to 0.151, remaining significant at the 1% level with an R-square of 0.237. Log GDP per capita and its squared terms retain their signs and significance. The level of urbanisation at time t has a negative and significant impact on youth unemployment at t+15 at the 10% level of significance, while net migration has a positive and significant impact at the 1% level. Service share is not statistically significant. The persistence of the positive and significant coefficient across all three specifications suggests this finding is not driven by confounding macroeconomic or demographic factors. A likely explanation lies in the relationship between literacy and labour force participation, as shown in the appendix in Table A3, higher youth literacy rates are negatively associated with NEET rates, which in turn are positively associated with youth unemployment, suggesting that countries with higher literacy rates tend to have more young people actively participating in the labour force and being counted as unemployed when jobs are not immediately available, while in lower literacy countries a larger share of youth, most likely women, tend to be outside the formal labour force altogether, reducing measured unemployment rates independently of actual employment conditions.

Vocational Training:

Vocational Training Rate and Youth Unemployment (15 Years Lag)

	Dependent variable:		
	Youth Unemployment (t+15)		
	(1)	(2)	(3)
Vocational Training Rate (t)	-0.102**	-0.167***	-0.234***
	(0.051)	(0.057)	(0.058)
Log GDP Per Capita (t)		26.609***	28.407***
		(3.098)	(3.215)
Log GDP Per Capita-Sq (t)		-1.474***	-1.515***
		(0.176)	(0.178)
Service Share (t)			0.117***
			(0.044)
Urban Population (t)			-0.135***
			(0.032)
Net Migration 1000s (t)			0.011***
			(0.002)
Year FE	Yes	Yes	Yes
Observations	673	666	626
R^2	0.037	0.140	0.218
Note:	Standard errors in parentheses. *p<0.1; **p<0.05; ***p<0.01		

Table 3.

The results for vocational training are consistent with the hypothesis that occupation-specific skill provision reduces structural youth unemployment across all three specifications. In the baseline specification (Column 1), the coefficient is -0.102, significant at the 5% level, indicating that a one percentage point increase in the share of secondary students enrolled in vocational training is associated with a 0.102 percentage point decrease in youth unemployment 15 years later, with an R-squared of 0.037. In Column 2, the addition of log GDP per capita and its squared term increases the magnitude of the coefficient to -0.167, now significant at the 1% level, with log GDP per capita entering positively and significantly and its squared term negatively and significantly, consistent with the inverse U-shaped relationship documented in Appendix Figure A3. The R-squared rises to 0.140. In the full controls specification (Column 3), the coefficient increases further in magnitude to -0.234, remaining significant at the 1% level, with an R-squared of 0.218. Service share enters positively and significantly at the 1% level, urban population negatively and significantly at the 1% level, and net migration positively and significantly at the 1% level. The consistent strengthening of the vocational training coefficient as controls are progressively added suggests that the negative association between vocational training enrollment and youth unemployment becomes both larger in magnitude and more precisely estimated once macroeconomic and structural factors are held constant.

Mathematics Proficiency:

Math Proficiency and Youth Unemployment (8 Years Lag)

	Dependent variable:		
	Youth Unemployment (t+8)		
	(1)	(2)	(3)
Math Proficiency LS (t)	-0.099***	-0.027	-0.095***
	(0.023)	(0.030)	(0.034)
Log GDP Per Capita (t)		24.995***	32.156***
		(8.182)	(8.408)
Log GDP Per Capita-Sq (t)		-1.452***	-1.795***
		(0.427)	(0.431)
Service Share (t)			0.208***
			(0.066)
Urban Population (t)			-0.052
			(0.046)
Net Migration 1000s (t)			0.002
			(0.002)
Year FE	Yes	Yes	Yes
Observations	346	346	339
R^2	0.095	0.168	0.199
Note:	Standard errors in parentheses. *p<0.1; **p<0.05; ***p<0.01		

Table 4.

Table 4 presents the results for mathematics proficiency at the lower secondary level, regressed on youth unemployment at an 8-year lag, across three specifications. Results for the 10-year and 15-year lag specifications are reported in the appendix and show broadly consistent patterns. In the baseline specification (Column 1), the coefficient on mathematics proficiency is -0.099, significant at the 1% level, indicating that a one percentage point increase in the share of lower secondary students achieving minimum mathematics proficiency is associated with a 0.099 percentage point decrease in youth unemployment 8 years later, with an R-squared of 0.095. In Column 2, the addition of log GDP per capita and its squared term renders the mathematics proficiency coefficient statistically insignificant, falling to -0.027, with the R-squared rising to 0.168, driven primarily by the explanatory power of the income controls. Log GDP per capita enters positively and significantly and its squared term negatively and significantly, consistent with the inverse U-shaped relationship documented in Appendix Figure A3. In the full controls specification (Column 3), the coefficient on mathematics proficiency recovers to -0.095, significant at the 1% level, with an R-square of 0.199. Log GDP per capita and its squared term retain their signs and significance. Once urbanisation, service sector share, and net migration are additionally controlled for alongside income, the negative association between mathematics proficiency and youth unemployment re-emerges as statistically significant. Service share enters positively and significantly at the 1% level, consistent with the pattern observed in the vocational training specification, while urban population and net migration do not reach conventional significance levels, suggesting that once income and formalisation are accounted for, urbanisation and population flows do not independently explain variation in youth unemployment in this specification. Overall, these results suggest that higher mathematics proficiency at the lower secondary level is associated with lower youth unemployment 8 years later, but that this association is sensitive to the inclusion of income controls, indicating that the relationship between cognitive skill quality and youth unemployment is partly mediated by national income levels and the structural composition of the economy.

Limitations

First, the OLS framework employed does not establish causality. While the lagged structure and year fixed effects help address reverse causality and omitted time-varying factors, the estimates remain associative rather than causal. Country fixed effects were not included as they would have led to overcontrolling, leaving insufficient variation in the data to identify the education quality effects of interest, meaning unobserved country-specific factors may simultaneously influence both education quality and youth unemployment.

Second, data availability poses a significant constraint. The mathematics proficiency variable has only 528 observations across the full sample, reflecting the limited coverage of internationally comparable education quality assessments, particularly in lower income countries. As shown in Figure A7 in appendix, PISA countries are skewed toward higher income levels relative to non-PISA countries, meaning the regression results for mathematics proficiency may not be representative of the full cross-country distribution of education quality and youth

unemployment. The panel data is also unbalanced across all three education quality variables, as shown in the table A6 in appendix, the mean number of observations per country is low across all three variables, at 5.2 for mathematics proficiency, 12.6 for vocational training, and 6.6 for youth literacy, indicating that many countries contribute only a limited number of data points to the analysis.

Future Research

Future research could extend this analysis in several directions. First, future work could examine whether the effect of education quality on youth unemployment is moderated by labour market institutions, for instance, whether the negative association between vocational training and youth unemployment differs across countries with stronger employment protection legislation or higher union density, where wage compression may limit the extent to which skill differences translate into employment outcomes. The vocational training finding could also be explored in greater depth by disaggregating the type of vocational system, dual apprenticeship systems, which combine workplace training with school-based instruction, may have very different effects on youth unemployment compared to purely school-based vocational programs, and future work could examine whether the negative association is driven primarily by countries with strong apprenticeship traditions.

Second, future research could explore more granular dimensions of education quality that this paper is unable to capture due to data constraints. Variation in teacher quality, widely documented as one of the most important determinants of student learning outcomes, could be examined as a channel through which education systems produce different cognitive skill levels across countries (Rivkin, Hanushek & Kain, 2005). Similarly, the number of globally ranked universities and access to high quality education more broadly, including whether students in rural or lower income areas have the same access to quality institutions as those in wealthier urban areas, represent dimensions that aggregate country-level measures inevitably obscure, and that future subnational analyses could usefully examine.

Conclusion

This paper examines whether education quality, measured through youth literacy rates, mathematics proficiency at the lower secondary level, and vocational training enrollment, is associated with lower youth unemployment across countries over approximately 30 years. The results suggest that education quality does matter for youth unemployment, but that the relationship is more complex and channel-specific than a simple negative association.

The most robust finding is that vocational training enrollment is consistently and negatively associated with youth unemployment across all specifications, with the effect strengthening as controls are progressively added. Within comparable economic contexts, countries that invest more heavily in occupation-specific secondary education tend to exhibit meaningfully lower youth unemployment over the medium to long term, pointing toward the expansion of high quality vocational training systems as a potentially effective structural policy

response to youth unemployment, particularly in lower and middle income countries where enrollment remains low and youth unemployment is most severe (UNESCO, World Bank & ILO, 2023).

Mathematics proficiency shows a significant negative association with youth unemployment in the baseline and full controls specifications, suggesting that foundational cognitive skill development matters for youth labour market outcomes, but that its effect is contingent on the broader structural economic environment. This implies that improving cognitive skill quality alone is unlikely to be sufficient without simultaneously addressing the structural conditions that determine whether those skills are rewarded in the labour market.

The positive association between youth literacy rates and youth unemployment is a counterintuitive finding that most plausibly reflects a measurement phenomenon rooted in labour force participation, as documented in the appendix through the relationship between literacy, NEET rates, and youth unemployment, rather than a genuine negative effect of literacy on labour market outcomes. This further suggests that youth unemployment rates alone may present an incomplete picture of youth labour market exclusion and should be considered alongside NEET rates for a more comprehensive assessment.

Work Cited

Autor, D. H., Levy, F., & Murnane, R. J. (2003). The skill content of recent technological change. Quarterly Journal of Economics, 118(4), 1279–1333. https://doi.org/10.1162/003355303322552801

Acemoglu, D., & Restrepo, P. (2018). Robots and jobs: Evidence from US labor markets. Journal of Political Economy, 128(6), 2188–2244. https://doi.org/10.1086/705716

Bell, D. N. F., & Blanchflower, D. G. (2011). Young people and the Great Recession. Oxford Review of Economic Policy, 27(2), 241–267. https://doi.org/10.1093/oxrep/grr011

Blanchard, O., & Wolfers, J. (2000). The role of shocks and institutions in the rise of European unemployment. Economic Journal, 110(462), C1–C33. https://doi.org/10.1111/1468-0297.00518

Carneiro, P., & Heckman, J. J. (2002). The evidence on credit constraints in post-secondary schooling. Economic Journal, 112(482), 705–734. https://doi.org/10.1111/1468-0297.00075

Eurostat. (2023). Youth unemployment statistics. European Commission. https://ec.europa.eu/eurostat/statistics-explained/index.php/Youth_unemployment_statistics

Gregg, P. (2001). The impact of youth unemployment on adult unemployment in the NCDS. Economic Journal, 111(475), F626–F653. https://doi.org/10.1111/1468-0297.00666

Gust, S., Hanushek, E. A., & Woessmann, L. (2022). Global universal basic skills: Current deficits and implications for world development. NBER Working Paper No. 30566. https://doi.org/10.3386/w30566

Hanushek, E. A., Schwerdt, G., Wiederhold, S., & Woessmann, L. (2015). Returns to skills around the world: Evidence from PIAAC. European Economic Review, 73, 103–130. https://doi.org/10.1016/j.euroecorev.2014.10.006

Hanushek, E. A., & Woessmann, L. (2008). The role of cognitive skills in economic development. Journal of Economic Literature, 46(3), 607–668. https://doi.org/10.1257/jel.46.3.607

Hanushek, E. A., & Woessmann, L. (2015). The knowledge capital of nations: Education and the economics of growth. MIT Press. https://mitpress.mit.edu/9780262029179/the-knowledge-capital-of-nations/

ILO. (2023). World employment and social outlook: Trends 2023. International Labour Organization. https://www.ilo.org/wcmsp5/groups/public/---dgreports/---dcomm/---publ/documents/publication/wcms_865332.pdf

Lilien, D. M. (1982). Sectoral shifts and cyclical unemployment. Journal of Political Economy, 90(4), 777–793. https://doi.org/10.1086/261088

Mincer, J. (1974). Schooling, experience, and earnings. NBER.

Paul, K. I., & Moser, K. (2009). Unemployment impairs mental health: Meta-analyses. Journal of Vocational Behavior, 74(3), 264–282. https://doi.org/10.1016/j.jvb.2009.01.001

Pissarides, C. A. (2000). Equilibrium unemployment theory (2nd ed.). MIT Press.
Quintini, G., & Manfredi, T. (2009). Going separate ways? School-to-work transitions in the United States and Europe. OECD Social, Employment and Migration Working Paper No. 90. https://doi.org/10.1787/221717700447
Ryan, P. (2001). The school-to-work transition: A cross-national perspective. Journal of Economic Literature, 39(1), 34–92. https://doi.org/10.1257/jel.39.1.34
Scarpetta, S., Sonnet, A., & Manfredi, T. (2010). Rising youth unemployment during the crisis. OECD Social, Employment and Migration Working Paper No. 106. https://doi.org/10.1787/5kmh79zb2mmv-en
Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. Econometrica, 73(2), 417–458. https://doi.org/10.1111/j.1468-0262.2005.00584.x
UNESCO, World Bank & ILO. (2023). Transforming technical and vocational education and training for successful and just transitions. UNESCO. https://www.unesco.org/en/articles/new-report-unesco-world-bank-and-ilo-says-technical-and-vocational-education-and-training-must

Appendix

Math Proficiency and Youth Unemployment (15 Years Lag)

	Dependent variable:		
	Youth Unemployment (t+15)		
	(1)	(2)	(3)
Math Proficiency LS (t)	-0.136***	-0.081*	-0.163***
	(0.031)	(0.042)	(0.049)
Log GDP Per Capita (t)		21.674**	27.225**
		(10.959)	(11.389)
Log GDP Per Capita-Sq (t)		-1.263**	-1.511**
		(0.581)	(0.592)
Service Share (t)			0.069
			(0.099)
Urban Population (t)			0.033
			(0.065)
Net Migration 1000s (t)			0.006*
			(0.003)
Year FE	Yes	Yes	Yes
Observations	182	182	176
R^2	0.113	0.162	0.213
Note:	Standard errors in parentheses. *p<0.1; **p<0.05; ***p<0.01		

Table A1.

Math Proficiency and Youth Unemployment (10 Years Lag)

	Dependent variable:		
	Youth Unemployment (t+10)		
	(1)	(2)	(3)
Math Proficiency LS (t)	-0.097***	-0.021	-0.096**
	(0.026)	(0.034)	(0.039)
Log GDP Per Capita (t)		23.850***	30.241***
		(8.868)	(9.142)
Log GDP Per Capita-Sq (t)		-1.401***	-1.709***
		(0.464)	(0.469)
Service Share (t)			0.205***
			(0.075)
Urban Population (t)			-0.030
			(0.052)
Net Migration 1000s (t)			0.003
			(0.003)
Year FE	Yes	Yes	Yes
Observations	279	279	272
R^2	0.090	0.165	0.198
Note:	Standard errors in parentheses. *p<0.1; **p<0.05; ***p<0.01		

Table A2.

	Youth Literacy Rate	NEET	Youth Unemployment
Youth Literacy Rate	1	-0.264***	0.333***
NEET	-0.264***	1	0.345***
Youth Unemployment	0.333***	0.345***	1

Table A3.

	Math (Lower Sec)	Math (Primary)	Reading (Lower Sec)	Reading (Primary)
Math (Lower Sec)	1	0.925***	0.93***	0.876***
Math (Primary)	0.925***	1	0.744***	0.804***
Reading (Lower Sec)	0.93***	0.744***	1	0.858***
Reading (Primary)	0.876***	0.804***	0.858***	1

*Notes: * p<0.1; ** p<0.05; *** p<0.01*

Table A4.

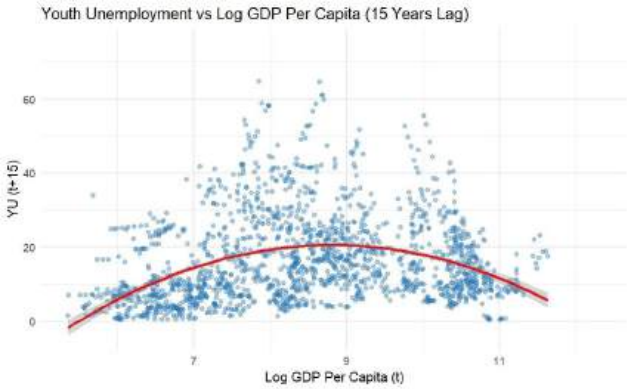


Figure A5.

	Countries	Means Obs	Median Obs	Min	Max
Math LS	103	5.2	6	1	8
Voc Training	176	12.6	10	1	26
Youth Lit	166	6.6	4	1	25

Table A6: Observations per Country by Education Quality Variable

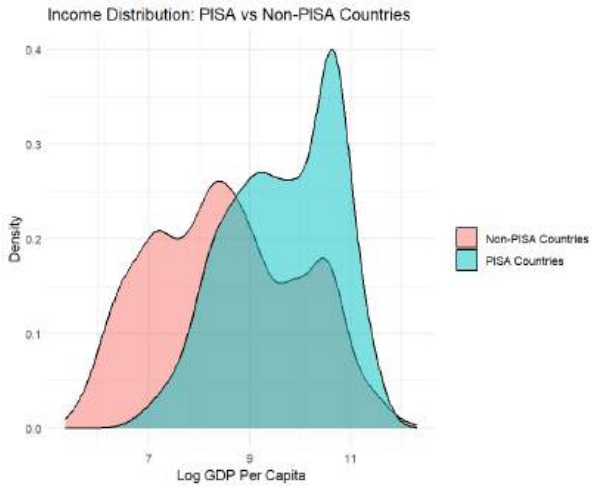


Figure A7 : Income Distribution: PISA vs Non-PISA Countries

Young Women and Grocery-Brand Marketing Tactics By Macy Hart

Abstract

In recent years, food choices have been linked with consumeristic and materialistic tendencies, and food consumption is often related to socio-economic status. Marketing and advertising teams must use visual or aesthetic components on the labels of food products to promote and sell food products as high quality and fit for the wants and needs of specific audiences. Consumerism is often seen as a preoccupation held by women, especially of younger age groups. This paper measures the extent of efficiency of ten tested marketing strategies used on grocery item labels, specifically among an age group of high school and college-aged women. A two-part quantitative study was conducted, collecting 38 grocery item labels and their marketing strategies used in a content analysis table. This data was then applied to a survey study for women of the tested age group, where 92 responses were collected. Results found that of all tested strategies, color was the most efficient, followed by font and logo, images and symbols, and brand name. These findings suggest that customers of this age group should be conscious of the use of these strategies on food product labels when grocery shopping and reveal to marketing and advertising teams what components of a label are most enticing to this audience, potentially shifting focus in the advertising world.

Keywords consumerism, aesthetic appeal, marketing, manipulation, demographic, target market group, wellness

Introduction

"Consumerism: a preoccupation with and an inclination toward the buying of consumer goods" (Meriam-Webster, n.d.). Today we live in a consumer culture where society functions on the pressures to buy anything that will serve as an upgrade, or at least any illusion of one. Pietrykowski finds, "People engage in 'lifestyle' shopping where goods function as symbols that convey messages and produce identities (Shields 1992)." (2004, p. 3). With that, consumer culture revolves especially around women and elements regarding materialistic tendencies- or as Ivashkevich quotes, "…contemporary girls' and women's empowerment 'is often directly tied to consumer culture and the ability to purchase, with [female] agentive powers premised upon and enabled by the consumption of products and services'" (2011, p. 7). Of course, materialistic tendencies in feminine culture have become stereotypical, but in recent years, food has begun to present itself as a new sort of luxury- and choosing what to buy in grocery stores has become commercialized. "…consumer health-related activities since at least the beginning of this century have been linked with their desire to display a social class distinction through their food purchases and consumption" (Wiedenroth and Otter, 2021, p. 2). Bearing in mind the consumer pressures towards food, marketing and advertising industries must strategically promote food products to captivate the customer. This presents the question: To what extent do marketing tactics used by food, drink, and supplement brands influence the purchasing choices of young

adult women in grocery store settings? The goal of this study will be for young women to be able to identify and realize how marketing tactics may influence them in daily situations and to help give further insight to marketing and advertising industries.

Literature Review

Marketing Strategies Regarding Aesthetics

The deciding factor that leads consumers to purchase an item comes down to what effect it has on the consumer. First impressions are crucial, and what catches the eye of a customer is often the deciding factor in purchasing or not purchasing an item. From the Journal of Consumer Psychology, Reimann and his team conducted a psychological investigation regarding aesthetics of a product and consumer choices. In the experiment, 80 items were chosen for aesthetics, and the other 80 chosen were functional. Results showed, "Products in the aesthetics conditions were chosen 69% of the time (31% of the time they were not chosen), while products in the standardization condition were chosen only 57% of the time (43 % of the time they were not chosen) (%2 = 149.06, df=l, p < .001)" (2010, p. 4).

Aesthetics are subjective, however, and what is considered visually appealing varies among different audiences. That said, brands must rely on specific demographic trends to successfully capture a wide audience of people willing to purchase a product. "Zhining Liu, XiaoFan Niu, and Chenyi Zhuang suggested that for companies to efficiently capture the interest of their customers, it is crucial that they possess an in-depth and comprehensive comprehension of the target audience's preferences, behavior, and values [5]" (Chen, 2023, p. 3). Consumption refers to the purchasing of an item to fulfill an individuals' wants or needs. The wants and needs of an individual can be significantly based off demographics such as age, gender, race, social class, etc., meaning brands must select specific demographic groups, also referred to as a target market audience, to base the aesthetics of their product on, or as Mohebbi found from Kauppinen-Räisänen, "In actual fact, consumers' responses to the design and color of packaging is assumed to be converted into brand preference; simply put, the decision to opt for a brand is based on aesthetics of packaging" (2014, p. 5).

Consumer Culture

Consumer culture in the West is apparently due to short-term psychological gains. "…the human mind constantly dreams of new forms and new patterns. Each generation creates new desires and requirements for what is 'needed' (Csikszentmihalyi and Rochberg-Halton 1981)" (Zinkhan, 1994, p. 1). Current trends and aesthetics can often trick people into believing that if they are to purchase an item, they will benefit significantly from the impulsive purchase. "70 percent of all brand and purchase decisions are made in-store at the moment of buying, even if a consumer enters a store with the intention of purchasing specific products based on a shopping list (Kauppinen-Räisänen, 2014)" (Mohebbi, 2014, p. 2). Aesthetics play a huge role since trends are always shifting. The visual component is crucial for catching the eye of a customer and

appealing to their wants or needs. From a neurological standpoint, Reimann et al. quotes, "…aesthetic designs seem to trigger certain positive responses in consumers such as an immediate desire to own the product" (2010, p. 2).

Feminine Culture and Wellness Trends

Rising trends in wellness have become apparent, as quoted by Accenture, "The $502 billion consumer healthcare market is now largely driven by prevention and wellness products—especially vitamins, nutritional supplements, weight-management products, and fortified foods and beverages—and is expected to grow 50 percent over the next five years" (Walsh, 2015, p. 2). Walsh from the Journal of the American Society of Aging explains the promotion of wellness, "The term generally is understood to reach beyond physical health to encompass a more holistic mind−body approach that also considers good mental and emotional health"(2015, p. 1). The idea of wellness coincides with dieting or the idea that one must be aware of what goes into their body.

However, the growing variety of dieting trends presented in the media can create inaccurate or unrealistic concepts regarding nutrition, pressuring young women to keep up with the exponentially impractical expectations. Crowley explains, "As certain feminist theorists point out, 'most women in Western cultures' are unhappy with their bodies—an unhappiness that stems from the relentless societal pressures of what women's bodies should look like" (2002, p. 1). Thompson and Chad from the Canadian Journal of Public Health presented a study testing young girls from different age groups on eating disorder inventory (EDI), social physique anxiety scale (SPAS), and body image questionnaire (BIQ). The EDI scores suggested that as the girls matured, they became more dissatisfied with the way they looked, and similarly the SPAS scores suggested that the participants' anxiety over their body image increased as they physically grew. The BIQ results found, "34% of the pre-pubertal girls, 36% of the pubescent girls and 76% of post-pubescent girls chose a smaller body shape as their preferred body type" (2002, p. 2). They claim no girls in the study chose a large body image. Body dysmorphia in young women is seemingly an inevitable occurrence among current culture, and projected health and diet trends are potentially significant factors that contribute to self-conscious thoughts and eating behaviors of women.

Gap in the Research

In previous studies, it has been determined that companies select target audiences to capture the attention and brand loyalty from a specific group of customers with shared interests that align to the product. "Knowing the importance of the target audience during the marketing process, companies should conduct market research to understand the target audience's demographics, behavior patterns, pain points, and motivations, which can involve using a variety of research methods, such as surveys, focus groups, and social media listening, to gather data about the target audience." (Chen, 2023, p. 5).

Additionally, because of this idea, it is known that marketing and advertising teams must use marketing techniques and visual components that will attract and manipulate the target market group based on their priorities and values "In Stealth Marketing, people are unknowingly part of the campaign to support the promotion. When these promotional strategies target teenagers, it creates a most deceptive way and proves the unethical practices of an organization to advertise their product" (Tripathi, 2016, p. 2).

Considering the research, there has been a population gap in research that has not yet included how young women specifically are affected by marketing tactics used by grocery-brand items. In this paper, a content analysis approach will be conducted to analyze frequencies and trends in specific visual marketing tactics of grocery- brand items, which will then be used in a survey study, where high school and college- age girls will select the grocery brand items they would choose in a grocery store setting and what factors led them to that decision to discover which marketing strategies prove to be most effective on this audience. My hypothesis is that young women will lean towards grocery products that use specific displays of colors and imagery that evoke an aesthetic appeal to the product, as well as products that present a sense of health or naturalness that fit lifestyle appeals.

Research Design and Methodology

Study Design

This project was designed to expose the factors in the packaging of grocery products like food, drinks, and supplements that persuade a customer that the individual product is worth buying. I conducted a two-part quantitative study that included both a content analysis study which recorded the frequencies of how often certain marketing strategies were used on different products, and a survey study taken by young women to measure which of strategies proved the most effective for the specific target market audience. My content analysis was heavily based on the one conducted in a study by Jones et al., from the Journal of Advertising, which measured the frequency of violence depicted in advertisements and television commercials and found that fear factor was an intriguing marketing tactic for people, especially of younger audiences (2010). This study chose to focus on one marketing strategy, meanwhile mine focused on multiple different strategies competing against each other, and all were recorded simultaneously. Yet, they both are similar in the aspect of measuring the frequency of marketing strategies with the goal of determining what effect it has on a consumer. Choosing a survey was significantly inspired by Montoya and Scott's study design from the Journal of Public Policy and Marketing, in which the stress levels of teenagers was measured to determine how it altered their consumer behavior (2013). Although the survey asked about the daily stressors of teenagers to measure how drastic the effects were on consumer behavior, the study presented a similar goal to my project by measuring the outside factors that may influence consumer behavior of younger generations. My goal was to reveal the most successful marketing tactics to teach women the intent that goes into influencing a customer and to bring awareness to the ways they may be unknowingly influenced

every day. This additionally brings value into the marketing industry and is meant to display which marketing strategies are the most influential and effective to use for this target market group.

Hypothesis and Assumptions

Prior to any research, I assumed that products that follow the conformed trends and values of specific demographics are highly persuasive and therefore, the current trends and beauty standards often determine decisions making of women, especially in their eating habits. Additionally, I assumed that marketing teams must focus on a specific target market group and follow trends that occur in the given demographic. With that, many marketing teams tend to market products to appeal to younger audiences. This is because brands want to secure brand loyalty of younger customers to maximize profits for longer periods of time. Based on this context, I ultimately assumed that testing the products that best conformed to the trends and aesthetics of younger adult women would be a valuable experiment that could teach women about what goes into persuading them to purchase an item, while providing more information that can help advertising teams who may want to appeal to younger audiences such as the demographic I am testing.

Before I conducted my study, I hypothesized that of the ten strategies I chose to test, the ones that would prove most effective included color, images and symbols, and lifestyle appeal. From my research collected in the beginning stages of the project, I found that color and imagery were both highly influential to aesthetic appeal and provoke certain emotions and feelings that would likely have a strong contribution to persuading a customer. Furthermore, with background understanding of the current uprising trends of wellness and femininity, as girls grow up, they tend to become more exposed to high and often unrealistic standards. These can become extremely influential to women in their daily routines, convincing many that they should follow the socially acceptable or "healthier" lifestyles and leaving many vulnerable to following certain dietary trends that may or may not be reliable.

Subjects

Content Analysis

The subjects I analyzed in my content analysis included 10 frequently used marketing strategies used on grocery brand items as shown in Table 1. I chose the 10 most distinguishable strategies based on preliminary research and a prior pilot study through visiting grocery stores. Data was recorded to measure the frequency of how often certain strategies may show up in stores.

Survey Study

My study used human subjects, specifically women from ages 14-22 (freshmen in high school to seniors in college). I picked this group because girls at this age are increasingly

becoming more independent, and are typically impacted by the media, creating a higher likelihood for younger developing women to fall victim to advertising and marketing tactics. Simultaneously, many brands and marketing teams are intentionally trying to appeal to younger groups for the sake of gaining brand loyalty, and therefore many marketing trends revolve around enticing younger audiences.

Research Instruments

The content analysis table consisted of the ten marketing strategies I chose to test in my experiment, and recorded a mix of ten food, drink, and supplement items to measure the prevalence of each strategy. The sheet shown in Table 1 shows what I used to record data, where a total of thirty-eight items were recorded by the end of the study. For each item, the strategies that were explicitly depicted on the packaging were checked off, and any one strategy that may have stood out was circled and highlighted for the sake of creating survey questions in the next step of my study. Once I selected one item to represent each marketing strategy, I found three additional items that were similar to the original item, yet not representative of the strategy (for example, I chose four sodas of the same brand, but three of which did not have an exceptional amount of color). The ten initial items were randomly selected and implemented if they showed excessive use of one strategy to serve as test variables in the next step of my study, then the additional three were chosen based on their test variable to serve as controls.

Table 1

Content Analysis Sheet

	Food Item 1 _____	Food Item 1 _____	Food Item 1 _____	Drink Item 2 _____	Drink Item 2 _____	Drink Item 2 _____	Supplement Item 3 _____	Supplement Item 3 _____	Supplement Item 3 _____
Color									
Font/ Logo									
Images/ Symbols									
Images/ Symbols									
Lifestyle Appeal									
Nostalgia/ Personal Connection									
Credibility of Brand									
Limited Edition									
Brand Name									
Celebrity Promotion/ Association									

When results were collected from the content analysis study, I then designed a survey study based on what I had gathered. After the questions regarding consent and personal information, questions 5 and 6 displayed in Table 2 were repeated ten separate times, each time showing four different samples of grocery items for the participant to select, (question 5) followed by the list of marketing strategies for the participant to select from to identify which strategies led them to their decision (question 6). Afterwards, I left a space for any open comment, yet nothing was taken as qualitative data nor used in my experiment.

Table 2

Survey Questionnaire

Question:	Response:
1. Overview: This survey is intended to record the effectiveness of different marketing tactics used for grocery items and record the advertising aesthetics that appeal to young women. Your participation is completely voluntary, and all responses are anonymous. Thank you so much for your participation!	I have read the overview provided above, and I consent to participate in this survey: o Agree
2. Please enter your name and phone number (optional):	Open-ended response
3. Please select your age:	o 14 o 15 o 16 o 17 o 18 o 19 o 20 o 21 o 22 o I do not wish to say
4. Please select your grade:	Please select your grade: o High school freshman o High school sophomore o High school junior o High school senior o College freshman o College sophomore o College junior o College senior o I do not attend high school or college

5. You are going shopping. Based on VISUAL APPERANCE, please select one of the following items that you prefer:	Sample 1 Sample 3
6. Please select which of the following factors influenced your choice from the previous question. Select ALL that apply:	○ Color ○ Font ○ Text Size ○ Symbols/ Images ○ Lifestyle Appeal ○ Emotional Appeal ○ Credibility Appeal ○ Economic Appeal ○ Sense of Urgency ○ Celebrity Promotion/ Association
Any additional comments or concerns (optional):	Open-ended response

Procedures

Content Analysis

Before the initial study, I took ten of what I believed to be the most effective marketing tactics used on grocery-brand items based on a pilot study in which I visited a grocery store and noted the strategies that were most evidently used. Using a content analysis sheet, I visited multiple grocery stores and recorded food, supplement, and drink items along with which marketing strategies were used on each item's label. I decided to initially record ten grocery items, one to represent each strategy, serving later as a test variable. From there, three similar control items were chosen (apart from one exception). These control items were of the same product function with similar design concept yet did not show any excessive use or representation of the tested marketing strategy. In some cases, control items were selected to

represent a separate marketing strategy, therefore allowing me to test the two against each other in the next step of my experiment, the survey study. Once all items were recorded, I circled each item that showed a distinctive use of one strategy. From the circled items, I highlighted each item that would be implemented in my survey questionnaire as a test variable (Appendix).

Survey Study

With my synthesized data, I created a survey study acting as a simulation. For each marketing strategy, two questions were presented: four grocery items were shown, and the participant was to choose the item they would have selected in a real-life setting. The following question would ask the participant to choose any of the ten tested marketing tactics that may have influenced their choice. In each question, one of the four presented items was used to represent each strategy over the three additional items (for example, one item would have exceedingly more color than the rest of the shown items).

Ethical Considerations

To conduct a survey that would ensure absolute safety and convenience for my participants, I decided ideas revolving around social influence and unrealistic beauty standards that may have influenced purchasing choices was a contextual issue that did not need to be addressed to young women for mental health reasons. I found that by creating a shopping survey that was designed as if it were a game would leave the participants with a positive experience causing no stress to be involved, as well as encourage young women to want to take my survey. The survey asked each participant to list their age, grade, and name- all of which were completely anonymous. Listing a name was optional for giveaway purposes and kept confidential. The first question asked the participant to agree to a consent overview and there was additional space available for any comments that the participants may have wanted to add at the end. After the survey was created, I did a quick test run with 5 classmate participants before making it available to the public to confirm that there were no errors or anything that went unconsidered in the process of designing the survey.

Delimitations

When it came to conducting the experiment, I chose to prioritize my focus on the frequency and performance of each marketing strategy. One delimitation was the data collection method of qualitative interviews. I felt that many items I tested coincide with current health and wellness trends, which in some cases branch from insecurities. Interviewing participants on these types of sensitive topics seemed unethical and left the risk of mentally harming participants. Therefore, I decided quantitative data could sufficiently speak for itself.

Moreover, when recording data, I specifically focused on anything that was explicitly on the label itself- not what the product was nor the price or function of it. If it was a food, drink, or supplement, I recorded anything that used one of the listed marketing strategies of my project. If a product implemented use of a strategy even in the slightest bit, it was marked and any further

judgement narrowed products down into smaller sections later. I only focused on the ten general strategies I found to be the most frequently used and only selected one product to represent each strategy for the survey questions to keep the experiment realistic and manageable within a given timeframe.

Limitations

The experiment presented some imperfections along the way as well. Finding the items to record for the content analysis was solely based on my own judgement, and categorizing each product could have resulted in some mistakes. Furthermore, I was limited to finding one test variable, or grocery-brand item, for each tested marketing strategy, therefore making it difficult to find a single item to entirely represent one strategy. This was especially difficult to find for some strategies such as nostalgia and personal connection, where finding one item to represent a personal connection towards a broad and general audience seemed a bit impractical. When publishing my survey study, I had to consider the possibility that when participants were answering questions, some answers could have been dishonest or lacked effort and critical thinking which could have altered data collection. I additionally realized that my survey study, in which the studied age group was high school and college-age girls, was predominantly advertised in a high-school setting. This means it was much easier to gather results from students at the school I attend or ones in the surrounding area, and I was able to gather significantly more high-schooler participants rather than college-aged participants or any other responses from outside my area, suggesting a lack of diversity in specific age group and location.

Data Results and Analysis

Content Analysis Results

Quantitative data was collected to measure the frequencies of how often each marketing strategy showed up on all tested products. Data collection of content analysis served as a foundation to finding data to test each of the ten strategies and their efficiency later in the survey study: finding one test variable and three control variables (with one exception where only two variables were tested in the font and logo strategy), resulting in a total of thirty-eight recorded grocery products by the end of the survey. A completed chart can be seen in Table 3.

Table 3

Completed Content Analysis Table

	Drink: Olipop Tropical Punch Soda	Food: Pringles Original (new label)	Supplement: Force Factor Greens Powder	Drink: Alani Energy Sherbert Swirl	Food: Lesser Evil Popcorn	Food: Girl Scout Cookies Thin Mints	Supplement: Optimum Nutrition Creatine Powder	Drink: Alani Energy Winter Wonderl-and	Food: OREO Cookies	Supplement: Lemme Sleep Melatonin Gummies
Color	☑	✓	✓	☑	✓	✓	✓	☑	✓	☑

Font/ Logo	✓	☑	✓	✓	✓			✓	✓	✓
Text Size	✓		☑		✓		✓		✓	✓
Images/ Symbols	✓	✓	✓	☑	✓	✓		☑	✓	✓
Lifestyle Appeal	✓		✓	✓	☑		☑	✓		☑
Nostalgia/ Personal Connection	✓	✓				☑		✓	✓	
Credibility of Brand							☑			
Limited Edition						☑		☑		
Brand Name	✓	✓	✓	✓	✓	✓	✓	✓	☑	✓
Celebrity Promotion/ Association										☑

	Drink: Olipop Classic Root Beer Soda	Food: Pringles Original (old label)	Supplement: Bloom Greens Powder	Drink: Celsius Energy Retro Vibe	Food: Orville Popcorn	Food: Nestle Chocolate Chip Cookies	Supplement: Lineage Creatine Powder	Drink: Alani Energy Blue Slush	Food: 365 Chocolate Cookies (Whole Foods brand)	Supplement: Natrol Melatonin Gummies
Color	✓	✓	☑	✓	✓	✓	✓	☑	✓	☑
Font/ Logo	✓	☑	✓	☑	✓		☑	✓		
Text Size	✓									
Images/ Symbols	☑	✓	✓		✓	✓	✓	☑	☑	✓
Lifestyle Appeal	✓		✓	✓			☑	✓		☑
Nostalgia/ Personal Connection	✓	✓		☑	☑	☑				
Credibility of Brand			☑							

Limited Edition										
Brand Name	✓	✓	✓	✓	✓	✓	✓	✓		✓
Celebrity Promotion/ Association										

	Drink: Olipop Doctor Goodwin Soda	Supplement: Be Well Greens Powder	Drink: Celsius Energy Essentials Orangesicle	Food: Pop Secret Popcorn	Food: Siete Chocolate Chip Cookies	Supplement: Clean Nutra Women's Creatine Powder	Drink: Alani Energy Dream Float	Food: Trader Joe's Joe-Joe's Chocolate Cookies	Supplement: Publix Brand Melatonin Gummies
Color	✓	✓	✓	✓	✓	☑	☑	✓	✓
Font/ Logo	✓	✓	✓	✓	✓	✓	✓	☑	✓
Text Size	✓		✓			✓			
Images/ Symbols	☑	✓		✓	✓		☑	☑	✓
Lifestyle Appeal	✓	☑	☑		☑	☑	✓		☑
Nostalgia/ Personal Connection	✓			☑	✓			✓	
Credibility of Brand		✓							
Limited Edition									
Brand Name	✓	✓	✓	✓	✓	✓	✓		
Celebrity Promotion/ Association									

	Drink: Olipop Vintage Cola Soda	Supplement: Thorne Greens Powder	Drink: Reign Energy Rainbow Sherbert	Food: Smartfood Popcorn	Food: Pillsbury Christmas Tree Sugar Cookies	Supplement: Old School Creatine Powder for Women	Drink: Alani Energy Pink Slush	Food: Newman O's Chocolate Cookies	Supplement: OLLY Sleep Melatonin Gummies
Color	✓	✓	☑	✓	✓	✓	☑	✓	☑
Font/ Logo	✓	✓	☑		✓	✓	✓		✓
Text Size	✓				✓	☑		☑	☑

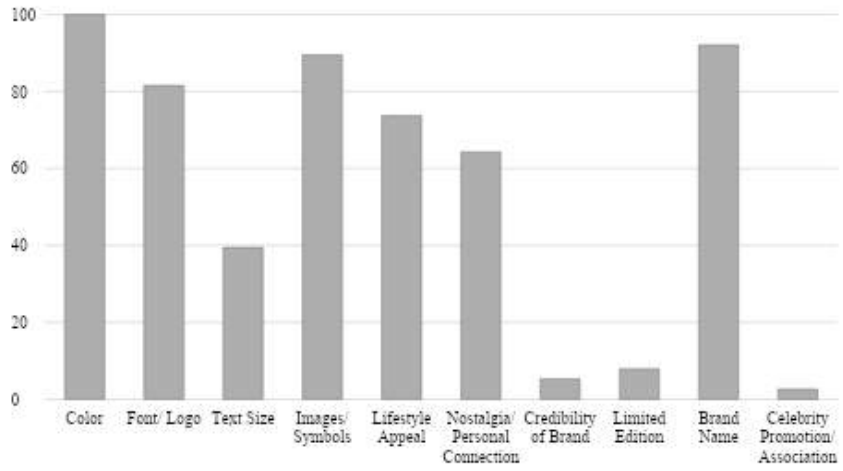
Images/ Symbols	☑	✓	✓	✓	☑	✓	☑	✓	✓
Lifestyle Appeal	✓	☑	✓	☑		☑	✓	☑	☑
Nostalgia/ Personal Connection	✓			✓	☑	✓			
Credibility of Brand									
Limited Edition					☑				
Brand Name	✓	✓	✓	✓	✓	✓	✓	✓	✓
Celebrity Promotion/ Association									

Marketing Strategy Frequencies

When content analysis data was gathered, products that showed use of any marketing tactics on a label were marked, regardless of if the strategy was or was not used well on a product. Therefore, nearly every product was marked for strategies like color and symbols, because nearly every item had color and some sort of image or symbol on the label. Of the thirty-eight recorded items, 100% were marked to have color present in the label. Similarly, brand name had 92% (thirty-five of the thirty-eight) of the items marked, where brand name was determined as anything that was not a product of the grocery store's own brand. Therefore, color and brand name showed the highest frequency, with images and symbols next in line having 89% of items marked (thirty-four of the thirty-eight items). These strategies are basic foundations to creating a product label, unlike other advertising strategies such as limited edition, credibility of the brand, and celebrity promotion and association, which are not as fundamental to the makeup of a label and therefore were marked in 8%, 5% and 3% of all recorded grocery products respectively. A chart of all recorded frequencies can be seen in Figure 1.

Figure 1

Percentage Frequency of Each Marketing Strategy

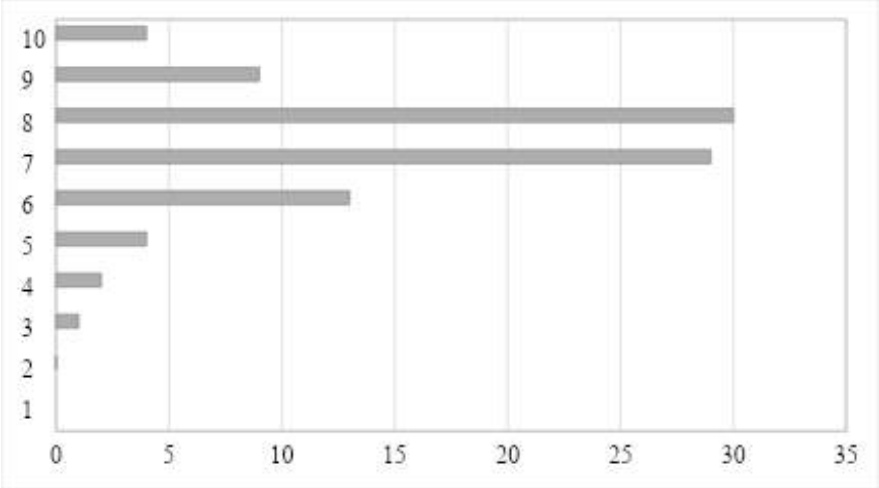


Survey Study Results

Quantitative data was collected in the survey study through a series of multiple choice and select all questions given to teenage girls and young women between the ages of 14-22 years old. Over the span of approximately three weeks, a total of 92 responses were collected, most of which were in the 16 to 17-year-old range, and juniors in high school. A configuration of all participants and their ages can be seen in Figure 2.

Figure 2

Age Breakdown of Participants



Tested Item Versus Outcome

Of the ten multiple choice questions asking participants to select their preferred item, seven showed overall favoritism of the test variable item. Some of the most selected items included the Sherbert Swirl Alani, selected by 78% of participants, which served as a test variable for images and symbols, and OREO cookies, selected by 79% of participants, the test variable for brand name. Furthermore, items like Force Factor Smarter Greens Powder,

representing text size, were only selected by 1% of participants. Their charts regarding what participants selected are seen respectively in Figures 3, 4 and 5.

Figure 3

Result Breakdown of Images and Symbols Question

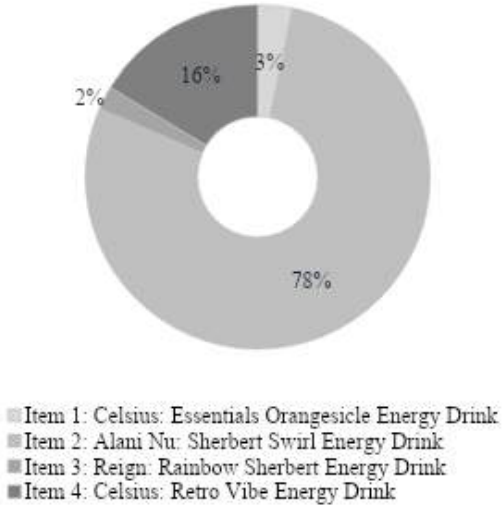


Figure 4

Result Breakdown of Brand Name Question

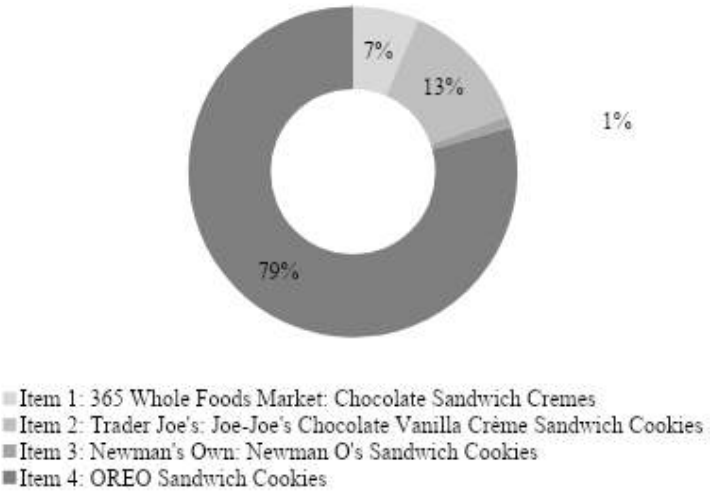
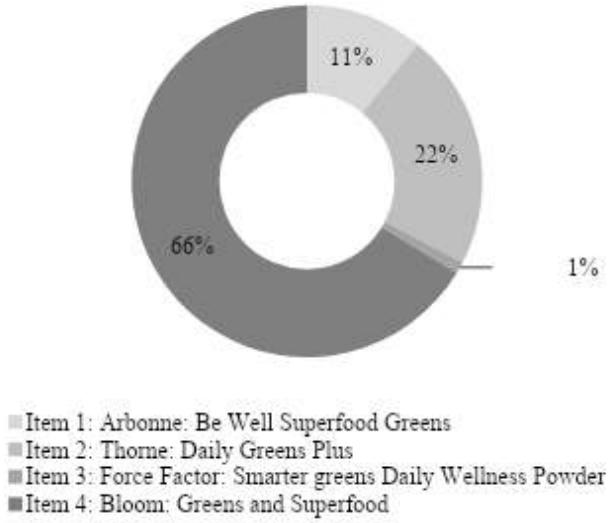


Figure 5

Result Breakdown of Text Size Question



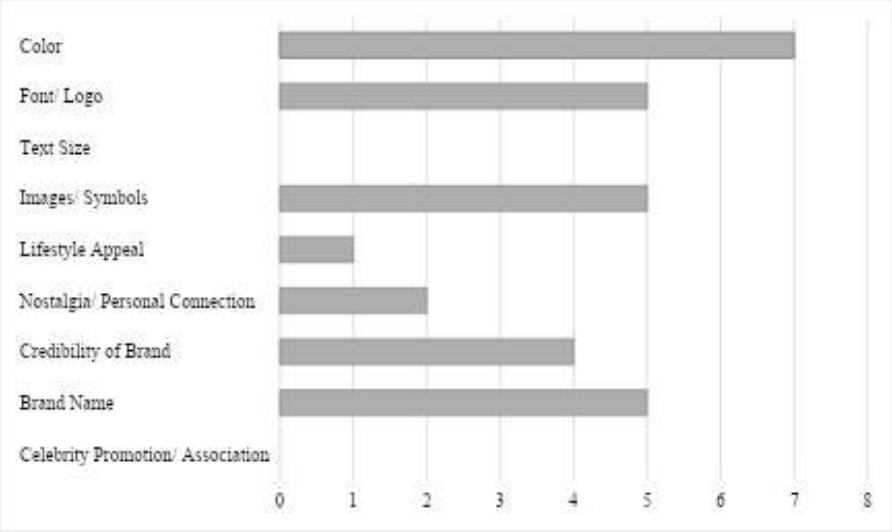
Effectiveness of Each Strategy

When participants selected an item, the following question asked them to select all marketing strategies that they felt influenced their decision. Strategies were deemed efficient if they resulted in the top three strategies selected by participants. Of the ten marketing strategies, six proved effective under their corresponding question. These included color, font and logo, images and symbols, lifestyle appeal, nostalgia and personal connection, and brand name. Color and nostalgia and personal connection both seemed to be represented the best, with color being selected 77 times under its corresponding question, and nostalgia and personal connection selected 62 times under its corresponding question.

It was also important to note that out of ten questions, color was in the top three of selected strategies a total of seven times; six of which color came out to be the most selected strategy- over half of the survey. Additionally, the strategies font and logo, images and symbols, and brand name all resulted in being in the top three selected strategies a total of five times throughout the survey. Strategies like lifestyle appeal were less popular, and shown in the top three responses once, along with text size and celebrity promotion and association, which did not make up any of the top three responses at all in any of the ten questions. A total breakdown of how often each strategy was recorded to be in the top three strategies selected by participants can be shown in Figure 6. The four strategies that were shown most frequently in the survey results were contemporaneously the same four strategies with the highest frequencies in the content analysis table as seen back in Figure 1. These four strategies are essential to creating a product label, and therefore their consistency among all the tested products and frequent exposure to an audience likely put these strategies at an advantage. Additionally, competing strategies did not perform as well if they were shown among grocery items with emphasis on any of the four most effective strategies, and even if they were shown to be effective, these four strategies had the most influence and control over the participants' decisions.

Figure 6

Frequency of Efficiency in Marketing Tactics



Conclusion

New Understanding

The goal of my research project was to find how efficiently certain marketing strategies displayed on grocery item labels worked to attract a young female audience through data collection methods of content analysis and a survey questionnaire. By measuring the results from the survey study, I came to the understanding that grocery products were most successfully chosen by this demographic when strategies including color, followed equally by font and logo, images and symbols, and brand name were displayed on the logo. This partially supports the hypothesis created in the preceding stage of my data collection, which stated that color, images and symbols, and lifestyle appeal would prove to be the most effective. Lifestyle appeal was a marketing strategy that was surprisingly not popular among this demographic, as it showed up frequently in the content analysis study, and considering the evidence that suggests a rise in health and wellness trends for women.

Fulfillment of the Gap

As stated earlier, it has been determined previously that companies focus on attracting demographics with shared interests and values to base their products off, and how it is crucial that companies determine these customers' values. "By analyzing the data, companies can gain a better understanding of who their customers are, what they care about, and how they behave" (Chen, 2023, p. 5).

Moreover, since different target market groups typically have differing values and viewpoints, marketing teams take these values into consideration when designing a label of a product, making it seem as if the product reflects the exact want or need of an individual in the

target market group. "…beautiful aesthetics may elicit notions of pleasure and hedonics. Activation of these concepts may lead people to view a food as more indulgent" (Hagen, 2021, p. 1).

Through content analysis and a survey questionnaire, my study successfully measured the frequencies of marketing strategies and how efficiently these impacted female teens and young adults, and therefore, addressed the population gap, stating that these women and their purchasing decisions were most enticed by color, font and logo, images and symbols, and brand name; the basic foundations to a product label.

Implications

This project was primarily conducted for young women of the studied age group to teach them about the tendencies of marketing teams, and how influential the presentation of a product can really be to someone in this demographic. This teaches female consumers what components of a product label to look out for when shopping, and to be conscious of the impact it has on them, ensuring the customer is grocery shopping for efficiency, nutrition and quality; not solely based on how the packaging of a product looks. Additionally, this can be beneficial to marketing teams who base their target market group on a similar demographic. This experiment has taught the customer to particularly be aware that color, font and logo, images and symbols, and brand name all have great control over one's purchasing decisions, while simultaneously showing those in marketing and advertising fields the value of each of these strategies and the impact they have on the more modern generation, shifting direction of the processes or priorities in the advertising world.

Limitations of the Conclusion

Unfortunately, my research presented a few flaws within the data collection process. The main concern being that I was unable to collect enough data from the older age range of the tested group in the survey questionnaire, and about half of my data is from peers in a younger, high school age range. This study was especially intended for college age women because grocery shopping applies to this group significantly. While this does still apply to women who are in high school, this age tends to be less likely to be grocery shopping independently.

Additionally, individually selecting one grocery product to represent each strategy in my study was a difficult process. Choosing a multitude of products to represent each strategy instead of just one would have likely led to a deeper and more comprehensive understanding of my research question, and perhaps a different result.

There were certain marketing strategies that my study was initially designed to dial in on, but due to some inconsistencies throughout the paper, I did not get to focus on these strategies to the same extent I planned on in the beginning stages of my research.

Areas for Future Research

Replication of the Current Study

To replicate the current study, the researcher must use the content analysis form to record the frequencies of the ten chosen marketing strategies by marking the use of them on various grocery products, and the survey questionnaire to test high school and college aged women from ages 14-22 from Tables 1 and 2. When finding products to mark in the content analysis coding form, the researcher must consider that this is a two- part study, and this data will subsequently be implemented into the survey questionnaire, making it important to record data based on what is needed for the next part. Therefore, it is important to record both grocery products that are well representative of the tested marketing strategies (test variables) along with additional products that are not well representative of the marketing strategies (constant variables). Moreover, when considering what products will be tested together in a survey question, it is crucial to try to keep the products and their specific function similar, (same brand, same flavor, same type of supplement, etc.) while only testing a difference in their use of marketing tactics.

Different Directions

To widen the understanding from my findings, future researchers could further investigate this topic by selecting a different set of grocery products that may result in other data when surveying a group of women. Additionally, due to limitations I had in my project regarding lack of survey data from the older age range, I believe it would be extremely useful to test this topic exclusively on college students, to whom this subject applies heavily to because this group is newly independent and naïve, allowing higher chances of being lured in by brands who use these marketing tactics to sell products and achieve brand loyalty of young audiences. These results could be compared to the results of my own research, potentially demonstrating a greater or subordinate effect marketing strategies hold on this demographic.

Researchers could further use this study to explore one specific marketing strategy, such as the focus on marketing tactics appealing to a healthy lifestyle; the original basis of my project before I had to redirect my research. This could additionally be done through the same methods, yet coding the usage of one marketing strategy across grocery products instead of ten, and creating a survey questionnaire that focuses all questions on testing the one strategy, giving a better idea to how effective it is comparatively across different products in a grocery store to find under what circumstances the tested strategy is more or less intriguing to a customer.

Works Cited

Chen, L. (2023). Effective Strategies for Attracting and Engaging Target Audiences in Today's Competitive Market. Journal of Education, Humanities and Social Sciences, 16,

141-147.https://scholar.google.com/scholar?hl=en&as_sdt=0%2C10&q=Chen%2C+L.+%282023%29.+Effective+Strategies+for+Attracting+and+Engaging+Target+Audiences+in+Today%E2%80%99s+Competitive+Market.+Journal+of+Education%2C+Humanities+and+Social+Sciences%2C+16%2C+141-147.&btnG=

Crowley, K. (2002). Gender on a Plate: The Calibration of Identity in American Macrobiotics. Gastronomica, 2(3), 37–48. https://doi.org/10.1525/gfc.2002.2.3.37

Hagen, L. (2021). Pretty Healthy Food: How and When Aesthetics Enhance Perceived Healthiness. Journal of Marketing, 85(2), 129–145. https://www.jstor.org/stable/27114287

Ivashkevich, O. (2011). Girl Power: Postmodern Girlhood Lived and Represented. Visual Arts Research, 37(2), 14–27. https://doi.org/10.5406/visuartsrese.37.2.0014

Jones, T., Cunningham, P. H., & Gallagher, K. (2010). Violence In Advertising: A Multilayered Content Analysis. Journal of Advertising, 39(4), 11–36. http://www.jstor.org/stable/25780657

Meriam-Webster. (n.d). Consumerism. In Meriam-Webster.com dictionary. Retrieved November 4, 2025, from https://www.merriam-webster.com/dictionary/feminine

Mohebbi, B. (2014). The art of packaging: An investigation into the role of color in packaging, marketing, and branding. International journal of organizational leadership, 3, 92-102.https://scholar.google.com/scholar?hl=en&as_sdt=0%2C10&q=Mohebbi%2C+B.+%282014%29.+The+art+of+packaging%3A+An+investigation+into+the+role+of+color+in+packaging%2C+marketing%2C+and+branding.+International+journal+of+organizational+leadership%2C+3%2C+92-102&btnG=

Montoya, D. Y., & Scott, M. L. (2013). The Effect of Lifestyle-Based Depletion on Teen Consumer Behavior. Journal of Public Policy & Marketing, 32(1), 82–96. http://www.jstor.org/stable/43305555

Pietrykowski, B. (2004). You Are What You Eat: The Social Economy of the Slow Food Movement. Review of Social Economy, 62(3), 307–321. http://www.jstor.org/stable/29770263

Reimann, M., Zaichkowsky, J., Neuhaus, C., Bender, T., & Weber, B. (2010). Aesthetic package design: A behavioral, neural, and psychological investigation. Journal of Consumer Psychology, 20(4), 431–441. http://www.jstor.org/stable/20778659

Thompson, A. M., & Chad, K. E. (2000). The Relationship of Pubertal Status to Body Image, Social Physique Anxiety, Preoccupation with Weight and Nutritional Status in Young Females. Canadian Journal of Public Health / Revue Canadienne de Sante'e Publique, 91(3), 207–211. http://www.jstor.org/stable/41993147

Tripathi, V. (2016). The role of stealth marketing strategy to manipulate teenagers. Asian Journal of Management Research, 7(2), 189-199.

Walsh, B. (2015). America's Evolution Toward Wellness. Generations: Journal of the American Society on Aging, 39(1), 23–29. https://www.jstor.org/stable/26556093

Wiedenroth, C. F., & Otter, V. (2021). Who Are the Superfoodies? New Healthy Luxury Food Products and Social Media Marketing Potential in Germany. Foods, 10(12), 2907. https://doi.org/10.3390/foods10122907

Zinkhan, G. M. (1994). From the Editor: Advertising, Materialism, and Quality of Life. Journal of Advertising, 23(2), 1–4. http://www.jstor.org/stable/4188923

Do Mega-Events like the LA28 Olympics Create New Long-Term Benefits for Host Cities when Considering Economic, Social, and Environmental Impacts Together?

By Matthew Chiang

Abstract

The Olympic Games stand as one of the world's most prestigious international sporting events. These mega-events are often seen as opportunities for host cities to increase global recognition and generate revenue, but the high costs and infrastructure burdens often offset these benefits. This paper examines whether hosting the Olympic Games creates long-term economic, social, and environmental benefits. To investigate this, a qualitative literature review was conducted to compare past Olympic Games, including the Athens 2004 Olympic Games and the Rio 2016 Olympic Games, and to gather insights for examining the 2028 Olympic Games. The findings indicate that while the Olympic Games can stimulate short-term economic growth and temporary global attention, host cities often experience cost overruns, environmental damage, and social displacements. However, findings show that the LA28 Olympics have strategies, including their sustainability policy and reuse of existing venues, which could help mitigate these pitfalls. Overall, the evidence suggests that although many previous Olympic Games struggled to deliver lasting benefits, LA28's planning approach may provide a more sustainable model for future host cities.

Introduction

The Olympic Games are among the largest and most expansive international sporting events in the world. Host cities often justify the high financial cost with hopes of economic stimulation, global recognition, job creation, and long-term success. Supporters argue that the Olympics can transform a city's infrastructure and increase tourism.

However, historical evidence suggests that hosting the Olympics does not always result in the promised benefits. Several host cities have experienced severe cost overruns, long-term debt, vacant venues, and social displacement. For example, some of the previous Games were criticized for environmental damage, including harm to natural areas caused by the development of infrastructure and other venues necessary for the country to host them. Sometimes, the money spent on the Olympics led to significant debt. These outcomes highlight the delicate balance between financial and social reward and long-term sustainability concerns.

As the 2028 Summer Olympics approach, Los Angeles boasts a new model to tackle these challenges - adaptive reuse of existing venues, controlling costs, and prioritizing sustainability, creating long-term benefits for the community. One example of this that demonstrates their strategy is the reuse of SoFi Stadium and existing college sporting infrastructure. This paper argues that while past Olympic Games have often failed to deliver long-term benefits due to economic inefficiencies and social and environmental costs, the LA28 Olympics represent a more sustainable approach.

Methodology

This paper aims to explore whether mega-events like the Olympics generate long-term economic, social, and environmental benefits for host cities. To investigate this, a qualitative literature review and comparative case study analysis were conducted. A qualitative approach was selected to provide a deeper understanding of the social, economic, and environmental impacts of hosting the Olympics, which cannot be fully captured through quantitative data alone.

Scholarly articles were collected from databases such as Google Scholar and JSTOR using keywords including "Olympic financials 2028," "LA 2028 Olympics," and "impacts of Olympic Games." Non-academic sources, including government reports, International Olympic Committee publications, and reputable news articles, were also used to supplement the analysis with current data and projections. Sources were selected for relevance, credibility, and recent publication to ensure accurate information, particularly for the Los Angeles 2028 Olympic Games. The collected articles were then organized into thematic categories, including economic impact, social effects, and environmental sustainability. These themes were analyzed to identify patterns, trends, and key insights regarding the long-term impacts of hosting the Olympic Games.

To strengthen the analysis, this paper uses a comparative case study approach alongside a qualitative literature review. Three Olympic Games were selected because they represent different approaches to hosting the event and show comparisons to LA28. The Athens 2004 Olympic Games were chosen because they show long-term financial challenges and underutilized infrastructure following the Games. The Rio 2016 Olympic Games were selected because they illustrate the social consequences, including the displacement of local communities, as well as difficulties maintaining newly constructed venues. Finally, the LA28 Olympic Games were chosen because they represent a different hosting strategy that emphasizes the reuse of existing venues, financing, and sustainability. Comparing these three case studies allows this paper to examine how changes in planning, funding, and infrastructure investment may influence the long-term economic, social, and environmental outcomes of hosting the Olympic Games.

One limitation of this methodology is its reliance on secondary sources, including financial projections and planning documents for LA28. Because the 2028 Olympic Games have not yet occurred, projected outcomes may change due to factors such as inflation, natural disasters, changes in tourism demand, or unforeseen economic conditions. As a result, conclusions regarding LA28 should be interpreted as evidence-based projections rather than confirmed outcomes.

Analysis/Results

Bidding Process and the Economics of Hosting

The Olympic Games require host cities to undergo a competitive bidding process organized by the International Olympic Committee (IOC). This process typically lasts about two years and involves multiple phases, including initial applications, detailed planning submissions, and final evaluations. Candidate cities submit proposals outlining infrastructure capacity,

financing strategies, security plans, and long-term development goals. While the bidding process is intended to identify the most capable host city, research suggests that it has become increasingly difficult to attract candidates because of the financial risks associated with hosting the Olympic Games.

Scholarly research indicates that rising costs and financial uncertainty have discouraged many cities from submitting bids. According to the Oxford Olympic Study, the Olympic Games consistently experience cost overruns, with many host cities spending far more than originally projected. The study found that the Olympics have one of the highest average cost-overrun rates of any mega-project, exceeding 150% on average. These findings suggest that cities often underestimate the true costs of hosting, increasing the financial risks associated with the Games and making governments more hesitant to participate in the bidding process. This evidence supports the argument that long-term economic success depends not simply on hosting the Olympics but on effectively managing financial risk.

The decline in bidding competition reflects these growing concerns. Earlier Olympic Games attracted numerous applicant cities, whereas recent Games have seen far fewer candidates. For example, the 2012 Summer Olympics had nine applicant cities, while the 2024 Summer Olympics ultimately had only two remaining candidates, Los Angeles and Paris, after several cities withdrew. This decline of more than 75% suggests that many governments no longer view hosting the Olympics as a worthwhile financial investment. Instead, increasing costs and uncertainty have made the bidding process significantly more financially risky than in previous decades.

Although the International Olympic Committee contributes substantial financial support through broadcasting rights and sponsorship revenue, host governments remain responsible for many of the largest expenses. Olympic funding typically comes from several sources, including IOC contributions, domestic sponsorships, ticket sales, broadcasting agreements, private investment, and public funding from local, state, and national governments (International Olympic Committee, 2024). Public funding is generally used to finance transportation improvements, security, utilities, and other infrastructure projects that remain the responsibility of the host city (SAJE, 2026). As a result, taxpayers often assume much of the financial burden if costs exceed original estimates (Flyvbjerg et al., 2024). This distribution of costs helps explain why cities have become increasingly cautious about pursuing Olympic bids despite the potential for economic growth.

Beyond direct spending, economists emphasize the concept of opportunity cost when evaluating the Olympics. Money invested in Olympic venues, transportation systems, and security cannot simultaneously be spent on other public priorities such as education, healthcare, or affordable housing. Even if the Games generate new economic activity, governments must consider whether those same investments could have produced greater long-term benefits if allocated elsewhere. Research suggests that these financial trade-offs have become another important factor contributing to declining interest in hosting the Olympic Games because they increase the uncertainty surrounding long-term returns.

Overall, the evidence suggests that the financial burden of constructing venues, upgrading infrastructure, and providing security often outweighs the projected economic benefits when planning is ineffective. These financial risks have contributed to declining interest in hosting the Olympic Games and reinforce this paper's argument that long-term success depends on responsible financial planning, cost control, and infrastructure strategies rather than hosting the event alone.

In contrast, successful host cities have demonstrated that the Olympics can contribute to long-term economic benefits when investments are incorporated into broader urban development plans. Barcelona's 1992 Olympic Games are frequently recognized as one of the strongest examples of this approach. Rather than constructing facilities solely for the event, the city integrated Olympic investments into a long-term redevelopment strategy that improved transportation networks, revitalized the waterfront, and supported continued tourism growth after the Games concluded. These investments continued benefiting residents long after the Olympic competition ended, demonstrating that long-term success depends on planning beyond the event itself.

Athens 2004 and Rio 2016 illustrate the opposite outcome. In both cities, many venues became underutilized shortly after the Games, creating ongoing maintenance costs while providing limited economic value (Flyvbjerg et al., 2024). These examples suggest that without a clear long-term strategy for adaptive reuse, Olympic infrastructure can become a financial burden rather than a lasting community asset.

Economic Consideration

Temporary Economic Growth

Hosting the Olympic Games is often promoted as a way to increase economic growth and bring investment to a region. A government report discussed by NBC Los Angeles estimates that the 2028 Summer Olympics could generate between $13 billion and $17 billion in economic activity across Southern California. (Southern California Association of Governments [SCAG], 2025)

The primary projected sources of Olympic revenue include broadcasting and media rights, corporate sponsorships, ticket sales, licensing and merchandise, hospitality programs, and tourism-related spending such as hotels, restaurants, retail, and entertainment. A significant portion of this projected value is expected to come from tourism-related spending, including hotel stays and entertainment, as well as media- and broadcast-related revenue, which has expanded in importance due to digital streaming platforms and global broadcasting rights. Unlike earlier Olympic Games, revenue streams are increasingly shaped by online viewership and sponsorships, which allow economic growth to extend beyond physical attendance.

Long-Term Economic Growth

However, temporary economic boosts from the Games do not necessarily result in

long-term growth. Long term growth depends on whether Olympic investments continue to provide value once the games are complete. Lasting value highly depends on how infrastructure is used after the event. Cities that successfully integrate Olympic venues into everyday community use are more likely to experience sustained economic benefits than those that construct facilities with limited long-term demand.

Barcelona's successful hosting of the 1992 Summer Olympics is frequently cited as one of the strongest examples of long-term Olympic success. Investments in transportation networks, waterfront redevelopment, and urban renewal transformed previously neglected areas into thriving residential and tourist destinations, while tourism continued growing after the Games (Brunet, 1995).

Cost Overruns/Debt

The example of Barcelona shows long term growth is possible, but cost overruns and overloading debt can often offset these potential gains. Despite projected economic gains, past Olympic Games consistently demonstrate that cost overruns are the norm, often leading to long-term financial strain for host cities. For example, the 2016 Summer Olympics in Brazil were originally estimated at about $2.8 billion but ultimately cost more than $20 billion, leaving the city with significant debt and underused venues. Similarly, the 2020 Summer Olympics experienced massive increases from its original budget, with total costs rising to nearly $13 billion, over six times the initial estimate. These examples demonstrate that initial Olympic budgets underestimate the true financial demands of hosting, increasing the likelihood that long-term costs will outweigh projected economic benefits.

The Oxford Olympics Study further supports this pattern by concluding that cost overruns occur consistently across Olympic Games due to expanding infrastructure projects, transportation improvements, and heightened security requirements. Because many of these expenditures are directed toward facilities and services that generate little direct revenue, host cities often struggle to recover their investments after the Games conclude. This evidence suggests that the financial success of the Olympics depends less on projected revenue and more on a city's ability to effectively control costs throughout the planning and implementation process.

Research also indicates that these financial risks disproportionately affect developing countries. A study published by the *National High School Journal of Science (NHSJC)* explains that the high costs of bidding, infrastructure development, and event operations create significant barriers for countries with fewer financial resources. Although hosting the Olympics can improve international recognition and attract investment, the study argues that the current hosting model favors wealthier nations with existing infrastructure and greater financial capacity. As a result, many developing countries face greater financial risk because they must invest substantially more in new construction and public infrastructure before the Games can even begin. This finding reinforces the argument that long-term Olympic success is closely tied to existing infrastructure and responsible financial planning rather than the event itself.

Additional analysis from an AECbytes viewpoint article on Olympic-related development discusses how large-scale events influence architectural design and long-term infrastructure decisions in host cities. Without a clear post-Olympic strategy, expensive facilities can become financial liabilities instead of long-term community assets.

In conclusion, these studies show an economic picture: while the Olympics can generate billions in short-term activity and global attention, long-term financial outcomes depend heavily on cost control, infrastructure planning, and the ability of host regions to manage risks that have not yet been mitigated in past games.

Social/Community Welfare & Environmental Sustainability

Social Challenges

Hosting the Olympics can create significant social challenges, particularly in housing affordability, community displacement, and public services. Olympic-related redevelopment often increases demand for land while reducing the supply of affordable housing, causing property values, rents, and home prices to rise. In some cases, governments also redirect public funding toward Olympic infrastructure, transportation, and security, placing additional financial pressure on taxpayers. These changes can contribute to gentrification, where redevelopment increases living costs and forces lower-income residents to relocate. Rather than benefiting all members of the community equally, Olympic investment can disproportionately affect vulnerable populations, limiting the long-term social benefits of hosting the Games.

The 2016 Summer Olympics in Rio illustrate these challenges. Preparations for the Games displaced thousands of residents to make room for new roads, buildings, and Olympic facilities. Many families were forced to move farther from their workplaces, schools, and established communities. Although these projects supported Olympic development, they also disrupted existing neighborhoods and created lasting social consequences. This example demonstrates that economic development associated with the Olympics can come at a high social cost when community needs are not prioritized.

These challenges often continue after the Olympic Games conclude. Rising housing costs associated with redevelopment and increased tourism can make cities less affordable for long-term residents, particularly lower-income households. In addition, transportation systems expanded for the Olympics may require costly long-term maintenance, placing continued financial demands on local governments.

Environmental Challenges

Beyond their social impacts, the Olympic Games also create significant environmental challenges. Large-scale construction projects for venues and supporting infrastructure can contribute to deforestation, habitat destruction, and increased carbon emissions. The 2014 Winter Olympics in Sochi, for example, were criticized for environmental damage caused by construction in protected natural areas. This case demonstrates that efforts to maximize economic

development may sometimes come at the expense of environmental sustainability, limiting the long-term benefits of hosting the Games.

Transportation also contributes substantially to the Olympics' environmental footprint. Millions of athletes, spectators, officials, and media personnel travel internationally by air, making transportation one of the largest sources of Olympic-related carbon emissions. In addition, construction requires large quantities of concrete and steel, materials that produce significant carbon emissions during manufacturing. Recent estimates indicate that modern Summer Olympic Games can generate between 3 and 6 million metric tons of carbon dioxide equivalent throughout their lifecycle, with transportation and construction accounting for the majority of emissions.

Another long-term environmental concern is the existence of vacant or underutilized Olympic facilities after the Games conclude. Several venues constructed for the Athens 2004 Olympics later fell into disrepair because maintenance costs exceeded their long-term value. Similar patterns occurred following the Rio 2016 Olympics, where several facilities became underutilized only a few years after the Games ended. These facilities represent not only financial waste but also inefficient use of land, construction materials, and public resources. The continued existence of vacant venues demonstrates that environmental sustainability depends on long-term planning for infrastructure rather than construction for a single event.

Overall, the evidence suggests that while the Olympic Games can generate economic opportunities and international recognition, these benefits must be evaluated alongside their social and environmental consequences. The experiences of previous host cities demonstrate that without careful planning, the long-term costs of displacement, environmental damage, and underutilized infrastructure can outweigh the projected benefits. These findings reinforce this paper's argument that lasting Olympic success depends on balancing economic development with responsible social and environmental planning.

LA28 Analysis

The Los Angeles 2028 Olympic Games plan to use a different approach from many previous host cities by focusing on reducing costs and making better use of existing resources. Rather than building many new venues, LA28 plans to use existing stadiums, arenas, and university facilities across Southern California. This strategy is intended to lower construction costs while reducing the risk of abandoned venues, displacement, and wasteful spending.

Profitability and Funding

LA28 expects to generate revenue through several sources, including broadcasting rights, sponsorships, ticket sales, licensing, merchandise, and tourism. A government report estimates that the Games could generate between $13 billion and $17 billion in economic activity across Southern California (Southern California Association of Governments [SCAG], 2025). Although the operating budget is primarily funded through private revenue, public agencies will still pay for areas such as transportation improvements, public services, and security. Security is expected

to be one of the largest public expenses due to the event's size. The projected economic activity demonstrates that LA28 could provide substantial short-term economic benefits to Southern California in many different ways. However, these benefits do not automatically demonstrate long-term economic success. Much of the projected activity will occur during the period near the Games, meaning that the long-term benefit will depend on whether Olympic-related investments and business activity continue to benefit them after 2028.

Cost-Benefit Analysis

Compared with previous Olympic Games, LA28 appears to face lower financial risk because much of the required infrastructure already exists. By limiting new construction, organizers seek to reduce the cost overruns that affected cities such as Athens and Rio while also decreasing future maintenance costs associated with underutilized facilities. If expenditures remain close to projected budgets, the anticipated economic benefits from tourism, employment, and business activity are more likely to outweigh overall costs. This approach reflects one of the central findings of this paper: long-term Olympic success depends on effective planning and responsible financial management rather than hosting the Games alone.

Monetary Strategy

A key component of LA28's strategy is the adaptive reuse of existing infrastructure. Most competition venues already serve professional sports teams, universities, or entertainment events, allowing Olympic investments to build upon facilities that already provide value to local communities. LA28's updated venue plan explains that using existing venues can reduce the cost and complexity of temporary stadium construction and operations while also providing financial and operational benefits (Los Angeles Organizing Committee for the 2028 Olympic and Paralympic Games, 2024). This approach contrasts with previous host cities that invested heavily in new stadiums that later became underutilized. Similar to Barcelona's successful integration of Olympic investments into broader urban development, LA28 seeks to maximize the long-term value of existing infrastructure rather than constructing facilities designed primarily for the Games.

Diversified Revenue Streams

Unlike many earlier Olympic Games, LA28 projects revenue from a diverse combination of television and streaming rights, ticket sales, sponsorship agreements, merchandise, licensing, hospitality programs, and tourism spending. Diversifying revenue sources reduces reliance on any single income stream and may strengthen the financial stability of the Games. This strategy also reflects changes in the global sports industry, where digital broadcasting and international media rights contribute substantially to Olympic revenue.

LA28's revenue strategy differs from earlier Olympic Games because modern Olympic organizers can rely on both traditional television broadcasting and digital streaming platforms to reach global audiences. This creates additional opportunities for media rights, sponsorships, and

advertising compared with earlier Games that depended more heavily on traditional television audiences. A broader range of revenue sources may reduce financial risk as well, because weaker performance in one category could be balanced by revenue from another.

Cost Reduction Techniques

Organizers also plan to reduce costs by using temporary venues where permanent facilities are unnecessary and by allowing several venues to host multiple sporting events. (Los Angeles Organizing Committee for the 2028 Olympic and Paralympic Games). These strategies reduce construction expenses while limiting the risk that expensive facilities will become vacant after the Olympics conclude. The emphasis on adaptive reuse and temporary infrastructure directly addresses many of the financial and environmental shortcomings identified in previous Olympic Games.

Strategy for Long-Term Benefits

LA28 aims to create benefits that continue after the Olympics are over. By improving transportation, creating temporary jobs, attracting tourism, and using facilities that already serve local communities, organizers hope the Games will provide lasting economic and social value without leaving behind unused venues.

Future Use of Artificial Intelligence

Future Olympic organizers may also benefit from using artificial intelligence (AI) to improve planning and reduce many of the challenges identified throughout this paper. Data from previous Olympic Games could be used to develop AI models that project cost overruns, estimate security needs, optimize transportation systems, and improve resource allocation. These predictive tools could help organizers make more informed financial and operational decisions before the Games begin.

Although AI cannot eliminate every uncertainty associated with hosting the Olympics, it may improve forecasting and reduce some of the financial and logistical risks that have contributed to cost overruns in previous host cities. If used alongside careful planning and sustainable infrastructure strategies, AI could support future Olympic organizers in creating stronger long-term economic, social, and environmental outcomes. However, because AI is still developing, its effectiveness for future Olympic Games remains uncertain.

Social/Environmental

LA28 also differs from previous Olympic Games through its emphasis on using existing venues rather than constructing new permanent facilities. According to LA28's Impact and Sustainability Plan, organizers will make use of existing venues and prioritize reuse across the Games' temporary infrastructure in order to minimize the environmental impact of the Olympics (Los Angeles Organizing Committee for the 2028 Olympic and Paralympic Games, 2025). LA28 will also prioritize clean energy and transportation, including purchasing 100% renewable

electricity for venues and hosting a transit-first Games. These strategies could reduce some of the environmental impacts associated with Olympic construction and transportation.

The use of existing venues may also create social benefits because these facilities already serve local communities and can continue to be used after the Olympics. LA28 will build on infrastructure that already has a purpose. LA28's plan also includes a goal of reusing or recycling at least 90% of materials used for temporary infrastructure and venue overlays, which could reduce the amount of waste produced by the Games.

However, LA28's environmental strategy has also received criticism. Strategic Actions for a Just Economy (SAJE) reports that California enacted legislation exempting temporary Olympic and Paralympic venues from the California Environmental Quality Act (CEQA). These exemptions create concerns because LA28's use of temporary infrastructure does not mean that the Games will have no environmental impact. SAJE notes that temporary Olympic facilities will still include spectator seating, bathrooms, training facilities, media centers, fan zones, and other infrastructure on public land (Strategic Actions for a Just Economy, 2026). The organization also raises concerns about the potential loss of public green space, including the proposed temporary venues.

This creates a more complicated environmental picture for LA28. On one side, the use of existing venues, renewable electricity, public transportation, and material reuse could reduce some of the environmental problems associated with previous Olympic Games. On the other hand, exemptions from environmental regulations and the construction of temporary facilities could create environmental and social costs for communities. Therefore, LA28's approach may reduce some of the risks associated with hosting the Olympics, but it does not eliminate them.

Conclusion

This paper examined whether hosting the Olympic Games creates long-term economic, social, and environmental benefits for host cities. The evidence suggests that while the Olympics can generate short-term economic activity, tourism, and global recognition, many host cities have also experienced high financial costs, environmental damage, and social challenges. The case studies of Athens 2004 and Rio 2016 demonstrate how poor long-term planning can leave cities with debt, underused infrastructure, and lasting impacts on communities.

In contrast, the Los Angeles 2028 Olympic Games represent a different approach. By reusing existing venues, relying on diversified sources of revenue, and having a robust sustainability plan, LA28 aims to reduce many of the risks experienced by previous host cities. Future host cities can learn from LA28 by prioritizing existing infrastructure, and integrating social and environmental considerations from the beginning, instead of being an afterthought.

Overall, the findings indicate that hosting the Olympic Games does not automatically produce long-term benefits. Instead, success depends on careful planning, responsible financial management, sustainable infrastructure, and a clear strategy for how Olympic investments will continue to benefit communities after the Games have ended.

Works Cited

Ballester, Patrice. “Barcelona in the Face of Globalization, How to Think of the City Through the Organization and Evaluation of Major Events?” arXiv.Org, 2022, https://arxiv.org/abs/2212.13901.

Budzier, Alexander, and Bent Flyvbjerg. “The Oxford Olympics Study 2024: Are Cost and Cost Overrun at the Games Coming Down?” SSRN Electronic Journal, no. July 2024, 2024, https://doi.org/10.2139/ssrn.4850085.

deMause, Neil. Damned If You Do, Damned If You Don’t: The Risks to L.A. Of Hosting or Withdrawing from the 2028 Olympics. Feb. 2026, https://www.saje.net/wp-content/uploads/2026/02/Olympics-Report-FINAL-2.4.26.pdf.

Ferran, Brunet. “Economy of the 1992 Barcelona Olympic Games.” Olympics.Com, 2026, https://library.olympics.com/Default/doc/SYRACUSE/71699/economy-of-the-1992-barcelona-olympic-games-ferran-brunet?_lg=en-GB.

Georgiev, Petko, et al. “Where Businesses Thrive: Predicting the Impact of the Olympic Games on Local Retailers Through Location-Based Services Data.” arXiv.Org, 2025, https://arxiv.org/abs/1403.7654.

IEA. “IEA - the Global Energy Authority.” Iea.Org, 2023, https://www.iea.org.

International Olympic Committee. “IOC - Funding.” International Olympic Committee, 27 Apr. 2021, https://olympics.com/ioc/funding.

Kim, Katie. “Bot Verification.” Nhsjs.Com, 2025, https://nhsjs.com/wp-content/uploads/2025/04/The-Economics-of-the-Olympic-Games-Analyzing-the-Issues-that-Bar-Developing-Countries-from-Hosting.pdf.

“LA 2028.” La28.Org, https://la28.org. Accessed 27 Aug. 2026.

Linstroth, Tommy. “How to Make the 2028 Los Angeles Olympics a Long-Term Benefit for the City’s Infrastructure: AECbytes Viewpoint.” Aecbytes.Com, 2024, https://www.aecbytes.com/viewpoint/2024/issue_118.html.

Lloyd, Jonathan, and Mekahlo Medina. “New Report Estimates the Economic Impact of 2028 LA Olympics for 5 Counties.” NBC Los Angeles, NBC Southern California, 5 Dec. 2025, https://www.nbclosangeles.com/olympics/2028-los-angeles/olympics-la-county-economy/3811740/.

“OWL Home Page.” Library.Olympics.Com, https://library.olympics.com/. Accessed 27 Aug. 2026.

Pinto, Pedro Janela, and Gustavo Lopes dos Santos. “Olympic Waterfronts: An Evaluation of Wasted Opportunities and Lasting Legacies.” Sustainability, vol. 14, no. 4, Feb. 2022, p. 1968, https://doi.org/10.3390/su14041968.

“Southern California Association of Governments - SCAG.” Southern California Association of Governments, https://scag.ca.gov/. Accessed 27 Aug. 2026.

UNEP. “UNEP - UN Environment Programme.” UNEP - UN Environment Programme, 2021, https://www.unep.org/.

“Venues.” *La28.Org*, 2024, https://la28.org/en/games-plan/venues.html.

Evaluating the Effects of Spain's National Public Transport Pass on Congestion-Related Economic Losses in Madrid and Barcelona By Priya Barke

Abstract

This literature review examines how Spain's National Public Transport Pass, the EUR 60 adult and EUR 30 youth Abono Unico introduced in January 2026, is likely to influence congestion-related economic losses in Madrid and Barcelona. Synthesising evidence from Spain's existing policy mix (the Madrid Central low-emission zone, the 2022-2023 fare reductions, urban speed-limit reductions, and Barcelona's AI-based traffic forecasting) alongside international experience with flat-fare and congestion-pricing schemes, it assesses four channels of loss, specifically commuter time loss, missed business opportunity, the spatial distribution of growth, and environmental and health externalities. The evidence indicates that fare integration reliably increases ridership and delivers large, progressive savings, but produces weak and partly transient substitution away from private cars when deployed alone. Drawing on traffic-flow physics and urban economics, the review argues that the Pass can meaningfully reduce commuter time loss, fuel use, and CO2 emissions, chiefly through improved reliability and accessibility, and most strongly for lower-income, transit-rich districts, provided it is paired with complementary pricing, speed-management, and predictive-management measures. Its ultimate congestion dividend remains an empirical question requiring rigorous post-implementation evaluation.

1. Introduction

Urban traffic congestion imposes costs that extend well beyond the immediate frustration of delayed commuters. In dense metropolitan areas, the aggregate burden encompasses lost productive time, impaired freight and logistics performance, elevated air pollution with measurable mortality consequences, and a systematic redistribution of economic opportunity away from those least able to absorb it. Lower-income households, who are disproportionately exposed to traffic-related air pollution and least able to adjust their travel behaviour in response to congestion, bear a larger welfare burden than aggregate city-level statistics suggest.

Spain's two largest cities illustrate the scale of the problem with particular clarity. According to the INRIX 2025 Traffic Scorecard, Madrid ranks fifth among the most congested cities in Europe, with commuters losing over 98 hours annually to peak-period delays and total congestion-related economic losses exceeding EUR 6.2 billion (INRIX). Traffic-related emissions account for approximately 78% of local source emissions for NO2 in Madrid, contributing to an estimated 900 premature deaths per year in the capital and over 1,000 in Barcelona, representing between 3% and 7% of overall premature mortality in each city respectively. In 2023, official figures recorded 101,306 road accidents across Spain, with both casualty collisions and fatalities rising relative to 2022 (National Road Safety Observatory).

Over the preceding decade, Spain has responded to these pressures through a succession of regulatory and demand-side interventions. Madrid's low-emission zone, implemented in

November 2018, reduced NO2 concentrations within the restricted area by approximately 19% and car traffic by 16%, though both effects proved partly transient as fleet renewal and route displacement partially offset the initial gains (Galdon-Sanchez et al.. Tassinari). Spain's 2022-2023 public transport fare reductions generated a substantial increase in ridership and shifted the subsidised mode share from 11.9% to 15.4%, but produced weak substitution away from private vehicle use (Ministry of Transport and Sustainable Mobility). Barcelona has deployed AI-based congestion forecasting that projects a 20% reduction in congestion, though this projection has not yet been confirmed through post-implementation evaluation (Garcia et al.). Each of these interventions moves the relevant indicator in the desired direction but falls short of delivering durable city-wide congestion relief, for reasons this review examines in detail.

Against this background, Spain introduced the National Public Transport Pass, the Abono Unico, in January 2026. Priced at EUR 60 per month for adults and EUR 30 for those under 26, the pass provides unlimited travel on Cercanias and Rodalies commuter rail, medium-distance regional trains, and state-run intercity buses, and is explicitly modelled on Germany's Deutschlandticket (Hughes). The government estimates annual savings of approximately EUR 750 for a typical commuter and projects that around two million people will benefit from reductions in their monthly transport expenditure of up to 60% (Martinez). Whether these gains translate into meaningful reductions in congestion-related economic losses remains an open empirical question.

This review asks how Spain's National Public Transport Pass could influence congestion-related economic losses in Madrid and Barcelona, particularly in terms of commuter time loss, missed business opportunity, the spatial distribution of economic activity, and environmental and health externalities.

Answering this question matters for policy design. Congestion in high-productivity cities is not straightforwardly a problem to eliminate, as some degree of density-driven traffic reflects the agglomeration benefits that make large cities economically productive. The relevant policy objective is therefore to reduce the welfare losses imposed by congestion while preserving the density and accessibility that sustain urban productivity, and to ensure that both the costs of congestion and the benefits of relief are distributed equitably. Despite extensive research on individual congestion instruments, there remains a gap in the literature regarding the monetised welfare effects of nationwide fare integration policies within complex urban transport systems. As the Abono Unico has only recently been implemented, empirical evidence on its long-run congestion and welfare effects is necessarily limited, and a principal purpose of this review is to frame the methodology for the post-implementation evaluation that will resolve the key uncertainties.

This paper argues that Spain's National Public Transport Pass is likely to reduce congestion-related economic losses primarily through improved reliability and accessibility, and most strongly for lower-income, transit-rich districts, provided it is paired with complementary pricing, speed-management, and predictive-management measures. Deployed in isolation, the more probable outcome is a significant equity and ridership gain accompanied by only a modest

congestion dividend, as the international and Spanish evidence consistently shows that fare integration alone produces weak substitution away from private vehicles.

2. Existing Policy Interventions Addressing Congestion in Spain

Spain has previously administered regulatory restrictions and financial incentives in order to aid in reducing congestion. Rather than evaluating the new policy in isolation, this research considers how Spain's National Public Transport Pass is complemented by Spain's existing policies, focusing on their combined effects.

2.1. Low-Emission Zones and Traffic Restrictions

Madrid has a long history of violating NO2 limit values defined by European Union environmental regulations, specifically Directive 2008/50/EC, which sets a maximum annual average of 40 micrograms per cubic metre. The Low-Emission Zones (LEZs) and traffic restrictions were created primarily to combat severe urban air pollution, improve public health, and reduce greenhouse gas emissions. Implemented on 30 November 2018 in the central district (approximately 5 km2), it restricted access for high-polluting vehicles based on Spain's ecological label system (B, C, ECO, Zero) and had fines up to EUR 90 for unauthorised entry (Galdon-Sanchez et al.).

Spain's ecological label system assigns one of four stickers to every vehicle, determining access rights, parking costs, and enforcement outcomes across all LEZ-implementing cities. B-labelled vehicles (petrol Euro 4 and diesel Euro 6 models) face restricted access in most city centres. C-labelled vehicles (petrol Euro 6d and diesel Euro 6d-TEMP models) are generally permitted entry but may face temporary restrictions during high-pollution episodes. ECO-labelled vehicles (hybrids with limited electric range and gas-fuelled Euro 6d models) are permitted access to LEZs with reduced parking fees and, in some municipalities, bus lane access. Zero-emission vehicles receive unrestricted LEZ access along with free or discounted parking and, in some cities, exemptions from congestion-related charges (Galdon-Sanchez et al.). Enforcement is camera-based, with automated systems that verify licence plates against the national environmental sticker registry at LEZ entry points. Under Spain's Climate Change Law, all municipalities with more than 50,000 residents are required to establish a LEZ, a mandate that encompasses over 150 cities across the country (Ministry for the Ecological Transition and the Demographic Challenge).

2.1.1. Environmental Costs, Car Traffic and Air Quality

Galdon-Sanchez et al. estimated the effects on traffic flow and NO2 levels using the following difference-in-differences regression model in Madrid Central (MC).

$$Yswy = B\ MCswy + d\ X'swy + usw + twy + eswy$$

The dependent variable Yswy represents the traffic or air pollution outcome measured at station s in week w of year y. MCswy is a binary indicator equal to 1 if station s is located within the MC area during a given year-week in which MC is in effect. The vector X'swy includes

control variables for meteorological conditions, with coefficient d capturing the effect of weather on air pollution levels. The term usw denotes station-week fixed effects accounting for season-specific patterns, while twy controls non-parametrically for time trends and year-week-specific shocks. The error term eswy reflects unobserved factors and may be serially correlated at the station level.

Galdon-Sanchez et al. found that the average number of cars fell by 16.1% (56.5 cars), and the LEZ reduced the frequency of road segment usage by cars by 21%. However, once penalties were put in place near the end of March 2019, the effects were partly reversed, the time a road segment was occupied was then only 10.6% smaller than its level prior to Madrid Central. As congestion eased, downtown travel became relatively more attractive for drivers with compliant vehicles, inducing substitution effects that offset part of the initial traffic decline. The LEZ caused a decrease of 18.6% in NO2 in the restricted area immediately after the policy was introduced, with a further drop to as much as 41.4% once illegal entry was penalised (Galdon-Sanchez et al.).

Table 1. Summary of selected empirical studies and policy interventions on congestion, emissions, and mobility.

Study	Intervention / policy	Location	Method	Key quantitative results
Galdon-Sanchez et al. (2023)	Low-emission zone (Madrid Central)	Madrid	Difference-in-differences	Cars -16.1%, NO2 -18.6% (to -41.4% after fines), outside-buyer retail spending -20.6%, online spending +12.1%
Tassinari (2024)	Low-emission zone (Madrid Central)	Madrid	High-frequency pre/post analysis (~4,000 sensors)	Flow -8.1%, density -17.7%, +2-6% displacement outside, metro entries +5.4%, effect fades ~7 months
Monzon and Guerrero (2004)	Road-transport emissions costing	Madrid	Emission inventory x unit damage cost	Local pollution cost ~0.4% of regional GDP, ~30,650 m pesetas/yr (1992 prices), NOx dominant
Garcia et al. (2024)	AI-based congestion forecasting	Barcelona	ARIMA vs. XGBoost + SHAP on Open Data BCN	ARIMA assumptions violated, XGBoost captures spatio-temporal structure, supports demand management / variable tolls
Rodriguez-Rey et al. (2021)	Congestion-emissions link	Barcelona	Traffic-emissions modelling	Up to +65% NOx under congested, stop-and-go conditions
Loder et al. (2024)	Flat nationwide fare (EUR 9/month)	Germany	Three-wave panel + GPS travel diary	Regular PT use 29% to 38% (32% after), only ~10-20% replaced a car trip, PT distance -37% post-expiry
Spanish fare reductions (2022-23)	Free Cercanias/Rodalies + 30-50% multi-trip discounts	Spain	Government / operator administrative data	Subsidised-mode share 11.9% to 15.4%, ~1.7 m free commuter passes, ~1 Mt CO2 and ~350 Ml fuel saved (2023 est.)
Congestion charges (Santos, Eliasson)	Congestion pricing (cordon / area)	London / Stockholm / Milan	Charging schemes, Stockholm reversal design	London entering traffic -18%, congestion -30%, Stockholm trips -22%, Milan accidents -21.3%, PM10 -13 to -15%

Sources are summarised from the studies reviewed in this paper, figures are reported as stated in the original sources and are not strictly comparable across studies owing to differences in method, period, and geography.

2.1.2. Effect on Consumption Spending

Beyond traffic and air quality, the Madrid Central scheme produced measurable distortions in consumer behaviour, illustrating that congestion policies carry private costs as well as social benefits. Galdon-Sanchez et al. document that brick-and-mortar spending by buyers residing outside the restricted perimeter fell by roughly 20.6%, accompanied by an approximately 3% daily decline in the number of active merchants and payment cards used inside the zone, representing an estimated annual reduction in physical retail activity on the order of EUR 24 million. The sectors most exposed were fashion, automotive services, and travel, categories that depend disproportionately on car-borne, non-resident shoppers. Online spending within the affected area rose by about 12.1% and the number of digital transactions by 9.4%, indicating that a substantial share of foregone in-store purchases migrated to e-commerce rather than disappearing. This substitution from physical to digital channels suggests that consumers adapt quickly to transaction-cost shocks and that small firms in central Madrid lacking an online presence bear the heaviest burden.

The largest spending reductions were concentrated among wealthier districts and car-dependent postal codes, whereas lower-income, transit-served neighbourhoods were comparatively less affected. Galdon-Sanchez et al. interpret the consumer-spending decline as the private cost incurred to correct a pollution externality. Residents inside the zone gained cleaner air, quieter streets, and improved health outcomes, including an estimated 88 lives saved annually and approximately EUR 3.4 million in avoided medical costs, so that the net efficiency effect of the policy is plausibly positive even where retail revenue fell. For the National Public Transport Pass, this evidence is instructive in two ways. First, because the Pass lowers the cost of reaching central districts by public transport rather than restricting car access, it may partially reverse the access penalty that LEZs impose on non-resident consumers, supporting rather than depressing downtown commerce. Second, it underscores that the welfare consequences of any mobility intervention are unevenly distributed and must be evaluated at the neighbourhood level rather than in aggregate.

2.2. 2022-2023 Fare Reductions

Spain's 2022-2023 fare reduction programme was the most extensive demand-side transport intervention in the country's recent history, emerging primarily as a macroeconomic response to the cost-of-living crisis rather than as a purpose-built congestion policy. Enacted through Royal Decree Law 25/2022 on 25 June 2022, the Council of Ministers approved a 50% reduction on multi-trip tickets for Renfe rail services, encompassing Cercanias commuter services, Media Distancia routes under 300 kilometres, and Avant high-speed services for regular travellers. Cercanias and Rodalies commuter services were subsequently made free of charge for multi-trip pass holders through to the end of 2022, and the subsidy was extended to state-owned long-distance bus services from January 2023. Between September 2022 and December 2024, the Spanish Ministry of Transport allocated EUR 3.4 billion to finance these measures, of which

EUR 1.41 billion was attributable to the free commuter and medium-distance rail component alone (Ministry of Transport and Sustainable Mobility).

2.2.1. Ridership and Modal Share Effects

The programme produced measurable and statistically significant increases in public transport ridership. The subsidised modes' share of total journeys rose from 11.9% in 2022 to 15.4% in 2023, while the road share declined from 84.1% to 79.8% over the same period (Ministry of Transport and Sustainable Mobility). By 2024, the cumulative effect had produced a 33% increase in public transport trips relative to 2022 levels. The Spanish Ministry of Transport projected that the free train measure alone was expected to prevent approximately one million tonnes of CO2 in 2023. These figures, however, must be interpreted with caution, as they are largely model-derived projections that do not fully account for induced demand or the counterfactual trajectory of modal shares absent the subsidy.

2.2.2. The Induced Demand Problem and Limits of Car Substitution

Despite the headline ridership gains, the 2022-2023 fare reductions illustrate a fundamental tension at the core of flat-fare policy. Strong price elasticity in public transport demand does not automatically translate into equivalent reductions in private vehicle use. As documented by Loder et al. in Table 1, the German 9-Euro-Ticket saw regular public transport use rise from 29% to 38% of trips during its active period, yet only 10-20% of these additional journeys replaced a car trip. The dominant effect was mobility generation rather than modal substitution. Broader comparative evidence from the ITF's synthesis of fare-free experiments found that additional trips typically represented induced demand rather than substitution from private car use (International Transport Forum). For Spain specifically, works focused on Madrid by Matas and Jalon et al. identify a structural limitation, namely the low cross-price elasticity of demand for private transport with respect to the price of public transport reduces the effectiveness of fare subsidies as a congestion-reduction mechanism (Rubensson).

The observed road modal share decline from 84.1% to 79.8% is also partly attributable to factors independent of the fare intervention, including the decline of remote working and the overall strength of the Spanish economy (Smith). Furthermore, where petrol prices fell in real terms during the same period, the implicit price advantage of public transport relative to driving was further compressed, weakening the substitution incentive (Litman, "Transit Price Elasticities").

2.2.3. Distributional and Fiscal Dimensions

The equity implications of the 2022-2023 programme are complex and partly in tension with one another. The policy delivered large absolute savings to lower-income commuters who were already dependent on public transport, but flat-rate, non-targeted subsidies tend to deliver proportionally larger absolute benefits to higher-income groups who travel more frequently and over longer distances. Borjesson et al. ("Distributional effects") found that low- and

medium-income households can experience welfare losses from poorly designed subsidy structures. The fiscal cost was also substantial relative to its congestion dividend. At EUR 3.4 billion allocated between September 2022 and December 2024, the programme achieved measurable modal share gains but produced little documented reduction in urban traffic volumes or congestion indices in Madrid and Barcelona. No peer-reviewed evaluation has, at the time of writing, estimated the monetised congestion relief attributable to the fare reduction programme using traffic-flow or travel-time data, representing a significant gap in the evidence base.

2.2.4. Implications for the National Public Transport Pass

The 2022-2023 fare reductions constitute the most direct precursor to the National Public Transport Pass introduced in January 2026. The precursor programme demonstrates that price reductions of this magnitude generate robust ridership responses, particularly among regular commuters. The Pass, at EUR 60 per month for adults and EUR 30 for those under 26, will extend these gains to a broader multi-modal network, consolidating the Cercanias, regional rail, and bus components under a single integrated fare instrument. Government projections estimate annual savings of approximately EUR 750 for a typical commuter and anticipate an estimated 120 million car trips shifted to rail and bus services. These figures are indicative rather than empirical, but they reflect the same behavioural elasticity patterns observed during 2022-2023. A substantial portion of the additional journeys generated by fare reductions represents induced demand rather than modal substitution, however, and the structural barriers to car abandonment, vehicle ownership costs being largely sunk, car use conferring time and flexibility advantages, and parking remaining cheap relative to many European comparators, are not addressed by the Pass in isolation. Its ultimate effectiveness will depend on whether it is paired with complementary measures that alter the relative generalised cost of driving, such as congestion pricing, parking management, or urban access restrictions. Where temporary fare discounts may encourage marginal trip-making without inducing durable modal habit change, the longevity of the Pass as a structural policy creates conditions for longer-run shifts in residential location, vehicle ownership, and commuting patterns (Loder et al.).

2.3. Speed-Limit Reductions and Traffic-Flow Physics

A distinct and often underappreciated lever for managing congestion is the speed limit, whose effects are best understood through the physics of traffic flow. The macroscopic state of a road is described by the fundamental diagram, which relates traffic flow q (vehicles per hour) to density k (vehicles per kilometre) and mean speed v through the identity q = k * v. In the canonical Greenshields formulation, flow is a concave function of density, q(k) = vf * k * (1 - k / kjam), where vf is the free-flow speed and kjam the jam density. The diagram has two regimes, namely an uncongested branch on which flow rises with density, and a congested branch on which flow falls as density increases, so that beyond the capacity-maximising apex, adding vehicles reduces the number that actually pass a given point per unit time. Travel time responds non-linearly to the volume-to-capacity ratio, as captured by the Bureau of Public Roads function,

T = T0 * [1 + a(V/C)B], with typical parameters a = 0.15-0.5 and B = 4, so that delay escalates steeply as a link approaches capacity.

By harmonising speeds across a road section, lower speed limits compress the variance of headways and suppress the perturbation amplitude that would otherwise trigger stop-and-go waves. A phantom jam occurs without the existence of a physical bottleneck and is caused by the imperfect driving style of road users under metastable traffic conditions, previous studies have shown that dynamic speed limits are a successful instrument for stabilising traffic flow and dissolving formed phantom jams (Suijs et al.). A statutory urban speed limit operates through the same mechanism at a population level. This matters directly for the National Public Transport Pass. By shifting some vehicle trips onto rail and bus, the Pass lowers road density, which moves congested links toward the uncongested branch of the fundamental diagram and makes the stabilising effect of speed management more powerful. The two instruments reinforce each other through the same physical mechanism.

2.3.1. Spain's National Reform and Barcelona's Superblocks

In Spain, the November 2020 law imposing a default speed limit of 30 km/h on single-carriageway two-lane urban roads represented a shift away from the historical regulatory priority given to motor vehicles (Gonzalo-Orden et al.). According to the DGT, the measure affects between 60% and 70% of roads nationally. Barcelona has gone considerably further through its superblocks programme. Superblocks consist of grids of blocks approximately 400 m by 400 m, with vehicles restricted to outer perimeter roads and interior speeds capped at 10 km/h for residents and delivery drivers (European Bank). The city has simultaneously expanded cycle lanes and initiated an orthogonal-grid bus network using main thoroughfares only. The 2024 Urban Mobility Plan targets a reduction in private vehicle journeys from 26.04% to 18.48% alongside the eventual creation of 503 superblocks citywide (Postaria). For the National Public Transport Pass, this architecture is directly relevant. The superblocks perimeter network is explicitly designed with segregated bus lanes, meaning that any additional bus ridership generated by the Pass benefits from a dedicated running environment. The Pass and the superblocks programme are, in this sense, designed for each other.

2.3.2. Empirical Evidence and Implications for the Pass

A systematic review of 40 European cities found that speed-limit reductions produced an average 18% decrease in emissions, a 2.5 dB reduction in noise, a 7% fall in fuel consumption, and a 23% average decrease in road crashes, with no documented increase in travel time or overall congestion for vehicular traffic in the majority of cities assessed (Yannis and Michelaraki). In Bilbao, where 100% of streets were converted to 30 km/h by September 2020, atmospheric pollution fell by 11.4% in NO2, 17.1% in NOx, and 19.1% in PM10, while bus services became faster, arriving on average four minutes ahead of schedule, and delivery operations were unaffected (Dragonetti). Bus schedule reliability is one of the most significant determinants of public transport attractiveness to potential modal switchers (Paulley et al.). If

speed management improves bus punctuality at no additional operational cost, it lowers the effective generalised cost of transit travel, compounding the fare reduction the Pass delivers. A commuter weighing whether to abandon their car will respond to both price and reliability. The Pass addresses the former, and speed management addresses the latter.

The interaction between density reduction and traffic physics is further demonstrated by Tassinari's analysis of Madrid Central. Following LEZ implementation, flow fell by 8.1% and density by 17.7% within the restricted area, producing a post-implementation average speed of 30.16 km/h, a 10.98% increase relative to the pre-policy baseline. This result illustrates that even a modest reduction in vehicle accumulation, of the kind the Pass might plausibly achieve by diverting commuters onto rail and bus, is sufficient to return a near-capacity network to its uncongested branch and generate disproportionate travel-time improvements. The Bureau of Public Roads function's steep exponent near capacity means that the commuter time savings attributable to the Pass would be non-linearly larger for highly congested corridors than a linear extrapolation would suggest. Mueller et al.'s health impact assessment of Barcelona's full superblock expansion estimated that full implementation could prevent 667 premature deaths annually, generating an annual economic saving of EUR 1.7 billion and average life expectancy gains of almost 200 days for the adult population.

2.3.3. Limitations, Displacement and Consistent Findings

Speed-limit reductions applied selectively share the displacement challenge identified in both the LEZ and fare reduction evidence. Rodriguez-Rey et al. ("To what extent") found that superblocks alone, without additionally implementing low-emission zones or demand management, will not achieve the necessary reduction in car traffic and emissions. In Grenoble, following the 30 km/h implementation, there were 9% fewer light vehicles and 20% fewer heavy vehicles overall, but travel times increased by 3% to 5% (Yannis and Michelaraki). The evidence across all three existing Spanish policy instruments reviewed in this section, the LEZ, the 2022-2023 fare reductions, and speed-limit management, points to a consistent structural finding. Each policy in isolation moves the relevant indicator in the desired direction but produces weaker-than-anticipated effects on total congestion-related economic losses because no single instrument addresses the full generalised cost of private vehicle use. The LEZ restricts access but not cost, the fare reduces lower transit cost but not the cost of driving, and speed management improves network efficiency but does not reduce vehicle ownership or the incentive to drive. The National Public Transport Pass, evaluated against this background, is best understood as the demand-side complement to a supply-side and regulatory architecture that Spain has been assembling incrementally over the past decade.

2.4. Congestion Pricing

Whereas Madrid Central restricts access by vehicle type, congestion pricing addresses the externality directly by charging drivers for the social cost they impose. The economic rationale is Pigouvian, and Russo et al. estimate that the marginal external cost of road congestion is roughly

two-thirds of the private time cost of travel, implying that drivers systematically under-price their own contribution to delay and that a corrective charge generating that wedge would yield substantial welfare gains. Existing evidence suggests that urban roads operating near capacity have especially large uncorrected traffic externalities, and that the introduction of congestion pricing can generate positive welfare gains even before revenues are recycled (Cook et al.). For the National Public Transport Pass, this framing is directly relevant. Fare reductions lower the generalised cost of transit but leave the cost of driving unchanged, whereas a congestion charge simultaneously raises the cost of driving, making the two instruments natural counterparts in a two-sided demand management strategy.

2.4.1. International Evidence

London's central charge, introduced in February 2003, cut traffic entering the zone by approximately 18% and congestion by roughly 30% in its first year, with traffic still around 10% below baseline a decade later (Santos). Peak period congestion delays declined by about 30%, bus congestion delays declined by 50%, bus ridership increased by 14%, and average traffic speeds within the zone rose by 37% (Litman, "Implications"). Approximately 80% of the net revenue generated between 2004 and 2006 was reinvested in improving buses in central London, forming part of the comprehensive policy package that produced the observed ridership and reliability gains (Reyes-Sanchez et al.). A modest cordon charge in Madrid or Barcelona, with revenues earmarked for the transit network that the National Public Transport Pass makes financially accessible, would simultaneously suppress car demand and fund the service quality improvements that translate fare savings into durable modal shifts.

Stockholm's cordon scheme provides unusually clean causal evidence through its reversal design. Traffic levels dropped immediately during the trial period, returned towards initial levels once pricing was abolished, then dropped again and remained roughly constant once congestion pricing was reintroduced, with the permanent traffic reduction holding at approximately 20% over the subsequent years (Eliasson). The Stockholm experience showed that public and political acceptability, initially low, rose substantially as objective improvements in travel times and air quality became apparent to the population (Borjesson et al., "Stockholm congestion charges"). This pattern carries a specific implication for Spain, where any congestion charge proposal would face initial political resistance. Spain's rollout of the National Public Transport Pass may, paradoxically, improve the political conditions for a future congestion charge by ensuring that an affordable, well-marketed transit alternative is in place before the charge is introduced (Hysing and Isaksson).

2.4.2. Revenue-Recycling and the National Public Transport Pass

The distributional consequences of congestion pricing depend critically on the revenue-recycling design. Analyses of congestion pricing schemes generally find that they are vertically regressive in isolation, disproportionately affecting low-income groups, though congestion pricing can be progressive if higher-income households are predominantly located in

suburban areas and commute by car to a central business district (Lindsey). Without revenue recycling, congestion pricing is regressive, with appropriate revenue recycling, low-income households can also be better off under congestion pricing than in the no-policy scenario (Nguyen and Muesgens). A revenue-recycling arrangement that directs congestion charge proceeds toward subsidising or expanding the National Public Transport Pass would simultaneously address two weaknesses in the current Spanish policy architecture, addressing the regressivity risk of the charge, and the fiscal sustainability problem of the Pass. Spain has not adopted a true congestion charge. Madrid Central and the proliferating LEZs operate through access bans and emissions criteria rather than prices, which is precisely why their traffic benefits proved susceptible to erosion through fleet renewal. A congestion charge closes this gap by ensuring that every car trip into a congested zone carries a price reflective of the external costs imposed on other road users. A predictive air pollution model applied to a hypothetical urban road pricing scheme in Madrid by Pedro Munoz Miguel et al. found highly significant effects in decreasing air pollution during peak operating hours. Bel and Rosell have identified Barcelona as a clear candidate for congestion charging, noting that the spatial distribution of income and car ownership implies that a central cordon charge would be broadly progressive rather than regressive.

2.5. Measuring Congestion-Related Economic Losses

Quantifying the economic burden of congestion requires translating heterogeneous physical effects, lost time, foregone transactions, wasted fuel, and damaged health, into comparable monetary units, as summarised in Table 2. The largest component is commuter time loss, valued by multiplying hours lost in excess of free-flow conditions by a Value of Travel Time (VTT). TomTom's 2018 index reports that Madrid drivers lose roughly 17 minutes for every 30-minute peak trip, placing Madrid at the upper end of the global range (INRIX). A refinement increasingly emphasised in the literature is the reliability premium. Travellers value the predictability of travel time over and above its average level, and empirical estimates of the Value of Reliability have varied from as low as 0.55 times to as high as 3.22 times the VTT itself (Carrion-Madera and Levinson). Because frequent, high-capacity public transport compresses the variance of journey times, much of the welfare benefit of a measure such as the National Public Transport Pass may accrue through improved reliability rather than through reduced mean travel time alone. Zhu et al. show that travel time variance is highly volatile when demand approaches or exceeds capacity. The Pass, by shifting commuters onto rail and bus corridors where schedule adherence is measurable and enforceable, converts unpredictable car journey times into the more stable distribution characteristic of scheduled public transport.

Table 2. Components of congestion-related economic loss and their valuation methods.

Loss component	Definition	Valuation method	Illustrative estimate / source

Commuter time loss (CTL)	Hours lost relative to free-flow travel	Value of Travel Time (VTT) x lost hours	~17 min per 30-min peak trip, >98 h/yr per driver (Madrid), ~EUR 6.2 bn losses (INRIX)
Reliability premium	Value of reduced travel-time variability	~0.5-1.5 x VTT applied to the variance reduction	Principal channel through which frequent public transport delivers welfare gains (Carrion-Madera and Levinson)
Missed business opportunity (MBO)	Productivity and logistics losses	Freight delay, inventory and supply-chain cost, foregone retail	Retail -20.6%, ~EUR 24 m/yr foregone (Galdon-Sanchez et al.)
Fuel and operating cost	Excess fuel use and vehicle wear	Fuel volume x price, stop-and-go penalty	~350 Ml fuel saved (2023 Spanish fare cuts, est.)
Environmental cost	Air-pollution and CO2 damages	Emission inventory x unit damage cost, ~0.4% of GDP benchmark	~30,650 m pesetas/yr (Madrid, 1992), NOx dominant (Monzon and Guerrero)
Health cost	Mortality and morbidity from traffic-related air pollution	Value of a Statistical Life, avoided medical cost	88 lives + ~EUR 3.4 m/yr saved (Madrid Central), ~900 deaths/yr (Madrid), >1,000 (Barcelona)
Marginal external cost (benchmark)	Under-priced social cost of an additional trip	Share of the private time cost of travel	~two-thirds of private time cost (Russo et al.)

Estimates are illustrative and drawn from the sources reviewed, they are intended to show how each loss channel is monetised rather than to provide a single comprehensive cost figure for Madrid or Barcelona.

Missed business opportunity captures the productivity losses transmitted through freight and logistics, including longer delivery times, higher inventory carrying costs, and supply-chain delays, as well as the foregone retail activity documented for Madrid Central. Environmental losses are monetised by applying unit damage costs to emission inventories or by benchmarking against GDP. An early valuation for Madrid attributed total annual damages on the order of 30,650 million pesetas at 1992 prices to carbon monoxide, nitrogen oxides, and sulphur dioxide, with nitrogen oxides dominant (Monzon and Guerrero). Air-quality modelling for Madrid finds that a 10% reduction in car numbers lowers ambient NO2 concentrations by between 1.6% and 8.7% depending on meteorological conditions, with road traffic responsible for approximately 78% of local NO2 emissions (Garcia et al.). If the Pass achieves a 5% reduction in car vehicle kilometres on central corridors, the implied NO2 benefit lies in the range of 0.8% to 4.4% reduction in concentrations, a potentially significant avoided mortality dividend. Health losses are valued through the Value of a Statistical Life and avoided medical costs (Huib van Essen et al.). A study of health-related social costs of air pollution in 432 European cities found that in 2018 on average every inhabitant of a European city suffered a welfare loss of over EUR 1,250 per year attributable to air pollution (CE Delft). The interaction between these four loss channels and the Pass is not additive but partially overlapping. The environmental and health channels are the most non-linear, as density reductions in the most congested corridors generate health dividends that are disproportionately large relative to the number of trips shifted.

2.6. Spatial Distribution of Growth and Transit-Oriented Economic Activity

The third channel of congestion-related economic loss examined in this review concerns not the aggregate magnitude of costs but their spatial incidence, specifically which districts and which populations capture the economic benefits of improved mobility, and which continue to bear the costs of congestion and poor access. The foundational insight from urban economics is that accessibility drives agglomeration. Improved transit access lowers the effective distance between locations, raising the productivity of firms and the wages of workers in well-connected districts (Ahlfeldt et al.). The National Public Transport Pass, by reducing the cost of reaching central Madrid and central Barcelona by rail and bus, is expected to increase the effective accessibility of transit-served districts and, through the mechanism of agglomeration, raise the productivity and commercial vitality of those areas.

The spatial pattern of these gains, however, is likely to be uneven in ways that reflect the pre-existing distribution of transit infrastructure. Research on travel pass usage in Madrid by Cadena et al. found that pass uptake depended primarily on income level and accessibility to public transport, with accessibility having a positive effect and income a negative influence. This implies that the Pass's benefits, both the direct fare savings and the indirect agglomeration gains from improved access, will be concentrated in districts that already have good transit coverage. Peripheral, lower-density areas of the Madrid and Barcelona metropolitan regions, where car dependence is highest and the generalised cost of driving relative to transit is lowest, are least likely to see modal shift and therefore least likely to capture the spatial economic gains the Pass delivers. Addressing this tension would require extending the Pass's coverage to metro and urban bus networks, which are currently excluded from the Abono Unico but which serve the short-distance, intra-district trips that are most relevant to economic activity in peripheral areas. The government has stated an intention to encourage regional and municipal authorities to integrate local networks into the national pass framework (Autocares), and the degree to which this integration is achieved will be a critical determinant of the Pass's spatial equity outcome.

2.7. Technological and AI-Based Traffic Management in Barcelona

Barcelona illustrates a supply- and information-side approach to congestion built on open data and machine learning. The city's mobility profile makes the problem tractable. Of roughly six million daily trips recorded in 2017, 40.1% were made by public transport and 35.3% by active modes such as walking and cycling, leaving only 24.6% to private vehicles (Garcia et al.). Garcia et al. compare classical time-series forecasting with modern gradient-boosting methods, finding that ARIMA models fail to satisfy the statistical assumptions of independence, homoscedasticity, and normality and cannot capture the spatial and temporal dependencies inherent in urban traffic, whereas an XGBoost model performs substantially better, with Shapley Additive Explanations applied to render the model's predictions interpretable for planners. City projections associated with such forecasting suggest the potential to reduce congestion by around 20%. Predictive systems are most powerful as a coordinating layer over the other instruments examined here. The same models that anticipate congestion can identify when and where a

flat-fare pass is diverting demand, where residual road capacity should be priced, and where speed or access measures would be most effective. For the National Public Transport Pass, AI-based management offers a means of allocating the additional ridership efficiently across the network and of monitoring, in close to real time, whether the Pass is translating into measurable congestion relief or merely into additional travel.

Barcelona's AI system predicts congestion up to 30 minutes in advance and adjusts traffic-light cycles dynamically, while separate AI tools support the optimisation of bus headways, the prediction of overcrowding, and route adjustments during disruptions (Mayors of Europe). In 2024, the Barcelona Innova Mobility Lab challenge selected two projects. The first provided bus stops with real-time data on the city's bus network and surrounding roadways, and the second used artificial intelligence, computer vision, cameras, and sensors with 5G connectivity to detect the occupation of bus lanes from the bus itself (Mobility and transport).

The second project, developed by Hayden AI, has been shown in comparable deployments to increase bus speeds by 5% while lowering collisions by up to 30% (Hayden AI). For the National Public Transport Pass, these improvements in operational speed and lane integrity act as a supply-side multiplier. The Pass reduces the monetary cost of taking the bus, and AI-enabled bus lane enforcement reduces the time cost, compressing the generalised cost of transit travel from both directions simultaneously. Significant challenges remain in the deployment of AI traffic management systems, however, including data quality, model interpretability, and the risk of generating digital divides if vulnerable or marginalised populations lack access to the technologies or underlying services (Zemmouchi-Ghomari).

Ensuring that Barcelona's AI infrastructure is extended equitably across the metropolitan area is therefore a prerequisite for the AI layer to amplify rather than undermine the progressive distributional objectives of the Pass. The critical limitation remains the gap between the city's projected 20% congestion reduction and any peer-reviewed, post-implementation empirical verification of that figure. What the AI evidence does establish is that Barcelona possesses the data infrastructure and technical capacity to conduct a rigorous evaluation of the National Public Transport Pass that earlier Spanish interventions could not.

3. Comparative Evidence

Placing Spain's policy in an international context clarifies both its likely effects and its design risks, since several European jurisdictions have implemented comparable flat-fare or fare-free schemes, summarised in Table 3. Germany's 9-Euro-Ticket is the closest natural experiment. A roughly 90% fare reduction that drove a large, rapid increase in ridership but only modest car substitution, and whose success in revealing latent demand led directly to the permanent Deutschlandticket, introduced in May 2023 at EUR 49 per month and since raised toward EUR 58-63 (Loder et al.). Austria's nationwide KlimaTicket and Tallinn's fare-free public transport for residents tell a consistent story. Ridership rises, but the modal shift away from private cars remains limited unless the fare measure is paired with disincentives for driving (Wallimann). Portugal's EUR 20 monthly national pass and comparable schemes elsewhere

reinforce the pattern, while the wide international price range situates Spain's EUR 60 adult and EUR 30 youth fare toward the more affordable end of the spectrum. Two lessons stand out for Madrid and Barcelona. First, fare integration reliably simplifies the user experience, expands access, and is strongly progressive, but on its own it is an unreliable instrument for reducing car traffic. Second, the durability of the effect depends on permanence and complementary service quality. Temporary schemes such as the 9-Euro-Ticket and Spain's own 2022-2023 discounts produced ridership that partly evaporated on expiry, whereas the permanence of the Deutschlandticket, and now of Spain's national pass, is designed to convert episodic experimentation into a stable change in travel habits (Loder et al.). The Spanish pass therefore enters a well-mapped policy space. Its demand effects are highly predictable from international experience, while its congestion and welfare effects will hinge on the complementary measures with which it is combined.

Table 3. Comparative international flat-fare and fare-free public-transport schemes.

Scheme (jurisdiction)	Type / price	Key ridership effect	Car substitution	Lesson
9-Euro-Ticket (Germany, 2022)	Flat EUR 9/month for 3 months (~90% cut)	Regular PT use 29% to 38%	Only ~10-20% replaced a car trip, PT distance -37% post-expiry	Strong demand response, weak and transient car shift
Deutschlandticket (Germany, 2023-)	Permanent EUR 49 (rising to ~EUR 58-63)	Sustained higher PT use	Modest, made durable by permanence	Permanence converts a trial into a habit
KlimaTicket (Austria)	EUR 1,095/year (now EUR 1,400), nationwide	Ridership up	Limited without driving disincentives	Fare integration alone is not mode shift
Fare-free PT (Tallinn)	Free for residents	Ridership up	Limited modal shift, +1.2% passenger demand	Needs complementary car restraint
National pass (Portugal)	EUR 20/month, residents	Ridership up	Recent, limited evidence	Affordable-end benchmark
Spanish fare reductions (2022-23)	Free Cercanias + 30-50% discounts	Subsidised-mode share 11.9% to 15.4%	More travel rather than less car use	High elasticity, weak substitution
Abono Unico (Spain, 2026)	EUR 60 adult / EUR 30 under-26	Projected large uptake	To be evaluated post-implementation	Effect on congestion hinges on complements

Prices and figures reflect the most recent values reported at the time of writing, the Deutschlandticket price has been adjusted upward since its 2023 launch.

3.1. Evidence from the Deutschlandticket

The Deutschlandticket provides the most instructive international parallel. Analysis from the Mercator Research Institute found that the Deutschlandticket produced 30% more rail journeys, but that the announced price increase from 49 to 58 euros per month could almost halve the emission reductions achieved, with train journeys projected to decrease by 14% and car kilometres to increase by 3.5% in response (Wettengel). At EUR 60 per month for adults, the

Abono Unico is priced slightly above the Deutschlandticket's launch price but below its current rate, and the German evidence on price elasticity suggests that the choice of fare level is not merely a fiscal decision but a critical determinant of the Pass's modal shift potential. Loder et al. found that the Deutschlandticket resulted in a 10% increase in ticket subscriptions for local public transport services, but that only one fifth of new subscribers reduced their use of automobiles at the same time, confirming the same structural finding documented for Spain's 2022-2023 programme. By early 2026, the Deutschlandticket had attracted approximately 14.6 million subscribers, contributing to a measurable rise in public transport ridership estimated at around 21% of trips linked to its availability.

3.2. Evidence from Austria's KlimaTicket and Tallinn

Austria's KlimaTicket, introduced in October 2021 as a EUR 1,095 per year nationwide unlimited pass, provides a contrasting case. From October 2021 to June 2022, some 170,000 passengers purchased the KlimaTicket, with approximately two-thirds of customers indicating they started using public transport more frequently and 85% reporting they had already replaced some car journeys with public transport (OECD). However, causal evaluation using synthetic control methods by Wallimann found that the KlimaTicket had no significant short-run passenger growth effects on the Austrian Federal Railways. This divergence between self-reported behavioural change and measured passenger demand underscores a methodological caution central to evaluating Spain's pass. Survey-based evidence on modal shift is likely to overstate the substitution effect relative to what revealed preference and sensor-based measurement will ultimately confirm. Tallinn's fare-free public transport scheme for residents, introduced in 2013, provides a further reference point. Cats et al. found that the fare-free measure accounted for an increase of only 1.2% in passenger demand, with the remaining demand growth attributable to other concurrent changes in supply and urban development. The finding that even full fare elimination produces only marginal modal shift in the absence of disincentives for driving reinforces the consistent pattern running across every row of Table 3.

3.3. Distributional Implications Across the Comparative Evidence

Across all three international comparators, a consistent distributional pattern emerges that is directly relevant to Spain's policy design. Research using propensity score matching to evaluate the 9-Euro-Ticket and Deutschlandticket found that both tickets drastically reduced the complexity of the fare system and lowered the cost of public transport, with qualitative research suggesting that the 9-Euro-Ticket provided financial relief and increased social inclusion for low-income households (Waldorf et al.). For Spain, Cadena et al. found that travel pass usage in Madrid depended primarily on income level and accessibility to public transport, with accessibility having a positive effect and income a negative influence on uptake, suggesting that lower-income households in transit-accessible districts are disproportionate beneficiaries of flat-fare passes. The EUR 30 youth tier reinforces this progressive architecture. Young people are disproportionately dependent on public transport, more responsive to price signals, and more

likely to form durable modal habits. Table 3's positioning of the Abono Unico as an affordable-end benchmark among its international peers is accurate in fiscal terms, but the comparative evidence makes clear that affordability at the lower end of the price spectrum does not by itself resolve the substitution problem that every scheme in the table has encountered to varying degrees.

4. Implications for Spain's National Public Transport Pass

Synthesising the evidence reviewed above, Spain's National Public Transport Pass is best understood not as a standalone congestion remedy but as the demand-side anchor of a broader mobility strategy. The strongest claim the literature supports is that the Pass will substantially increase public transport ridership and deliver large, progressive cost savings, potentially reducing a typical commuter's monthly transport expenditure by up to 60% according to official government estimates (Reyes-Sanchez et al.), with the greatest gains accruing to lower-income households in transit-rich districts. This is consistent both with the demonstrated price elasticity of Spanish demand during the 2022-2023 fare reductions and with the international experience of the 9-Euro-Ticket and Deutschlandticket.

Its effect on congestion-related economic losses, however, is contingent. The 9-Euro-Ticket and the 2022-2023 Spanish discounts show that fare reductions generate weak car substitution and partly transient ridership unless made permanent and paired with driving disincentives (Loder et al.). Madrid Central demonstrates that access-based traffic gains erode through fleet renewal and induced substitution (Galdon-Sanchez et al.. Tassinari), the congestion-pricing record from London, Stockholm, and Milan shows that directly pricing the externality delivers durable traffic and emissions reductions, with public acceptance growing as benefits become visible (Eliasson), the traffic-flow physics of speed management indicates that removing vehicles from dense networks moves links back toward the productive region of the fundamental diagram, amplifying time and emissions savings (Yannis and Michelaraki), and Barcelona's AI forecasting offers the coordinating intelligence to allocate ridership and price residual capacity efficiently (Garcia et al.).

Where the Pass is combined with these complements, the LEZs now mandatory in every Spanish municipality above 50,000 residents, urban speed management, selective congestion pricing, and predictive traffic management, the analysis supports this paper's thesis that it can meaningfully reduce commuter time loss, fuel consumption, and CO2 emissions, principally through improved reliability and accessibility, while redistributing retail and productivity gains toward well-connected, transit-oriented areas. Where it is deployed in isolation, the more likely outcome is a welcome equity and ridership gain accompanied by only a modest congestion dividend, as additional discretionary travel offsets part of the mode shift.

5. Conclusion

The central argument of this review is that Spain's National Public Transport Pass represents a necessary but insufficient condition for reducing congestion-related economic losses

in Madrid and Barcelona. The evidence assembled across the four channels of commuter time loss, missed business opportunity, the spatial distribution of growth, and environmental and health externalities converges on a consistent structural finding. No single instrument, whether a LEZ, a fare pass, speed management, or congestion pricing, addresses the full generalised cost of private vehicle use, and it is the interaction of multiple instruments that produces welfare gains of the magnitude that Spain's congestion problem warrants.

The Pass's most secure contribution is to the reliability and equity dimensions of the problem. By consolidating the Cercanias, regional rail, and state bus network under a single, affordable monthly instrument, it lowers the transaction cost of transit use, provides a stable price signal that can support longer-run behavioural adaptation, and delivers its largest absolute benefits to the lower-income, transit-dependent households who suffer most acutely from the health and time costs of congestion. As Cadena et al. demonstrate, accessibility level has a positive effect on pass uptake while income level has a negative influence, indicating that lower-income households in transit-accessible districts are disproportionate beneficiaries of subsidised pass schemes. This progressive incidence is reinforced by the EUR 30 youth fare, which targets the demographic group most sensitive to transport costs and most likely to form durable transit habits that persist into adulthood.

The Pass's contribution to the congestion-relief channels is more uncertain and more contingent on the complementary policy architecture. Translating the Pass's ridership gains into density reductions requires that those gains represent genuine car substitution rather than induced demand, a condition that the 2022-2023 evidence and the international comparators consistently show will be only partially satisfied in the absence of pricing and access disincentives for driving. Congestion pricing remains the critical missing instrument in Spain's current policy mix. Without a price signal that raises the marginal cost of an additional car trip to reflect the social externality it imposes, the Pass will continue to perform more strongly as a welfare and equity instrument than as a congestion management tool.

The decisive empirical questions raised by this review, the realised rate of car substitution, the persistence of ridership beyond the novelty period, and the spatial incidence of benefits across central and peripheral districts, can only be resolved through rigorous post-implementation evaluation. Spain possesses, in Barcelona's Open Data BCN sensor network and AI forecasting infrastructure, the technical capacity to conduct such evaluation in near real time (Garcia et al.). Establishing a formal evaluation framework around the Pass's January 2026 introduction, and applying the valuation methods summarised in Table 2 to the observed changes in vehicle kilometres, modal shares, NO2 concentrations, and travel-time distributions in the months following implementation, constitutes the most urgent priority for post-implementation research identified in this review. Given the scale of Spain's congestion-related economic losses, quantified at over EUR 6.2 billion annually in Madrid alone, the welfare stakes of getting the complementary policy architecture right are considerable, and the window for learning from the Pass's early implementation is narrow.

Works Cited

Ahlfeldt, G.M., et al. "The Economics of Density: Evidence From the Berlin Wall." Econometrica 83.6 (2015): 2127-2189.

Autocares, A. y. "El gran reto del Abono Único es extenderlo a las comunidades autónomas." Autobuses Autocares, 2026, www.autobusesyautocares.com.

Bel, G., and J. Rosell. "The impact of socioeconomic characteristics on CO2 emissions associated with urban mobility." Energy Economics 64 (2017): 251-261.

Borjesson, M., et al. "The Stockholm congestion charges, 5 years on." Transport Policy 20 (2012). ["Stockholm congestion charges"]

Borjesson, M., J. Eliasson, and I. Rubensson. "Distributional effects of public transport subsidies." Journal of Transport Geography 84 (2020): 102674. ["Distributional effects"]

Cadena, P.C.B., et al. "Social and Distributional Effects of Public Transport Fares and Subsidy Policies: Case of Madrid, Spain." Transportation Research Record 2544 (2016): 47-54.

Carrion-Madera, C., and D. Levinson. "Value of Travel Time Reliability: A Review of Current Evidence." University of Minnesota Nexus Research Group Working Paper, 2010.

Cats, Oded, et al. "Public Transport Pricing Policy: Empirical Evidence from a Fare-Free Scheme in Tallinn, Estonia." Transportation Research Record 2415 (2014): 89-96.

CE Delft. Health costs of air pollution in European cities and the linkage with transport. CE Delft, 2020.

Cook, C., et al. "NBER Working Paper: Assessing Congestion Pricing Welfare Effects." National Bureau of Economic Research, n.d.

Desk, N. "Spanish government extends free train routes to continue throughout 2023." Spain English, 8 Oct. 2022.

Dragonetti, W. "Bilbao: Slow and steady for the win." Eurocities, 2020.

Eliasson, J. "Lessons from the Stockholm congestion charging trial." Transport Policy 15.6 (2008): 395-404.

European Bank. "Urban planning with superblocks: Barcelona, Spain." EBRD Green Cities, n.d.

Galdon-Sanchez, J.E., et al. "Social Benefits and Private Costs of Driving Restriction Policies: The Impact of Madrid Central." Journal of the European Economic Association 21.4 (2023): 1227-1267.

Garcia, E., et al. "Predictive Analyses of Traffic Level in the City of Barcelona: From ARIMA to eXtreme Gradient Boosting." Applied Sciences 14.11 (2024): 4432.

Gonzalo-Orden, H., et al. "Why is it necessary to reduce the speed in urban areas to 30 km/h?" Transportation Research Procedia 58 (2021): 209-216.

Hayden AI. "Barcelona launches automated bus lane and bus stop enforcement pilot with Hayden AI." Hayden AI Blog, 2025.

Herzog, I. "The city-wide effects of tolling downtown drivers: Evidence from London's congestion charge." Journal of Urban Economics 144 (2024): 103714.

Hughes, R. "Which European countries offer monthly public transport passes?" Euronews, 16 Dec. 2025.

Huib van Essen, CE Delft, et al. Handbook on the external costs of transport: version 2019. Publications Office, 2019.
Hysing, E., and K. Isaksson. "Building acceptance for congestion charges, the Swedish experiences compared." Journal of Transport Geography 49 (2015): 52-60.
INRIX. INRIX 2025 Global Traffic Scorecard. INRIX, 2025.
International Transport Forum. The Economics of Fare-Free Public Transport. OECD/ITF, 2022.
Lindsey, R. Distributional effects of urban transport policies to discourage car use: A literature review. OECD Environment Working Papers, No. 211, 2023.
Litman, T. "Implications for Other Cities." Victoria Transport Policy Institute, 2011. ["Implications"]
Litman, T. "Transit Price Elasticities and Cross-Elasticities." Journal of Public Transportation 7.2 (2004): 37-58. ["Transit Price Elasticities"]
Loder, A., et al. "Observing Germany's nationwide public transport fare policy experiment '9-Euro-Ticket.'" Case Studies on Transport Policy 15 (2024): 101148.
Martinez, E. "Spanish government to launch new unified transport travel pass in January for 60 euros per month." Sur in English, 2025.
Mayors of Europe. "Barcelona's AI-Driven Mobility Transformation." Mayors of Europe, 2025.
Ministry for the Ecological Transition and the Demographic Challenge. "Las zonas de bajas emisiones en Espana." MITECO, n.d.
Ministry of Transport and Sustainable Mobility. "El uso del transporte publico sube un 33% en toda España." MITMA, 2025.
Mobility and transport. "Smart stops and AI to improve circulation, the winning projects in the urban innovation challenge." Ajuntament de Barcelona, 2024.
Monzon, A., and M.-J. Guerrero. "Valuation of social and health effects of transport-related air pollution in Madrid (Spain)." Science of the Total Environment 334-335 (2004): 427-434.
Mueller, N., et al. "Changing the urban design of cities for health: The superblock model." Environment International 134 (2020): 105132.
National Road Safety Observatory. Main Figures on Road Traffic Accidents Spain 2023. Spanish Interior Ministry, 2024.
Nguyen, T.N., and F. Muesgens. "Fuel tax loss in a world of electric mobility: A window of opportunity for congestion pricing." arXiv preprint arXiv:2409.20033, 2024.
OECD. "Austria's 'KlimaTicket' to promote low-carbon mobility." OECD IPAC, 2022.
Paulley, N., et al. "The demand for public transport: The effects of fares, quality of service, income and car ownership." Transport Policy 13.4 (2006): 295-306.
Pedro Munoz Miguel, J., et al. "A forecast air pollution model applied to a hypothetical urban road pricing scheme: An empirical study in Madrid." Transportation Research Part D 55 (2017): 21-38.
Postaria, R. "Superblock (Superilla) Barcelona, a city redefined." Cities Forum, 2021.
Reyes-Sanchez, A., S. Sharifi-Asl, and L. Mozaffarian. "Case Study II: Congestion Charging in London: The Western Expansion." 2024.

Rodriguez-Rey, D., et al. "A coupled macroscopic traffic and pollutant emission modelling system for Barcelona." Transportation Research Part D 92 (2021): 102725. ["A coupled macroscopic"]

Rodriguez-Rey, D., et al. "To what extent the traffic restriction policies applied in Barcelona city can improve its air quality?" Science of the Total Environment 807 (2022): 150743. ["To what extent"]

Rubensson, I. "Distributional effects of public transport subsidies." Journal of Transport Geography 84 (2020): 102674.

Russo, A., et al. "Welfare losses of road congestion: Evidence from Rome." Regional Science and Urban Economics 89 (2021).

Santos, G. "London Congestion Charging." Brookings-Wharton Papers on Urban Affairs (2008): 177-234.

Smith, K. "Spanish public transport soars." International Railway Journal, 2025.

Suijs, L.C.W., et al. "Phantom Jam Avoidance Through In-car Speed Advice." Transportation Research Procedia 8 (2015): 227-236.

Tassinari, F. "Low emission zones and traffic congestion: Evidence from Madrid Central." Transportation Research Part A 185 (2024): 104099.

Waldorf, I., et al. "Low fare public transport as social policy? A study of Germany's 9-Euro-Ticket and Deutschlandticket via propensity score matching." Transportation, 2025.

Wallimann, H. "Austria's KlimaTicket: Assessing the short-term impact of a cheap nationwide travel pass on demand." Transport Policy 159 (2024): 201-214.

Wettengel, J. "Flat-rate train ticket reduced Germany's transport emissions by 5% in first year." Clean Energy Wire, 2024.

Yannis, G., and E. Michelaraki. "Review of City-Wide 30 km/h Speed Limit Benefits in Europe." Sustainability 16.11 (2024): 4382.

Zemmouchi-Ghomari, L. "Artificial intelligence in intelligent transportation systems." Journal of Intelligent Manufacturing and Special Equipment 6 (2025): 26-42.

Zhu, S., G. Jiang, and H. Lo. "Capturing Value of Reliability through Road Pricing in Congested Traffic under Uncertainty." Transportation Research Procedia 23 (2017): 664-678.

China's Pet Economy: Development, Digital Growth, and Current Challenges

By Xinhe Wang

Abstract

China’s pet economy has grown rapidly as changing social attitudes, rising incomes, urbanization, and digital technology have increased demand for pet-related products and services. This article examines the development of China's pet industry, focusing on pet food, health care, services, products, e-commerce and social media. Analysis shows that consumer demand has expanded beyond basic feeding, while digital platforms have accelerated product discovery and online purchase. However, rapid growth has also brought challenges, especially the regulatory gap and the imbalance in the quality of products and services. The paper believes that clearer regulations, stricter quality standards, better product information and more responsible digital platforms are crucial to building consumer trust and supporting the sustainable development of China's pet economy.

Introduction

China's pet economy has grown from a small market into an important part of consumer spending. According to the China Pet White Paper, people living in Chinese cities are expected to spend about 312.6 billion yuan on pets in 2025, which is a 4.1% increase from 2024. The report also predicts that China's pet market will reach 405 billion yuan by 2028. This growth shows that pet spending has become a visible part of China's consumer economy.

However, compared with mature markets, China's pet ownership rate remains relatively low. Although China ranked above Japan and South Korea, its rate remained below that of many countries in Europe and the Americas (GfK). The difference suggests that a further increase in pet ownership could support future demand in China.

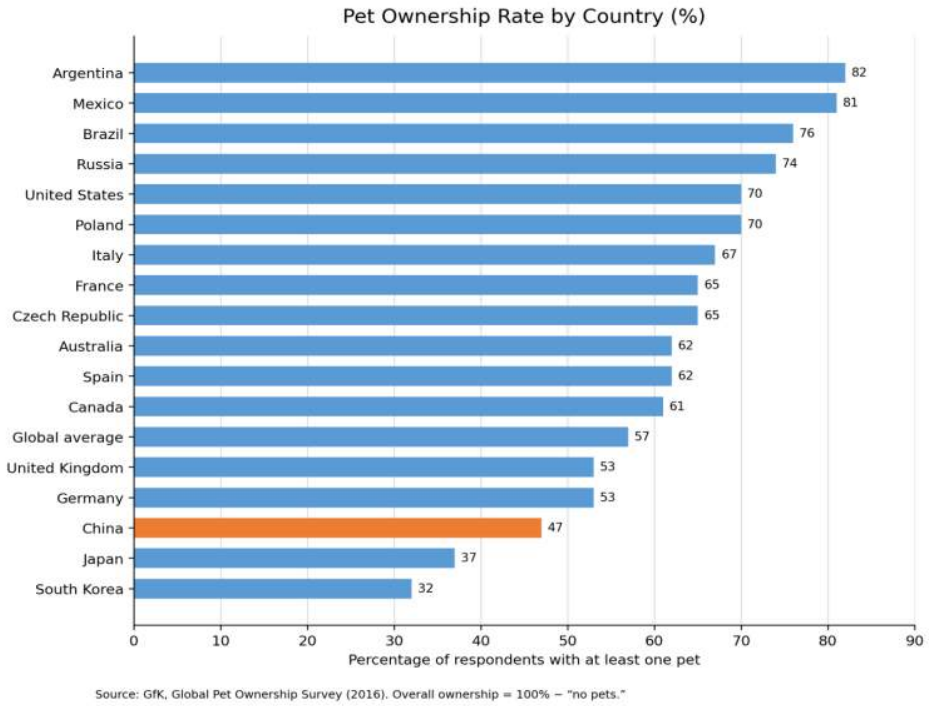


Figure 1. Pet ownership rates in selected countries. Source: GfK, Global Pet Ownership Survey, 2016. Overall ownership equals 100% minus the share of respondents who reported "no pets."

Against this background, this paper examines the development of China's pet industry and the current state of the market. The following discussion covers pet food, healthcare, services, products, online retail, and social media. It also considers two problems that may slow future growth: gaps in regulation and uneven product and service quality. Together, all these developments and challenges show how China's pet economy has evolved and what may shape its future growth.

1. The Development and Current Structure of China's Pet Industry

China's pet industry started later than the pet industries of many Western countries, but the Chinese market has developed and changed rapidly. Over a relatively short period, many Chinese people began to treat pets as companions and family members rather than useful tools. This change has created demand for a wider range of pet-related products and services.

1.1 Historical Development

Pets in ancient China often served practical purposes. During the Han dynasty, people used dogs to hunt, guard homes, control rodents, and protect property. Historical evidence also shows that some people kept dogs as companions. Companion keeping therefore existed, but practical roles remained more common throughout China's long agrarian period (Devienne 365–82).

This pattern of pet keeping began to change as China entered the modern period. During the Maoist period, pet keeping became much less visible in Chinese cities because the government associated it with a bourgeois lifestyle. Pet ownership slowly returned after economic reform began (Doi and Pettier 196–223). The change did not happen all at once, but it created the social conditions for a modern companion-animal market. From the 1990s onward, rising incomes, urban growth, smaller families, and changing attitudes made pet keeping more common. The industry expanded more quickly after 2010. Online shopping and social media later made pet products and pet-related information easier to access. Consumers gradually moved beyond basic feeding and began to spend more on nutrition, healthcare, play, convenience, and companionship (China Financial Information Network).

1.2 Current Structure

Through the historical development described above, China's pet industry has evolved into a diversified market encompassing pet food, healthcare, services, and products.

i. Pet Food

Pet food remains a central part of China's pet economy. This category includes dry staple food, wet food, snacks, freeze-dried food, fresh food, and formulas designed for different ages or health needs. Recent industry reports show that consumers increasingly pay attention to product quality, nutrition, and specialized formulas (Eastmoney; Sina Finance).

This trend can be seen in the growing interest in freeze-dried food, which preserves the texture and flavor of the original ingredients. The share of owners who bought freeze-dried staple food increased from 45.1% in 2022 to 47.2% in 2023 (Petdata). Although the increase was small, it shows that consumers are paying greater attention to the quality, ingredients, and nutritional value of the food they give their pets.

ii. Pet Healthcare

Pet healthcare has become another important part of the market. In 2024, the market reached 84.1 billion yuan and made up 28% of all pet spending. From 2017 to 2024, it grew by an average of 16.7% each year (Sohu). As pet owners spend more on healthcare, they are also seeking a wider range of medical services, including diagnosis, treatment, examinations, medication, and preventive care. Pet diagnosis and treatment made up 37.8 billion yuan, pet medicine 29.1 billion yuan, physical examinations 9.6 billion yuan, and vaccines 7.5 billion yuan. The supply of these services has also expanded. By early 2025, China had 34,339 pet hospitals, which were 16.6% more than the year before (Sina Finance). This growing demand shows that many owners view medical spending as a normal part of responsible pet care rather than an unusual expense.

However, the rapid growth in the number of pet hospitals does not guarantee consistent quality. Hospitals differ in equipment, staff training, prices, and service standards. These

differences make it difficult for owners to judge value and may weaken consumer trust. The industry therefore needs clearer standards for both chain hospitals and independent clinics.

iii. Pet Products

Traditional pet products include everyday items such as litter, cleaning supplies, beds, bowls, toys, clothing, leashes, and carriers. Beyond these basic products, industry reports show that Chinese consumers are buying a wider range of daily supplies and specialized products for their pets (Guangming Daily). Smart feeders, water fountains, cameras, and tracking devices are examples of this shift. Smart devices help owners care for pets when no one is at home, and many devices connect directly to mobile apps. As these products become more advanced, companies are now competing on safety, durability, useful data, and convenience instead of relying only on appearance and price (China Financial Information Network).

iv. Pet Services

As pet owners spend more on convenience and professional care, pet services have become another part of the market. These services include grooming, boarding, training, photography, transport, insurance, and funeral services. For owners with busy schedules, professional services can provide support that is difficult to manage on their own.

Pet training provides a clear example of how this service market works. A 2026 China Consumer News report says that pet training is changing from an add-on service into a separate spending category. Different programs offer different course lengths and service levels, which creates a wide price range. For example, some programs cost about 1,980 yuan for one week and 8,800 yuan for one month (China Economic Net).

The growing demand creates business opportunities, but it also increases the need for professional training and clear service standards (China Economic Net). Training providers do not always have clear qualifications and training methods, making it difficult for consumers to assess the quality of different programs. As the market continues to expand, addressing these issues will become increasingly important for the industry.

As the range of pet products and services expands, so do the ways consumers find and purchase them. While physical stores and animal hospitals remain important, the emergence of online platforms gives consumers a wider product selection, easier price comparisons, and home delivery. This evolving retail landscape provides important context for understanding the distinctive role of e-commerce in China's pet market.

2. Distinctive Traits of China's Pet Market: A Digital Development Path

Digital platforms strongly shape China's pet market. E-commerce influences where owners buy products, while social media influences what owners notice, discuss, and trust. Digital platforms help brands reach consumers across the country. However, the same platforms also create concerns about information quality and platform responsibility.

2.1 E-commerce-driven: The Growth of Online Pet Retail in China

Chinese online retail has become an important shopping channel for Chinese pet owners. Consumers can compare prices, read product information, and buy goods that may not be available in nearby stores. These advantages make online retail particularly suitable for standardized pet products.

The growth of online retail benefits both consumers and smaller brands. For consumers, large platforms lower search costs, provide home delivery, and allow buyers to compare products through reviews and sales records. For brands, online platforms provide access to a wider consumer base without requiring a large physical retail network. During Tmall's Double 11 shopping festival, domestic brands made up 68% of pet businesses with sales above 10 million yuan (Sina Finance), showing the importance of online platforms for smaller and emerging brands. However, low entry costs can allow poor-quality products and misleading claims to spread. E-commerce improves access and competition, but the market also needs stronger platform checks and clearer product information (Founder Securities).

2.2 Social Media and Livestream Shopping

As online shopping becomes more popular, digital platforms are becoming an important source of information for pet owners. KPMG reports that many Chinese pet owners get information from short videos and social media. This gives pet owners opportunities to share experiences, read product reviews, and learn from other consumers. Social media therefore supports product discovery and purchasing, as well as advertising (China Financial Information Network).

One reason social media can have this influence is that it shapes how pet owners think about "good" pet care. Content about "scientific pet care," "refined pet parenting," reviews, and product suggestions shows how other owners feed, entertain, and care for their pets. As a result, consumers can find new products and learn new ways to care for pets. For example, content about healthier ingredients may make owners notice claims such as "no artificial additives." Videos about smart feeders, pet toys, and other products also show new ways to improve pets' daily lives. In this way, emotional attachment can create demand for products that owners did not think about before. A 2026 report from Sina Finance notes that sales growth in smart pet products and smart toys reached 1.56 times and 1.47 times the earlier level (Sina Finance). Livestream shopping has also become an important part of this process. Consumers can learn more about products through pet-related content, reviews, and livestreams, and quickly access a purchase link (KPMG China).

In this way, social media and e-commerce work together. Social media builds emotional ties, introduces new ways to care for pets, and creates demand, while e-commerce makes it easier to turn that demand into purchases. The process is content discovery → emotional connection → new ideas about pet care → product demand → quick purchase.

2.3 Summary

E-commerce and social media have made China's pet market faster, wider, and easier to enter. Digital platforms help consumers find products and allow brands to reach customers across the country. However, digital platforms can also spread poor-quality products and unclear claims at great speed. The market therefore needs stronger checks, clearer information, and more responsible platform management.

3. Current Challenges in China's Pet Industry

China's pet economy still has strong growth potential, but rapid expansion has created several problems. The industry faces gaps in regulation and uneven quality in a fragmented market. These problems can harm animal welfare, weaken consumer trust, and limit the development of reliable brands.

3.1 Gaps in Pet Industry Regulation

China's fast-growing pet market has created problems with regulation, especially in the sale of live animals. One example is "weekend pets." These animals look healthy when they are sold but become sick soon after purchase. In a 2025 case at the Minhang District Court in Shanghai, a pet store sold sick animals for more than six years. It attracted consumers online and then pushed them to buy expensive packages in the store. More than 1,000 complaints were made in Shanghai, more than 170 people filed criminal complaints, and reported losses were over 1.44 million yuan. This was China's first criminal case that directly targeted the sale of "weekend pets" (Dong). The case shows a wider problem: regulation has not grown as fast as China's pet market. Doi and Pettier state that China has no national law that directly protects companion animals.

Many current rules are local and focus more on controlling pets than protecting their welfare. The United States offers one possible example of a more developed regulatory approach for China. Many U.S. states have pet purchase protection laws, often called "puppy lemon laws." These laws let buyers of sick animals or animals with birth defects ask for a refund, a replacement animal, or payment for vet costs. Buyers usually need a certificate from a licensed vet. Recent cases also show that enforcement matters. State authorities have acted against pet sellers for illegal sales, false information, and violations of pet purchase rules. Japan offers another example. It faced similar problems as its pet industry grew, so it gradually strengthened its rules. For example, Japan changed its rules in 2012 to ban pet stores from showing dogs and cats at night. It also reviews and improves its animal welfare system and pays more attention to rules for pet owners and animal businesses. Together, the U.S. and Japanese examples show the importance of clear rules and strong enforcement. China could require sellers to give clear health and source information, define what sellers must do when animals become sick, and regularly review and enforce these rules more strongly.

3.2 Uneven Quality and Weak Brand Development in a Fragmented Market

A market with many small companies is not necessarily unhealthy. Small companies can increase choice and support innovation. However, problems emerge when a fragmented market also has weak standards, limited information, and intense price competition. Chinese securities research describes a competitive pet-food market in which stronger brands are gaining share while smaller brands face greater pressure (Founder Securities; Eastmoney).

These conditions can make it difficult for the market to build consistent quality and consumer trust. First, consumers may struggle to compare product safety, ingredients, and service quality. Second, companies may focus on low prices and short-term promotions instead of research, employee training, and stable quality. Third, weaker brands may depend too heavily on temporary attention from digital platforms. The market can therefore increase its sales without building lasting consumer trust.

The solution is not simply to reduce the number of companies. Regulators and businesses should instead improve product testing, labeling rules, professional qualifications, and complaint records. Online platforms should also remove false claims and display clearer information about sellers. These basic standards would allow both small and large companies to compete through quality, service, and innovation.

Conclusion

China's pet economy has developed rapidly because social attitudes, urban life, household income, and digital technology changed together. Consumer demand has moved beyond basic feeding and now includes nutrition, healthcare, services, products, e-commerce, and social media. Pet food remains the largest category. At the same time, growing spending on healthcare, services, and smart products shows that owners want more complete care.

The future of China's pet economy will depend on consumer trust as well as sales growth. Clearer national regulations can define the duties of breeders, sellers, hospitals, and service providers. Better product information, stronger professional qualifications, and more responsible online platforms can protect both animals and consumers. Addressing these areas will help China's pet market continue to grow while becoming more stable, professional, and innovative.

Works Cited

China Economic Net. “Pet Training Market Expands as Service Standards Remain Unclear.” China Economic Net, 21 Apr. 2026, city.ce.cn/tjyd/202604/t20260421_2917851.shtml.

China Financial Information Network. “Pet Industry Market and Consumption Trends.” 27 Oct. 2025, m.cnfin.com/wx/share?url=//m.cnfin.com/zs-lb//zixun/20251027/4323902_1.html.

Citywide Law Group. “Buying a Puppy: Lemon Laws for Companion Animals.” Citywide Law Group, citywidelegalgroup.com/guides/buying-a-puppy-lemon-laws-for-companion-animals/. Accessed 20 Aug. 2026.

Devienne, Elsa. “Dogs in Early Imperial China: Anthropo-Zoological Reading of Iconographic Sources from the Han Dynasty (206 BC–AD 220).” Dogs, Past and Present: An Interdisciplinary Perspective, Archaeopress Publishing, 2023, pp. 365–82.

Doi, K., and Jean-Baptiste Pettier. “Animal Protection in China and Japan: The Ambiguous Status of Companion Animals in Rapidly Changing Societies.” Protecting the Weak in East Asia, Routledge, 2018, pp. 196–223.

Dong, Y. “Shanghai’s First Criminal Case Involving the Sale of ‘Weekend Pets’: The Group Was Involved in Fraud and Money Laundering and Was Identified as an Organized Criminal Group.” Zhengce Law Firm, 22 Oct. 2025, https://www.joint-win.com/Cn/new/news/id/977/catid/39.html.

Eastmoney. “Pet Food Industry Research Report.” 30 July 2025, data.eastmoney.com/report/zw_industry.jshtml?infocode=AP2025073017181815.

Founder Securities. “Pet Industry Research Report.” 8 Aug. 2025, www.foundersc.com/infoResearchMacro/8659704.jhtml.

GfK. Pet Ownership: Global GfK Survey. 2016, cdn2.hubspot.net/hubfs/2405078/cms-pdfs/fileadmin/user_upload/country_one_pager/ar/documents/global-gfk-survey_pet-ownership_2016.pdf. Accessed 20 Aug. 2026.

Guangming Daily. “‘China (Nanhe) Pet Food Series Index’ Released.” 29 Oct. 2025, economy.gmw.cn/2025-10/29/content_38375064.htm.

KPMG China. 2025 China Pet Industry Market Report: Consumption Upgrade and Market Insights of the “Pet Economy”. June 2025, assets.kpmg.com/content/dam/kpmgsites/cn/pdf/zh/2025/06/2025-china-pet-industry-market-report.pdf.

Petdata. “2023–2024 China Pet Freeze-Dried Consumption Report.” 2024, www.petdata.cn/infodetail_6963337131884544.html.

Sina Finance. “2025 Pet Food Industry Review and 2026 Outlook: Entering the Era of Scientific Pet Care.” vip.stock.finance.sina.com.cn/q/go.php/vReport_Show/kind/lastest/rptid/823974313481/index.phtml. Accessed 20 Aug. 2026.

Sohu. “China Pet Industry White Paper.” www.sohu.com/a/959684442_99921395. Accessed 20 Aug. 2026.

Volatility Regimes and the Dispersion–Return Relationship: Testing for Herding in U.S. Equities During the 2007–2009 Crisis By Hrishik Agarwal

Abstract

This article investigates whether the cross-sectional dispersion of U.S. stock returns behaves differently during periods of high volatility. To reach this conclusion, 439 trading days from the period of November 2007 through July 2009 were analysed, using a sample of 101 S&P 100 index stocks. The analysis of dispersion in stock returns is done by means of the cross-sectional absolute deviation methodology, and the high-volatility periods are classified as those days which remained in the top quartile according to the VIX index over the selected period of study. The pooled regression estimates suggest that the squared-return coefficient amounts to −0.664, thus being negative but not revealing any significance at conventional confidence levels (HAC $t = -1.79$), which does not provide convincing proof of herding effects if examining the data as a whole. Nevertheless, taking into account the volatility changes in two states yields the following results: firstly, on stress days, the level of dispersion increases by 0.0068, as confirmed by every regression model; secondly, the convexity of the relation between dispersion and return substantially decreases, and the squared-return coefficient falls from the ordinary days value of 10.85 to an implied 0.16 for stress days. However, the implied coefficients for stress days remain uncertain and do not demonstrate stable negative figures while different volatility thresholds and return treatments are utilized. All inference uses Newey–West standard errors, and the design is observational throughout.

Introduction

The financial crisis of 2008 thus created a setting under which behavioral aspects of equity market functioning can be examined. There was a significant increase in uncertainty and deterioration in the quality of information, as well as substantial daily fluctuations in the market. The recession in the USA that coincided with the crisis lasted from December 2007 to June 2009 (Labonte), and volatility in the S&P 500 during this period has been significantly different from what has been observed in the past (Akter and Nobi). The need for investors to rely on the actions of others rather than on private information becomes more relevant for analyzing the events of that time, as the signals from private information were very noisy and many events in the public news were difficult to interpret (Banerjee; Bikhchandani et al., “Theory of Fads”). It is important to determine whether or not these mechanisms left observable traces in the market data of those times.

Herding occurs in financial markets when correlation between investors’ actions arises that is not explained by fundamentals alone. Herding can happen either through informational learning or via reputation. Constraint-based effects, for example margin calls, can also make investors behave similarly in a state of crisis without any incentive to imitate (Bikhchandani and Sharma). In periods of crises, all these mechanisms become active at once. In practice, it is not easy to separate the impact of high volatility from that of herding, as each of them can be applied

to the same issue, and any macroeconomic news, liquidity shocks as well as large institutional repositionings will occur jointly. Testing for associations and regime differences, rather than causal contribution, is therefore the appropriate approach here.

> This paper focuses on the U.S. stock market during the crisis window and asks: To what extent are indicators of investor herding associated with higher equity market volatility in the U.S. stock market during the 2008 financial crisis window (2007–2009)?

In this research, dispersion is evaluated through the use of the cross-sectional absolute deviation (CSAD) that follows the regression arrangement of Chang, Cheng, and Khorana (2000), also known as CCK. CSAD measures how far individual stock returns deviate from the market return on a particular day. From a theoretical perspective, if there's no herding, the larger the absolute market return is, the higher the dispersion is, since companies possess diverse exposures and have faced several different news. In case herding moves the individual returns towards the return of the market, it means that the dispersion is not as high as expected in cases of abrupt changes on the market, or is significantly lower. According to CCK, a negative coefficient on $R^2_{m,t}$ indicates compression, and a positive one can be interpreted as continued growth in the dispersion.

According to the researchers, the VIX index is a good indicator of volatility, as it serves as a measure of market expected volatility for the upcoming days, and they associate the herding variables with the VIX variables, as herding fluctuates depending on the market conditions (Economou et al.).

The sample includes 439 observations in total covering the period from November 2, 2007, to July 31, 2009. Stress days are defined as those in the top quartile of VIX values, as during the period of investigation there were 110 highly stressful days ($VIX_t \geq 42.56$). The main conclusions include the following. First, the coefficient of the quadratic return isn't significant considering the entire sample, which means that the crisis by itself doesn't prove any herding. Second, separating different volatility regimes leads to the finding that the level of dispersion during the stressful days is higher by 0.0068. The last finding indicates a significant decrease of the convexity, from 10.85 on ordinary days to an implied 0.16 under stress, which stops short of the negative value the CCK criterion requires.

The paper consists of the literature review, data and methodology description, data analysis with robustness checks, discussion, limitations, and conclusion.

Background and Related Literature

Theoretical Foundations of Herding

Social learning models provide a framework through which the rational behavior of people can be related. The observers of the actions of others will tend to ignore their private signals when what those actions reveal outweighs what the observers know individually, and will thus follow the crowd, resulting in the formation of informational cascades. This process takes place swiftly and the behavior that results from it is highly fragile. A piece of new information is usually sufficient to stop the cascade. Bikhchandani et al. (1998) expanded the logic behind the informational cascades by showing that conformity can appear in markets whenever private information is limited, and not only in the narrow settings of the early cascade models.

There are various types of financial herding. Informational herding refers to the cases of herding described above. Reputational herding refers to those instances when people who manage funds are under the hazard of negative repercussions of deviating from other professionals' actions. Constraint-based co-movement occurs when various restrictions (in the form of concentration limits, calls for margin, or pressure to de-leverage) make a number of institutions liquidate the same assets at once, thus causing coordinated behavior with no imitation motive behind it (Bikhchandani and Sharma). One can argue that during 2007–2009 all the aforementioned types of herding took place simultaneously, even though the data cannot separate them.

Observing herding directly is highly challenging, as the data on the holdings that market investors have is usually late and not complete, while observing trade-level data is possible only with data access. Consequently, academic studies tend to infer signs of herding mainly from the market-level effects that can be traced to the return co-movement or dispersion, despite the fact that it could be caused by events not related to any imitation on behalf of the market investors (Bikhchandani and Sharma).

Measuring Herding Through Dispersion

In 1995 the first structured dispersion test was proposed by Christie and Huang, whereby it was suggested to identify the phase when the cross-sectional return dispersion is decreasing in conjunction with extreme market moves. If it is observed, the inference could be made about stocks clustering more closely around consensus rather than reflecting individual characteristics of the firms. This is the key feature of the study since the test employed only price data, allowing it to be practically executed and replicated.

CCK refined this method using regressed CSAD on both the absolute return and its square. A negative coefficient on the squared term is the criterion for identifying herding in their approach, indicating that each time extreme moves of the market occur, dispersion grows slower rather than faster. A positive coefficient would mean that rather than falling, dispersion continues to grow, thus contradicting the notion of herding. What they found became important for the rest of this work. Starting from testing this method in five markets, they found out that no herding behavior can be traced in the case of the US and Hong Kong, partial evidence is predominant in Japan, and credible herding was observed only in South Korea and Taiwan. The finding of no

herding signature for the US in the original study serves as the key reference point for this paper, which deals with US data gathered in a turbulent period.

Later experiments yielded different results depending on the market studied. Filip et al. (2015) in their study also relied on the same CSAD approach and Newey–West inference, and found that herding does take place in some Central and Eastern European countries but not in Poland or Romania. One reading of that inconsistency is that herding is not a stable feature of a market and occurs under certain conditions. The closest approach to this current research was taken by Economou et al., who included the use of implied-volatility fear indices in the CSAD method for the US, UK and Germany from 2004 to 2014, arriving at the conclusion that the presence of fear is statistically important for the herding estimates. Studies conducted by various authors state that clustering of the stocks during downward movement occurs more than during the upward move (Chang et al.; Christie and Huang). This theory is being tested in the current research.

However, one needs to be aware of the limitations of CSAD. When we speak about large market movements, compression could be explained by macroeconomic shocks, sector news affecting whole industries, or by index effects, so there is no need to talk about imitation patterns in this case (Christie and Huang; Bikhchandani and Sharma).

VIX and Stress-Regime Studies

The VIX is a standard volatility benchmark in empirical finance. The VIX is based on S&P 500 option prices and it is an indication of market expectations with regard to short-term volatility. The VIX level goes high if the perceived risk increases (Whaley). A VIX percentile threshold method for dividing the sample is in line with studies that see volatility as a regime variable instead of a control variable (Economou et al.).

As per research that links VIX levels to co-movement, high-VIX episodes differ structurally and not only in the magnitude of volatility (Wang et al.). With the rise in the VIX there are changes to the funding conditions, leverage and risk appetite, which is the condition when herding mechanisms are likely to emerge. Dating the recession (Labonte, National Bureau of Economic Research) verifies that the period 2007–2009 was a real period of stress and not a common period of high reported volatility.

Where This Paper Fits

Research suggests that herding is theoretically sound, that the CSAD approach is a possible way even if indirect to establish the existence of herding, and that volatility regimes are empirically meaningful states. However, there have not been sufficient studies investigating how these concepts are related. Most of the existing research measuring CSAD calculates one parameter throughout the entire sample period. If the nature of the relationship between returns and dispersion alters with volatility states, then a single estimate averages those states together and may describe neither of them. This paper aims to establish this relationship directly by

examining data from the USA during the economic crisis and calculating both the aggregate and state-specific coefficients so that they can be compared with each other.

Data

Sample Window and Stock Universe

The sample runs from 02/11/2007 to 31/07/2009. This is equivalent to 439 trading days covering the acute phase of the crisis until the end of the NBER-designated recession (Labonte; National Bureau of Economic Research). The analysed cross-section is a fixed list of 101 current S&P 100 stocks, which makes it possible to avoid mechanical dispersion changes arising from the fact that index membership changes over time. This means that the list is a present-day snapshot instead of the historical membership, which will be discussed later.

Returns and Market Proxy

The closing prices are used to calculate daily log returns. The return on stock i on day t is $R_{i,t}$ and $R_{m,t}$ is the log return of the market, which is calculated from the daily data on the S&P 500 index. Stocks that do not have returns for a given day are eliminated from that day's cross-section of stocks. This way of using the broad market return rather than a weighted average of the 101 stocks avoids constructing the market series mechanically out of the same cross-section used to build CSAD.

Dispersion Measure

CSAD is the average absolute distance between each stock's return and the market return on a given day:

$$CSAD_t = \frac{1}{N_t} \sum_{i=1}^{N_t} \left| R_{i,t} - R_{m,t} \right|$$

where N_t is the number of stocks with a return available on day t. Higher CSAD means returns are spread further from the market; lower CSAD means they cluster near it.

Units

All return variables ($R_{i,t}$, $R_{m,t}$, $\left|R_{m,t}\right|$, $R_{m,t}^2$) are daily log returns in decimal form, where a 1% move is approximately 0.01. CSAD uses the same units, so 0.0177 corresponds to an average absolute deviation of roughly 1.77 percentage points. Because $R_{m,t}^2$ is the square of a small decimal it is about two orders of magnitude smaller than $\left|R_{m,t}\right|$ (sample means 0.00054 and 0.0158). Its coefficient is correspondingly larger, so the two terms must be read jointly rather than compared by magnitude.

VIX Stress Indicator

High-volatility days are identified using a binary indicator:

$stress_{high,t} = 1$ if VIX_t is in the top 25 percent of VIX values within the sample window

$stress_{high,t} = 0$ otherwise

In this sample, the 75th percentile of the VIX is found to be 42.41, while the lowest measurement that can be categorized as stress is 42.56, which translates into 110 days of stress (25.1% of the total 439). The quartile cut captures a period of time that has been characterized by high volatility and is not tied to specific calendar dates. Robustness checks are done using the top 10% ($VIX_t \geq 54.34$; 44 days) and top 33% measures ($VIX_t \geq 36.19$; 145 days).

Methodology

Pooled CCK Specification

The starting point is the CCK regression applied to the full sample:

$$CSAD_t = \alpha + \gamma_1 |R_{m,t}| + \gamma_2 R_{m,t}^2 + \varepsilon_t$$

The herding test is the sign of γ_2. A negative and significant estimate indicates dispersion compressing as market moves grow extreme. A positive estimate indicates continued expansion. This model imposes one relationship on the entire window.

Regime-Interaction Specification

To let the relationship differ across volatility states, the model adds $stress_{high,t}$ both as a level term and interacted with each regressor:

$$CSAD_t = \alpha + \beta_1 |R_{m,t}| + \beta_2 R_{m,t}^2 + \beta_3 stress_{high,t} + \beta_4 (|R_{m,t}| \cdot stress_{high,t}) + \beta_5 (R_{m,t}^2 \cdot stress_{high,t}) + \varepsilon_t$$

The importance of the level term β_3 should be noted here. Mean CSAD is found at 0.0151 on non-stress days and 0.0254 on stress days, giving a difference of 0.0102 in values. Excluding the level term and keeping the interactions, we would require both regimes to share a common intercept, and that difference would have to be accounted for by the slope and curvature terms, distorting both. Missing the main component of an interaction is a standard specification error and has serious implications here. An F-test of the restriction yields $F = 38.16$ ($p < 0.001$), while the implied curvature for stress days moves from −1.48 when the level term is excluded to +0.16 when it is included in the model. Moreover, the unrestricted version of the model matches regime-by-regime separate regressions exactly, which cannot be said about the version where the level term is excluded.

Implied coefficients by regime are:

- Non-stress days: slope β_1, curvature β_2.
- Stress days: slope $\beta_1 + \beta_4$, curvature $\beta_2 + \beta_5$, intercept shifted by β_3.

Estimation and Inference

The estimation method used is OLS. Daily return data has a tendency to display heteroskedasticity and autocorrelation behavior. Both phenomena increase the misleading precision of the OLS standard errors, and for this reason all the reported estimates use Newey–West HAC standard errors (Newey and West). The lag used is 5 days, which is close to the Newey–West rule-of-thumb bandwidth of 4(T/100)^(2/9) for T = 439. Polynomial and interaction variables are collinear by their nature and variance inflation factors for the squared variables exceed 180. This inflation is mechanical and does not bias the coefficients. However, individual t-tests on the collinear variables should be interpreted in the light of joint tests and implied regime coefficients. All the results obtained are associations; causality is not being claimed.

Results

Descriptive Statistics

Table 1 reports summary statistics for the full sample and by regime.

Table 1. Descriptive Statistics, 2007–2009 (N = 439; 329 non-stress, 110 stress)

Note: Returns and CSAD are daily log returns in decimal form. VIX in index points. Stress threshold VIX ≥ 42.56 (top quartile).

Variable	Mean	SD	Min	Max	Non-stress Mean	Stress Mean
$CSAD_t$	0.0177	0.0076	0.0067	0.0516	0.0151	0.0254
$\lvert R_{m,t} \rvert$	0.0158	0.0171	0.0000	0.1356	0.0113	0.0292
$R^2_{m,t}$	0.00054	0.00146	0.00000	0.01837	0.00022	0.00149
VIX_t	33.30	13.79	16.30	80.86	26.52	53.57
$stress_{high,t}$	0.251	0.434	0	1	0	1

Stress days differ from ordinary days in every dimension. Mean $CSAD_t$ is 0.0254 against 0.0151, a gap of 68%. Mean absolute market return is 0.0292 against 0.0113, and the largest single-day move in the sample is 13.6%. VIX averages 53.6 in the stress regime and 26.5 outside it, with no overlap at the boundary by construction. The size of the CSAD level gap is what makes the specification choice in the previous section consequential.

Visual Pattern

Figure 1 plots CSAD against the absolute market return for each regime with separate linear fits.

Fig 1: CSAD against |R□,□| for non-stress days (N = 329, blue) and stress days (N = 110, red), with separate linear fits. Stress observations sit at a higher CSAD level throughout and extend across a much wider range of market moves. The stress-day fit is flatter, and the two lines cross near |R□,□| ≈ 0.06. All variables in decimal form.

Pooled and Regime-Interaction Estimates

Table 2. CSAD Regressions: Pooled CCK Model and Volatility-Regime Interaction Model

*Note: N = 439. Newey–West HAC standard errors in parentheses, lag = 5. Model 1 is the pooled CCK specification; Model 2 adds the stress level term and interactions. Asterisks mark statistical significance: *** p < 0.001, ** p < 0.01, * p < 0.05; no asterisk means the estimate is not distinguishable from zero.*

Term	Model 1 (Pooled CCK)	Model 2 (Regime interaction)
α $(intercept)$	0.0126*** (0.0005)	0.0134*** (0.0006)
$\lvert R_{m,t} \rvert$	0.3440*** (0.0334)	−0.0586 (0.0910)
$R^2_{m,t}$	−0.6639 (0.3699)	10.8510*** (2.6374)
$stress_{high,t}$	—	0.0068*** (0.0017)
$\lvert R_{m,t} \rvert \cdot stress_{high,t}$	—	0.2280* (0.1145)
$R^2_{m,t} \cdot stress_{high,t}$	—	−10.6958*** (2.6847)
$Implied\ stress\ -\ day\ R^2_{m,t}$	—	+0.155
R^2	0.4377	0.5712
Adjusted R^2	0.4351	0.5663

Model 1 pools the window. The linear term is positive and strongly significant ($\gamma_1 = 0.3440$, HAC t = 10.29), so dispersion rises with the size of market moves. The squared term is negative but falls short of conventional significance ($\gamma_2 = -0.6639$, HAC t = −1.79, p = 0.073). On the CCK criterion the pooled sample therefore yields no reliable herding evidence, which reproduces what CCK themselves found for U.S. data on a very different and calmer sample.

Model 2 allows the regimes to differ, and the fit improves markedly: adjusted R^2 rises from 0.4351 to 0.5663. The stress level term is positive and highly significant ($\beta_3 = 0.0068$, HAC t = 4.10), confirming that dispersion sits structurally higher when volatility is elevated. On ordinary days the relationship is strongly convex ($\beta_2 = 10.8510$, HAC t = 4.11) with a flat linear component. The curvature interaction is large and negative ($\beta_5 = -10.6958$, HAC t = −3.98), which almost exactly offsets the baseline convexity.

The implied stress-day curvature is $\beta_2 + \beta_5 = +0.155$. That value is close to zero and not distinguishable from it. The substantive reading is that the convex relationship visible on ordinary days disappears in the stress regime: dispersion stops accelerating as market moves grow. It does not become negative, so the CCK herding criterion is not met in either regime. What changes across regimes is the presence of convexity, not its direction.

Robustness

Table 3 re-estimates Model 2 under five alternative designs.

Table 3. Model 2 Under Alternative Thresholds, Outlier Treatment and Market Direction

*Note: S denotes the stress indicator. Coefficients estimated with Newey–West HAC standard errors, lag = 5. Winsorized clips market returns at the 1st and 99th percentiles (10 days). Direction splits use signed market returns. Asterisks mark statistical significance: *** p < 0.001, ** p < 0.01, * p < 0.05; no asterisk means the estimate is not distinguishable from zero.*

Term	Baseline Top 25%	Top 10% (44)	Top 33% (145)	Winsor. 1%	Down (215)	Up (224)
$\lvert R_{m,t}\rvert$	−0.059	0.198***	0.023	−0.059	−0.203*	0.077
$R^2_{m,t}$	10.85***	2.73**	7.67*	10.85***	12.74***	8.67
S	0.0068***	0.0128***	0.0071***	0.0075***	0.0082***	0.0055**
$\lvert R_{m,t}\rvert \times S$	0.228*	−0.211	0.165	0.136	0.267*	0.196
$R^2_{m,t} \times S$	−10.70***	−1.47	−7.61*	−8.74**	−11.78***	−9.19
$Implied\ stress\ R^2_{m,t}$	+0.155	+1.260	+0.062	+2.109	+0.963	−0.516
R^2	0.5712	0.5120	0.5997	0.5613	0.6221	0.5603

The stress level variable stands out in each of the six columns. The HAC t-statistics for this variable range from 2.90 to 5.66, making the stress term the only variable to generate these results in all columns. The results indicate that this dataset supports the conclusion that there is a big difference in the level of dispersion during periods of turbulence.

The results related to the interaction between curvature and stress are less robust. While the interaction is statistically significant for the baseline case, at the top-33% level, when the results are winsorized, and for the down days, it is not significant at the top-10% level ($t = -1.13$) or for the up days ($t = -1.69$). The top-10% column is the most informative failure: restricting the stress regime to the 44 most extreme days does not leave enough observations to identify a change related to curvature, and the baseline value of convexity itself drops to 2.73.

For the herding test, the quantity that matters is the implied stress-day curvature, and it is positive in five of the six columns, varying from +0.062 to +2.109. There is only one negative number, namely −0.516 on the up days, and it belongs to a column where the interaction is not significant. No specification produces a stress-day coefficient that is both negative and reliably estimated, which is a stronger statement than any single column can make on its own.

By dividing the days into up and down we get, respectively, 215 down and 224 up days. The results for both level shift and curvature show that these variables are larger and more precisely estimated for the down days, which is in line with the current evidence that there is a concentration of co-movement on the downward days (Chang et al.; Christie and Huang). Moreover, the model that fits down days best appears to be the one which is reported in the table above.

Discussion

Interpreting the Regime Difference

Volatility, when it is high, affects two aspects. One is that the level of dispersion increases by an amount close to 0.0068, or 45% of the non-stress mean level. The second aspect is that the relationship between dispersion and the size of market movements changes from a convex type to a flat one, in connection with a fall in the curvature coefficient from 10.85 to an implied 0.16. This, however, does not amount to CCK herding, because one of its conditions is that the curvature should be negative.

Still, the significance of this flattening is considerable. In normal situations, larger market movements would lead to a much larger dispersion effect, since larger shocks create more differences in how firms react. However, this amplification doesn’t take place in stress periods. In such situations, dispersion is high no matter what, so it is roughly as high on a moderate stress day as on an extreme one.

Relation to Existing CSAD Findings

CCK's conclusions for the U.S. fit neatly with the pooled results. They found no evidence of herding in the U.S. markets based on their own dataset, and the pooled estimate here is

negative but not statistically significant in a turbulent sample. If the crisis period does not produce evidence of herding in big U.S. stocks, it is not surprising that calmer periods also do not show it.

One way of interpreting the mixed results in the CSAD literature is that pooled estimates are too broad. This concern is valid, and it can be shown in Model 2, which demonstrates how different regimes behave differently with respect to the CSAD measure. At the same time, however, the division among the regimes does not confirm the herding hypothesis in our case. It reveals a different structure rather than a hidden herding effect.

The S&P 100 is important in this discussion. These companies are the biggest U.S. companies and are well studied, have high institutional ownership, and are widely traded. Herding effects are strong in emerging markets and among retail investors, which corresponds with what CCK found. A large-cap U.S. universe is among the least likely places to find dispersion compression, which has to be considered when interpreting the result.

Why Dispersion Might Behave This Way Under Stress

Numerous processes may generate increased levels of dispersion along with a decline in convexity. During periods of crisis, issues related to credit and solvency that are firm-specific receive attention. Financials were exceptionally different from the rest of the index in 2008, increasing the level of dispersion without requiring any relationship to the size of the market move. Constraint-led selling demonstrates the phenomenon as well: institutions, engaged in deleveraging, sell what is liquid rather than what is falling, separating individual movements of the stocks from the overall market (Bikhchandani and Sharma).

The decline in convexity may be closer to mechanical. On ordinary days, big moves in the market are unusual and usually follow significant news. However, during the periods of extreme market stress, large moves occur routinely, which means that an additional percentage point of market movement carries less information. Separating this from a behavioral explanation would require data beyond prices.

Answering the Research Question

The inquiry was focused on the degree to which herding indicators correlate with increased volatility during market crises. The measure of dispersion is seen as being related to the volatility regime: the CSAD measure presents at a higher level with a strikingly different shape under times of elevated VIX, and this difference is statistically strong and robust. However, the distinctive CSAD pattern associated with herding does not show up, in either regime or in any robustness specification. To sum it all up, there is no denying that volatility regimes affect the way dispersion behaves, and this particular test has failed to show evidence of herding in U.S. large caps even at the height of the crisis.

Implications

In the realm of risk management, the level result provides the greatest utility. When VIX is high, cross-sectional dispersion climbs sharply and in a sustained manner that is not well explained by the magnitude of the market move on any given day. Any model that has been calibrated using calm periods is bound to underestimate the dispersion during crisis times, even conditional on market volatility.

For the herding literature, the principal lesson relates to model specification. The restricted model that excludes the stress level term yields an implied stress-day coefficient of −1.48 and appears to indicate that herding is present. The correctly specified model gives a coefficient of +0.16 and shows none. The difference between the two models rests on a single dummy variable; one may easily overlook this discrepancy, since the restricted model still returns a respectable $R^2 = 0.53$.

Limitations

The CSAD is a proxy measure. Higher dispersion or a reduction in convexity does not automatically indicate herding, and may instead point to common macro shocks, sectoral news or even index composition effects (Christie and Huang; Bikhchandani and Sharma). A null result on the CCK criterion rules out one specific signature, but does not establish that there is no herding.

Design is purely observational. Dispersion and volatility both respond to the same shocks, such as freezing of funding markets, Lehman's default, or policy announcements, and there is nothing indicating the direction of effect here.

The cross-section uses a recent 101-ticker S&P 100 snapshot applied to a 2007–2009 window. Some companies in it were not index members during the crisis, and companies that were included back then are missing today. The firms that failed or got bought are the likely candidates for having shown the biggest dispersion, and their absence therefore suggests the measure is understated during the stressed period.

The cross-section refers to the S&P 100 while our market return refers to the S&P 500, which differ in terms of number of constituents and sector weights. The difference is small at the daily level, however it still exists.

Variance inflation factors on the squared and interacted terms exceed 180 by construction. This does not lead to bias in coefficients, however it makes the top-10% column difficult to interpret, because the stress regime there comprises only 44 observations.

At last, the stress regime is constructed in-sample. Setting a threshold within the crisis period identifies the most volatile quarter of an already volatile period, which means the non-stress comparison group is not a calm market. A pre-crisis benchmark would yield a sharper comparison.

Conclusion

This paper assessed if the dispersion among returns behaves differently when volatility is at high levels in U.S. equities from 2007 to 2009 by means of CSAD, a fixed 101-stock S&P 100

cross-section, and a stress threshold defined by the top quartile of VIX within the CCK regression framework.

By pooling the window together, the squared return coefficient is equal to −0.664 with HAC t = −1.79, which does not reach significance, so the crisis sample alone does not supply any evidence of herding with regard to the CCK approach. On the other hand, allowing for the difference in regimes enhances the adjusted R^2 from 0.435 up to 0.566 and provides two results. Firstly, on stressful days, the dispersion is above the ordinary level by 0.0068, which has been proven significant in every model tested. Secondly, the convexity of the dispersion–return relationship is lessened from 10.85 down to an implied 0.16, thus the amplification normal days show vanishes in stressful times. The implied coefficient on stress days is equal to or above zero in five out of the six robustness checks, with the one negative number coming from the column where the interaction being calculated is insignificant.

Thus, the volatility regime alters the way dispersion operates without producing the compression that would indicate herding. This conclusion is consistent with that made by CCK, who found no herding in the U.S. market, and it expands the scope of the study to a crisis period the CCK approach was not applied to. It also shows that the pooled estimate conceals a substantial difference between regimes in both level and curvature, even though the herding conclusion itself is unchanged.

The lack of the stress level term while keeping the interactions leads to an implied stress-day coefficient of −1.48, which appears to provide distinct evidence of herding. It thus appears to be necessary to report the main effect together with the interaction terms when conducting this kind of regression analysis.

There are several extensions that may be applied to improve the paper quality. Past constituent lists will eliminate the survivorship bias, which would be most essential for the financial companies that caused the volatility in the first place. The period before the crisis (2004–2006) should be analyzed to see if the regime difference is observed only in crisis time. A binary threshold should be replaced by the continuous VIX interaction to find out the point of the volatility distribution where convexity ceases to exist. Lastly, it would make sense to combine dispersion testing with volume or holdings-based herding measures to delineate movements driven by limitations from those generated by people’s beliefs only.

Works Cited

Akter, Nahida, and Ashadun Nobi. "Investigation of the Financial Stability of S&P 500 Using Realized Volatility and Stock Returns Distribution." Journal of Risk and Financial Management, vol. 11, no. 2, 2018, article 22, https://doi.org/10.3390/jrfm11020022.

Banerjee, Abhijit V. "A Simple Model of Herd Behavior." The Quarterly Journal of Economics, vol. 107, no. 3, 1992, pp. 797–817, https://doi.org/10.2307/2118364.

Bikhchandani, Sushil, and Sunil Sharma. "Herd Behavior in Financial Markets." IMF Staff Papers, vol. 47, no. 3, 2001, pp. 279–310, https://doi.org/10.2307/3867650.

Bikhchandani, Sushil, et al. "A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades." Journal of Political Economy, vol. 100, no. 5, 1992, pp. 992–1026, https://doi.org/10.1086/261849.

---. "Learning from the Behavior of Others: Conformity, Fads, and Informational Cascades." Journal of Economic Perspectives, vol. 12, no. 3, 1998, pp. 151–70, https://doi.org/10.1257/jep.12.3.151.

Chang, Eric C., et al. "An Examination of Herd Behavior in Equity Markets: An International Perspective." Journal of Banking & Finance, vol. 24, no. 10, 2000, pp. 1651–79, https://doi.org/10.1016/S0378-4266(99)00096-5.

Christie, William G., and Roger D. Huang. "Following the Pied Piper: Do Individual Returns Herd around the Market?" Financial Analysts Journal, vol. 51, no. 4, 1995, pp. 31–37, https://doi.org/10.2469/faj.v51.n4.1918.

Economou, Fotini, et al. "Investors' Fear and Herding in the Stock Market." Applied Economics, vol. 50, nos. 34–35, 2018, pp. 3654–63, https://doi.org/10.1080/00036846.2018.1436145.

Filip, Angela, et al. "The Herding Behaviour of Investors in the CEE Stocks Markets." Procedia Economics and Finance, vol. 32, 2015, pp. 307–15, https://doi.org/10.1016/S2212-5671(15)01397-0.

Labonte, Marc. The 2007–2009 Recession: Similarities to and Differences from the Past. CRS Report R40198, Congressional Research Service, 7 May 2010.

National Bureau of Economic Research. "US Business Cycle Expansions and Contractions." NBER, www.nber.org/research/data/us-business-cycle-expansions-and-contractions. Accessed 29 Aug. 2026.

Newey, Whitney K., and Kenneth D. West. "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix." Econometrica, vol. 55, no. 3, 1987, pp. 703–08, https://doi.org/10.2307/1913610.

Wang, Wei, et al. "CBOE Volatility Index (VIX) and Corporate Market Leverage." International Review of Financial Analysis, vol. 90, 2023, article 102837, https://doi.org/10.1016/j.irfa.2023.102837.

Whaley, Robert E. "Understanding the VIX." The Journal of Portfolio Management, vol. 35, no. 3, 2009, pp. 98–105, https://doi.org/10.3905/JPM.2009.35.3.098.

The Game Theory Behind the Evolution of the Korean Entertainment Industry
By Isaac Ahn

Abstract

Nearing the end of the 20th century, South Korea made a large pivot into exporting popular culture in an attempt to recover from the recent 1997 Financial Crisis. This pivot was massively successful as South Korea made itself a staple of mainstream media in contemporary global society. The permeations of Korean culture into the global atmosphere of pop culture became known as *hallyu*, quite literally translating to "Korean Wave." However, there is more than what meets the eye in the South Korea's entertainment industry. In order to truly understand what made South Korea's entertainment industry so successful on the global scale, *hallyu* cannot be seen as just a cultural phenomenon and instead has to be seen through the lens of political game theory. The entertainment industry is a complex strategic environment, with many stakes, payoffs, rewards, and information to be spread. Recently, there has been a sudden rise of independent artists in South Korea's entertainment industry. Here, political game theory is a necessary tool to help understand these recent trends. Through concepts such as coordination games and information asymmetry, game theory rationalizes the behaviors of the many actors in the environment such as artists, the companies that hire artists, the fans of the artists, and the government that oversees the companies.

Introduction

In the late 1990s, South Korea began one of the most successful pivots into becoming a cultural driver. The country had just exited out of a brutal financial crisis, known as the 1997 Asian Financial Crisis, and was now facing, along with the rest of Asia, a devastating recession. At the same time, censorship over South Korea's media had come to an end and satellite technology was on the rise. These three conditions made *hallyu*, the global rise in interest in South Korean entertainment and pop culture, nearly inevitable. Coining specific Korean media with the prefix "K"–such as K-dramas, K-pop, K-beauty, and K-food–South Korea was able to increase its international reach and immerse the world with Korean culture. As a result, Korean culture spread in two waves over the next two decades, the first wave marking its popularity among neighboring Asian countries, and the second wave marking its popularity among global audiences further out such as Europe and the Americas. The spread of K-pop in particular was extraordinarily dominant, with many K-pop groups and bands becoming mainstream in the music industry. K-pop bands like BTS boosted South Korea's economy greatly and also drew in many tourists. This further opened up the field to increased sales of Korean products of electronics and cosmetics. While the world consumed the polished and picture-perfect images of smiling and attractive idols, the arrival of K-pop scandals were soon to follow, with numerous lawsuits and exposés that revealed the terrible conditions that these artists were subjected to.

Now, in the modern day, there has been a major shift for artists within the South Korean entertainment industry to become independent artists rather than work under a contract for a

company. Keeping in mind both the history and the evolution of the Korean entertainment industry as well as the large volume of lawsuits that have exposed the exploitative practices in the industry, this paper explores the dynamics of the entertainment industry through the lens of game theory by analyzing the various players on stage and the strategic environment that motivates them towards certain trends.

Background

In the 1960s, Park Chung Hee had expanded South Korea's mass media by heavily investing into radio broadcasting infrastructure through his power as military dictator. (Kim) This expansion helped modernize South Korea and would set the stage for the first Korean Wave, or *hallyu*. This expansion, however, also came at the cost of strong censorship which included heavily cracking down upon independent journalism. (Kim) Under the Yushin Constitution, an authoritarian document that was meant to follow in the footsteps of Japan's successful Meiji Restoration plan, Park also heavily censored foreign imported media, such as American protest songs, under the guise of preventing foreign culture from sullying Korean culture. In reality, the censorship was an attempt to keep the country distracted from political matters while he sought to consolidate more power as dictator and suppress the masses. (Kim) Following Park's assassination in 1979, Chun Doo-hwan had presumed control over the military regime. Chun's unpopularity and heavy media backlash led to him forcing the merging of all 29 Korean broadcasting stations into the Korean Broadcasting System. With only one broadcasting system for all of South Korea, Chun had more control over the media, allowing him to suppress all critical views about him.

The dawn of the 1990s began with greater liberties being granted to Korea's media industry. One of them being the creation of the first private television network since Chun's broadcasting merger in the '80s, the Seoul Broadcasting System (SBS). Following the aftermath of the 1997 Asian Financial Crisis with the collapse of chaebols, which were family-owned business groups, the then-current administration under President Kim Dae-jung had come to understand that South Korea as a country and its economy could not uphold through traditional manufacturing means. (Kim and Wakabayashi) President Dae-jung saw culture as more than just an art form and believed it to be a key to establishing a domineering presence on the world stage. (PGP Capital) This government recognition of entertainment and Korean culture serving as a driver for the economy allowed for the dissolution of tight censorship and restrictions enacted decades prior. As a result, previously banned Japanese media was now available–having been banned after the liberation of South Korea from Japan–which provided much more creative insight and experimentation into creating new Korean media. All of this gave way to SBS being able to export more of its media abroad into neighboring countries.

Many countries that were affected by the Asian Financial Crisis of 1997 quickly accepted and aired the K-dramas being exported by South Korea as they were only a fraction of the cost that it took to import typical popular Japanese media. As a result, many East Asian broadcasters began airing K-dramas in the early 2000s, where they quickly began to gain traction due to the

success of various K-dramas such as *Firework* and *Jewel in the Palace*. These breakthroughs in global culture essentially marked the first wave of *hallyu*, when Korean entertainment became mainstream among its neighboring East Asian countries.

In the late 2000s, following the rise of social media and the internet, the second Korean Wave had begun. While the first *hallyu* wave was reliant on TV exports, the second *hallyu* wave primarily spread through digital platforms such as YouTube and Twitter. These digital platforms were essential as it further enhanced the global village, a concept where all people are linked into a single community by telecommunications, which allowed fans to bypass traditional radio limitations and access Korean content. The increased accessibility to Korean content gave way to a key turning point, when PSY's *Gangnam Style* gained huge traction and popularity worldwide without traditional market support. (Jin) Digital platforms during this era also allowed for fans to act as unpaid marketers: translating lyrics and variety shows into many languages, creating trendy hashtags that would catch the eyes of Western media, and organizing bulk-buying campaigns that would push Korean artists onto Western charts, such as the US Billboard 200. (Lee)

This rapid spike in popularity during the second *hallyu* wave as well as government support helped establish large entertainment companies in the Korean entertainment industry, some of which being SM Entertainment, YG Entertainment, and JYP Entertainment. By offering tax incentives, providing funding for startups, and constructing advanced technological concert venues, the South Korean government made it very easy for entertainment companies to thrive and compete on a global scale. These large entertainment companies were responsible for scouting for new talent, training promising recruits, evaluating their performances, and deciding whether or not to allow them to debut as a new group. Once under an entertainment company, artists typically have their autonomy restricted severely. They are placed on restrictive diets, given intense and frequent training sessions, and monitored very closely. Additionally, companies typically take 80-90% of the revenue generated and place them into debt for the training they received, leading to artists earning very little or essentially working for free. This style of training idols quickly became a standard for the industry and was essentially the framework for the *hallyu* wave. (Magnus-Sharpe)

In recent years, there has been a rise in artists becoming independent and circumventing working with a company. The main benefit being the increase in personal and creative freedom. Additional benefits included not having to suffer through the harsh conditions mentioned earlier and being able to have more autonomy in how they interact with audiences and showcase themselves. This freedom is what has made becoming an independent artist a very attractive option for both current and rising pop artists in South Korea.

Setting the Game Theory Stage

To better understand the trends in South Korea's entertainment industry and the rise of Korean culture in mainstream media, it must be understood that *hallyu* has more substance than just as a cultural phenomenon. The nature of *hallyu* lends itself to a deeper and more complicated

strategic environment between entertainment companies and artists. This is where game theory, the study of how individuals behave when other people's actions can have an effect on their outcome, can help explain how the Korean entertainment industry functions and evolves. In particular, the specific market behaviors in the Korean entertainment industry are entirely dictated by the foundational rules of political game theory. (Dixit) The actions of a player in game theory is never done in isolation, and usually holds consequences, both positive and negative, for other players in the environment. In fact, the many choices of each player interact and generate a matrix that include both sub-optimal and optimal outcomes for each player. These outcomes account for both short term and long term effects of a choice. For instance, an artist's decision to pursue a choice that guarantees financial stability within the short term will also have lasting consequences in the long term such as intellectual property or burnout.

In this strategic environment, there are players, rewards, and information. Players in this environment include entertainment companies, actors, artists, and fans. Rewards in this environment include fame, money, freedom, and reputation. Information in this environment signifies bits of information that various players may or may not be aware of. Every player in this environment behaves with the intent of obtaining distinct rewards, and will thus optimize their position within the given constraints, such as limited information or written contracts, in order to maximize their chances of obtaining the reward.

For fans and fandoms, their role in this environment can transcend to more than just consumers. Fans have the ability to organize and mobilize resources that can either push artists, and by proxy the company, to greater heights or cast them aside and make them lose relevancy in the pop industry. Especially with digital platforms, fandoms quickly evolved into massive global digital armies that hold significant power and dominance in the industry. Fans have often coordinated mass-streaming songs to help them break into top charts and break world records. Additionally, as mentioned previously, fans also help bridge language barriers by performing language translations and explaining cultural nuances to newer fans. These activities lower the barrier to entry for non-Korean speakers, allowing the spread of Korean culture even more prevalent.

The relationship between artists/actors and the companies that hire them is best modelled by the principal-agent problem, where companies are the principals, entities that hire individuals for their talents, and artists and actors are the agents, people that are hired to perform under the company and grant them prestige and money. Within the companies themselves, the system that best describes them is an oligopoly, where a few companies dominate the majority of the market. This oligopolic structure ultimately sets the stage for inevitable systemic conflicts that will eventually escalate and catch the attention of the public sphere. (PGP Capital)

The Strategy During Early *Hallyu* and Challenges to the Status Quo

In the early phases of the second wave of *hallyu*, as K-pop became more widespread, so did the exploitation of artists. However, fans were unaware of this severe exploitation, which introduces information asymmetry in the strategic environment. This information asymmetry

allowed for exploitation to have a high reward payoff for the entertainment companies, where they can continuously put their artists through terrible conditions and still gain much money and reputation. For artists, they are dependent on the company due to the very limited options they had on the path to stardom during the early phases of *hallyu*. If an artist attempted to terminate their contract to switch to a different one, it came with a high risk of being blacklisted. Since digital platforms were still in their early phases, an artist attempting to go solo would have little to no chance to become successful enough to make a living since they did not have the resources, ability, or outlet to make themselves known. Thus, the risk of going independent or switching companies was too high for the payoff to be worth it for the majority of artists. The government of South Korea is also incentivized not to make any changes to the game and to not interfere with the strategies of the entertainment companies, as the entertainment companies greatly drove South Korea's economy and helped establish South Korea as a global soft power with newfound potential to become a global leader.

However, the information asymmetry that allowed for the continuous exploitation of artists did not stay that way forever. As K-pop became more mainstream in the Western hemisphere, news of lawsuits that exposed the exploitation of artists also became more prevalent. The shock of the abuse artists were put through massively changed the strategic environment in the entertainment industry. With so much backlash now being received from fans, companies were forced to make changes to the main framework of the idol industry and adapt. (Lardner, Woo & Schuette). One such example was the lawsuit surrounding 3 members of the band TVXQ and SM Entertainment, the company that the members were contracted under. The members had come out publicly to state that they had been unfairly treated, citing extreme working conditions and large disparities in profit distribution. The legal battles that ensued, along with the media investigations and documentaries that brought to light the immense pressure and standards set upon artists, brought global criticism and backlash, and mobilized fans to become active advocates and demand better treatment and mental health support for artists in the industry. The media also amplified these scandals which further shifted the public's view on Korean entertainment industries.

This change in strategy was definitely not due to changes in cultural shifts but rather because the payoff of such severe exploitation was significantly reduced for these companies. In fact, such exploitation now came with high risks such as damage to reputation, expensive lawsuits, and being boycotted by fans. Additionally, the entertainment company also faced pressure from government lawmakers to reform the harmful practices within the industry. After the exposés, the government could no longer remain silent for these entertainment companies as this would risk also damaging the reputation of the South Korean government. This prompted the government to take actions and enact legal reforms. More specifically, the Korean government reformed contracts by shortening contract lengths and providing increased legal protections for artists. (Dutta)

During this time, artists also now had a lot more viable options to their path to stardom. Previously, going independent was very risky with the risk outweighing the benefits severely.

Since many artists that go into the industry young do not have standard schooling to pursue other career paths, going independent and failing to become famous would lead to financial ruin.

However, the choice of whether or not going independent is in fact a coordination game, a situation where the best option is one where all players make the same choice. Artists that successfully transitioned into becoming indie artists early established themselves as strategic focal points. (Dixit) Thus, as more artists went independent and became successful, the barriers to entry for future indie artists kept decreasing and the indie scene kept gaining traction. Eventually, with so many artists going independent, the payoff began to significantly outweigh the risks. Additionally, with digital platforms now being very widespread, such as the prominence of TikTok, it reduces the barriers of entry even more for many artists. This is important because digital platforms are free and very accessible to the public while also containing the potential for virality. This virality helps artists become popular and potentially set them on the path to stardom, which means it is a very attractive choice compared to working under a company.

These changes to the industry created shifts away from manufactured idols and towards more authentic artists who weave in their personal identities with their music. (Rosario) The newfound independence for artists empowered them to break from the standardized formula for success and are able to address social issues, personal struggles, and cultural identity without risk of censorship.

Conclusion

Hallyu is a powerful cultural force that helped South Korea establish itself in the global sphere and drive its economy. The entertainment companies and artists that are put out by such companies give South Korea its strong reputation and influence. However, when companies exploit artists for far too long and such exploitation comes to light, the information asymmetry that gave these companies an advantage disappears. Thus, the strategic environment in the entertainment industry shifts, rapidly changing the payoffs and risks for each player. This leads to reduced payoffs for both the entertainment companies and the government, leading to mass legal reform in regards to how the industry treats artists. Additionally, the widespread use of digital platforms provides artists with an alternative and viable option to stardom. Coupled with these digital platforms, many artists then decide to go independent, establishing an indie scene that further erodes the barriers to entry and further increases the chances for success. This, in turn, incentivizes artists to go independent than to work under an exploitative company. This is what explains the recent trends in the Korean entertainment industry. While there is still much work to be done to protect artists fully, the recent shift towards transparency and artist autonomy is a good step towards a more sustainable future for the industry. Ultimately, game theory reveals that changes in industry patterns and the status quo are often not a sign of shifting cultural values, but rather a strategic choice made by the player for self-preservation and to ultimately, continue playing the game.

Works Cited

Cho, Joan E., et al. "Media Exposure and Regime Support under Competitive Authoritarianism: Evidence from South Korea." Journal of East Asian Studies, vol. 17, no. 2, 9 May 2017, pp. 145–166, doi:10.1017/jea.2016.41.

Dixit, A., & Nalebuff, B. (2002). Game theory . econlib. https://www.econlib.org/library/Enc/GameTheory.html

Dutta, D. (2026, March 17). Longshot: Jay Park's newest rookies are calling their own shots. Rolling Stone India. https://rollingstoneindia.com/lngshot-rookie-jay-park-k-pop-debut/

Evelyn Woo. "K-Pop's Global Rise Tests Labor Protections." The Regulatory Review, 16 Apr. 2026, www.theregreview.org/2026/04/18/seminar-k-pops-global-rise-tests-labor-protections/

Jin, D. Y. (2012). Hallyu 2.0: The new Korean Wave in the creative industry. International Institute Journal University of Michigan. https://quod.lib.umich.edu/i/iij/11645653.0002.102/--hallyu-20-the-new-korean-wave-in-the-creative-industry?rgn=main%3Bview%3Dfulltext

Jin, Dal Yong. Cultural Production of Hallyu in the Digital Platform Era: Industry Perspectives. University of Michigan Press, 2025. Project MUSE, https://muse.jhu.edu/book/138613.

Kim, et al. "The Park Chung Hee Era: The Transformation of South Korea." Weatherhead Center for International Affairs, www.wcfia.harvard.edu/publications/park-chung-hee-era-transformation-south-korea

Kim, M. H. (2016). Film Censorship Policy During Park Chung Hee's Military Regime (1960–1979) and Hostess Films. IAFOR Journal of Cultural Studies, 1(2). https://doi.org/10.22492/ijcs.1.2.03

Kim, and Chang. "Kim & Chang's Representation of Entertainment Agency in Preliminary Injunction to Preserve Management Agency Status and Confirm Validity of Exclusive Agreement: IP Stars." Asialaw, 25 Mar. 2026, www.asialaw.com/NewsAndAnalysis/kim-changs-representation-of-entertainment-agency-in-preliminary-injunction-to/Index/2569.

Kim, V., & Wakabayashi, D. (2023, December). What to know about the chaebol families that dominate South Korea's economy (published 2023). https://www.nytimes.com/2023/12/18/business/chaebol-south-korea.html

"Korean Wave." Wikipedia, Wikimedia Foundation, 1 Aug. 2026, en.wikipedia.org/wiki/Korean_Wave

Lardner, A., Woo, E., & Schuette, G. (2026, April 18). K-pop's global rise tests labor protections. The Regulatory Review. https://www.theregreview.org/2026/04/18/seminar-k-pops-global-rise-tests-labor-protections/

Law, Imogen. "Evolution of K-Pop Generations and Shifts in Corporate Dominance | by

Imogen Law | Medium.” Medium, Imogen Law, 5 Nov. 2025, medium.com/@imogenlaw2/evolution-of-k-pop-generations-and-shifts-in-corporate-dominance-87a00cfc1046.

Lee, J. S. (2024). Editor’s introduction. Journal of Japanese and Korean Cinema, 16(1), 1–5. https://doi.org/10.1080/17564905.2024.2339507

Lee, S., & Prey, R. (2025). The Labor Process of Relational Labor: The Case of the K-pop Fan Platform “Bubble.” Popular Music and Society, 48(4), 416–436. https://doi.org/10.1080/03007766.2025.2492505

Magnus-Sharpe, S. (2025, July). The K-pop status shuffle: Producers, power and Reinvention. Cornell Chronicle. https://news.cornell.edu/stories/2025/07/k-pop-status-shuffle-producers-power-and-reinvention

PG Capital Advisors (2023). The K-pop industry: General overview and outlook Fall 2023. (PGP Capital). https://pgpcapital.com/wp-content/uploads/2023/09/K-Pop-Industry-Overview-Fall-2023vF.pdf

Rosario, D. D. (2022, January 19). A year after freedom, got7 flies even higher | Inquirer Entertainment. A year after freedom, GOT7 flies even higher. https://entertainment.inquirer.net/434308/a-year-after-freedom-got7-flies-even-higher

Sanchez, JC. “The K-Pop Industry: How the Business Really Works.” Orphiq, 23 July 2026, orphiq.com/resources/korean-music-industry-kpop-ecosystem.

The Price of Performance: The Inefficiency in Baseball's Payrolls
By Jyotish Jotwani

Introduction

Baseball is the only sport where a team can spend $350 million and miss the postseason entirely, while an $80 million roster can march all the way to a championship run. Baseball is unique in the fact that it doesn't have a hard salary cap. This gives the teams that have the money the ability to spend top dollar on talent, but on the other side, there are teams that don't spend as much. For the teams that don't spend a lot to obtain talent, they tend to draft well, spend money responsibly, and keep their roster young. The main question is: How strong is the relationship between MLB payroll and team performance, and how do efficiency, roster construction, and pitching vs. offense explain team outcomes?

Since Major League Baseball (MLB) doesn't have a hard salary cap, teams that generate more revenue can easily outspend the smaller-market teams and acquire the big-name players. This could lead to the assumption that if you were to spend more on these big-time players, this would mean you would win more games, but it doesn't always work that way. While higher-spending teams get better players, spending millions of dollars can't guarantee a championship. General data shows that if a $50 million team goes to $100 million, adding this much money to your payroll can help out a lot, and it usually makes a big difference to the overall team. On the other hand, going from $200 million to $250 million usually only adds a few wins, which doesn't justify the major spending. There's also a high chance for those high-spending teams that they end up picking up bad contracts. These bad contracts could include overpaying for older star players who don't perform as well as they used to.

So how do the lower-spending teams compete? They compete through the strategy called Moneyball. Moneyball is a strategy that general managers use to focus on efficiency and getting the maximum dollar of wins out of each dollar that they spend. This strategy comes from the famous book called Moneyball, and it shows how there are multiple ways, even if you don't spend a lot of money, to compete and make the playoffs. The main way they do this is by targeting undervalued statistics and smart roster management.

This paper argues that while spending more money gives MLB teams a higher baseline chance of winning, a team's ultimate ceiling depends on financial efficiency, smart roster construction, and cost-controlled talent. To prove this, the study analyzes 90 team observations across three consecutive MLB seasons using correlation metrics, R^2 determinations, cost-efficiency calculations, and a breakdown of hitting versus pitching spending. The rest of this paper is organized as follows: first, a review of existing research on sports economics and market efficiency; next, a detailed explanation of the methodology and data collection; followed by the statistical results and a discussion of what these findings mean for front-office strategy; and finally, an analysis of the study's limitations along with ideas for future research.

Literature Review

Previous research on Major League Baseball has looked at whether financial resources translate to better team performance or whether teams can overcome payroll disadvantages from a more efficient roster construction. The research generally indicates that payroll is an advantage, but that spending does not fully determine success. The way teams allocate resources and spot valuable talent does have a major effect upon performance.

Wiseman and Chatterjee (2003) looked at the relationship between team payroll and performance within Major League Baseball between 1985 and 2002. The study indicated the relationship strengthened over time and more particularly in the late 1990s and early 2000s. Teams with greater payrolls tended to show better performance, but the way that payroll was distributed among players was also critically important. Each of these evolving differences in team payrolls bears direct relation to concerns regarding competitive balance within MLB. This sort of research helps to confirm both the value of financial resources as providing an advantage,but also in what ways the structure and distribution of such resources matters.

Hakes and Sauer (2006) provide a different perspective into the Moneyball hypothesis through economic evaluation. The research into productivity and compensation of players from 1999 until 2004 demonstrates how certain skills in baseball were initially undervalued relative to the labor market. Teams that could identify such inefficiencies could acquire productive players at costs less than market value. The growth of knowledge about such undervalued skills throughout baseball, however, indicates that the market found a way to correct that original inefficiency. This finding is important in relation to how lower-payroll teams could compete through information and resource efficiency rather than simply outspending competitors.

Lu et al. (2018) expand the relationship between spending and performance to separate total payroll into pitcher and hitter payroll figures. Using MLB team level data from 1985 to 2015, they establish the same inverse U-shaped model of relationship between payroll and winning percentage. Similar curvilinear relationships are seen with both pitcher and hitter salaries separately. Such findings do establish that increased payroll can lead to improved performance, the relationship, however, is clearly not unlimited or perfectly linear. The consideration of pitcher versus hitter payroll is particularly relevant to the present study, as each type of spending creates different effects upon team performance.

Anzivino (2004) brings additional background information about the relationship between payroll and performance within professional sports leagues. His work considers how payroll relates to competitive performance by placing the discussion within the context of payroll inequalities across professional sports teams. This perspective does help to indicate both the relationship of financial resources to competitive outcomes and why such differences in payroll alone fail to completely explain team success. The issues of team building and financial resource allocation remain important considerations in achieving performance.

These studies demonstrate that payroll is one of the factors that impacts the performance of professional baseball teams, but it is not the only factor that impacts their performance. Wiseman and Chatterjee found that teams with higher payrolls experienced higher levels of

performance compared to teams with lower payrolls. Furthermore, Hakes and Sauer found that teams that can find undervalued players within the league can spend their resources more efficiently to compete against other teams in the league. Additionally, Lu et al found that the impact of team spending differs according to whether the teams spend their resources on hitters or pitchers within the team. Each of these studies provides the foundation for this study in examining the impact of MLB team payrolls on team wins, as well as the efficiency of that spending and whether it correlates more with offensive or pitching team performance.

Methodology

To see how team spending impacted the performance of each team, data was collected for all 30 MLB teams across three consecutive seasons. This created a sample size of 90. Using the 90 observations provides a large enough sample size to show meaningful patterns across multiple seasons, rather than one season where there could be multiple outliers. For each season, data was collected on the following variables: financial input, overall performance, roster breakdown, and financial efficiency. By tracking both the highest-level outcomes (wins and playoff spots) and smaller metrics such as pitching vs. offense, the dataset shows us how to measure not only whether spending money helps the team, but also how different roster strategies lead to success or failure.

To analyze how team spending affects how good a ball club is in the MLB, the data was used in a four-part statistical methodology to evaluate data from 30 teams across 3 consecutive seasons.

The first statistical correlation coefficient that was used was calculated using the built-in formulas from Google Sheets. It helped measure the strength and direction of the linear relationship between outcomes like small stats, wins, offensive rank, and pitching rank. For some context, a correlation value near 1 shows that spending directly leads to a higher output, but if you get a value that's near zero, it suggests that spending has little to no effect.

To build on this, the coefficient of determination, R^2, was also used for a scatterplot trend line to measure the exact percentage of variation in regular-season wins. For example, 40% of the variation in wins is statistically associated with variation in payroll in that model, while the remaining 60% could be explained by many other different factors.

The reason why total spending varies so widely across the MLB is that the data introduced a new statistic, an efficiency metric called wins per $100 million payroll. This efficiency metric was calculated by dividing total wins by team payroll and multiplying by 100, and then multiplying by 100. This allowed for a direct efficiency comparison between the lower-budget teams and the higher-budget teams, and showed which type of spending led to a better return on investment.

When evaluating the data based on secondary variables such as pitching and offense, data was used to find the rankings from 1, which is the best, to 30, which is the worst. This was used because a smaller rank represents superior performance. A negative correlation coefficient

indicates a positive relationship, meaning higher spending leads to a better rank in the league. The data showed this led to consistent postseason teams.

Even though statistical formulas cannot capture every part of a season, the findings were paired with an outlier analysis by comparing big spenders who were underperformers against low-budget contenders. We were able to see that key factors such as young talent pipelines and smart contract structures could be implemented even though they were not directly captured by the formulas.

Results

Figure 1: correlation between wins and payroll between all three seasons.

Across the 90-team season observation, the data shows a moderate positive correlation (r=+0.41) between the total team payroll and regular-season wins. This shows that higher spending teams tend to win more games, but the amount of money that they put in alone does not guarantee the success of the team. The strength of this relationship also fluctuated year to year, ranging from a weaker correlation in 2023 (r= +0.22) to a stronger one in 2025 (r=+0.55). Based on a scatterplot analysis, the coefficient of determination, R^2, showed that payroll spending directly accounts for roughly 17-30% of the variation in regular-season wins, depending on the particular season. While breaking down the overall performance metrics, spending money was significantly better at buying offensive production than it was with pitching success. Across the total dataset, total payroll showed a moderate to strong negative correlation with the run scored rank (r= -0.52) and batting average rank (r = -0.44), meaning that the higher-spending teams consistently secured better offensive ratings. In comparison, the correlation between payroll and team pitching rank (ERA) was far weaker (r =-0.18). To analyze team efficiency, the metric wins per $100 million payroll was used to highlight the major gaps in how efficiently franchises convert money into wins. Smart-market teams operating on tight budget constraints consistently achieved superior cost efficiency by relying on young, controlled talent and smart asset

valuation. There were two outliers: The low-payroll overperformers were the 2023 Baltimore Orioles, who had 101 wins despite being a bottom-tier budget team, and the 2025 Guardians, who had 88 wins and the division title with a $94 million payroll, which ranked 29th in the MLB. Franchises like the 2023 and 2025 New York Mets, who missed the postseason despite having top-two payrolls both seasons, showed the risks of overpaying for agent talent.

Discussion

The results from the data confirm that payroll plays a significant role in MLB success. However, it is far from a perfect predictor of how well a team will do. Spending money provides a distinct advantage by allowing teams to have deeper rosters, more complete pitching, and a better bullpen. This leaves the majority of winning performance to nonfinancial factors like coaching, roster construction, player productivity, and, more importantly, how lucky the team got that particular season. This gives the high-spending franchises a better baseline probability of winning. Eventually, spending only accounts for a small percentage of overall wins; money functions as a tool rather than a guarantee of success. Factors such as coaching, team chemistry, player health, and second-half momentum also play an equally critical role in determining whether a team reaches the postseason or falls short.

A main point that's derived from the research is that financial capital is significantly more effective at buying better offense than pitching success. Offense shows a much higher connection to spending because position players generally offer predictable year-to-year statistical consistency, making hitters a safer long-term bet. Pitching, on the other hand, is a lot more volatile due to high injury rates, how much their arm is used, and random performance dips. As a result, front offices often exercise caution when committing large capital to free-agent pitchers because it does not guarantee elite ERA rankings. Successful organizations tend to buy better offensive output while relying on internal development to develop their pitching. This demonstrates that financial capital reliably secures those proven hitting talents on the free agency market, but pitching performance is much harder to buy on the market due to how often the ranks change and how much more injury-prone a pitcher is compared to a hitter. Teams like the Orioles prove that the strategy of having a strong system for locking in core talent early can beat high-spending rivals. These outliers prove that even though money raises a team's baseline floor, poor efficiency and weak pitching can cause high payrolls to collapse.

Most importantly, the population of major outliers highlights the operational efficiency of getting the highest win output per dollar spent. This is the greatest equalizer in MLB. High-payroll underperformers demonstrate that spending money on aging or declining assets creates a severe financial burden that can hinder a roster and prevent the team from going in a direction they would normally not go in. On the other hand, low-payroll overperformers prove that the small-market strategy of maximizing cost-controlled talent can consistently beat the richer rivals. Ultimately, while larger payroll raises a team's floor, if a team is more disciplined and the roster construction is great, that will be what determines the overall ceiling.

Limitations

The primary limitation of the study is that correlation measures statistical association rather than direct causation. This is why higher-spending teams are associated with more regular-season wins, but a higher payroll alone does not cause a team to win. Extreme variables such as mid-season trade deadline acquisitions, roster health, and manager quality, as well as some random luck, can show a strong influence on a team's final record that stats alone cannot isolate.

Tracking total payroll only measures overall financial output and does not show how effectively that capital is distributed across a roster. For example, a team with a $250 million payroll might spend a lot of money on two injured superstars who always sat on the bench, leaving the rest of the active roster underfunded. On the other hand, a team spending the same amount might distribute its funds evenly and construct a deep, balanced squad. Total payroll numbers often fail to capture contract quality, how deep a roster is, or where money is tied up in a productive or unproductive player, as well as when some players get injured.

Using league ranks, ranging from 1 through 30, for the different types of variables like batting average, runs scored, and pitching ERA, simplifies multidimensional performance into a single ordering. The ranking variables remove important context regarding the raw statistical gaps. For example, the gap between the first and second-best pitching teams might be massive, while the gap between the 10th- and 15th-best might be extremely small. Ranking systems treat these gaps as equal increments, which, over time, loses the overall value of the numerical granularity.

This study also relies on a three-season dataset, which has 90 total teams. Even though analyzing three consecutive years provides strong insight into the current competitive trends, MLB financial dynamics, player luxury thresholds, and market valuations shift over long periods of time. A broader multidecade or ten-season analysis would be necessary to determine whether these patterns remain consistent over time. This can also lead to a disparity between buying offense or buying pitching, which normally remains stable over a long-term economic cycle.

Conclusion

In conclusion, while the study shows that MLB payrolls do have a moderately positive correlation ($r = +0.41$) with regular-season wins delivered by clubs, a large payroll does not automatically bring success. Total spending explains only between 17% and 30% of win variations; while higher budgets certainly improve the chance of success at baseline, they don’t determine performance ceilings.

Furthermore, spending is significantly more effective at securing offensive players and performance ($r = -0.52$) than pitching performance ($r = -0.18$). This indicates that teams can more effectively secure reliable offensive contributions with their resources, while pitching performance continues to remain more difficult to effectively predict or purchase. However, spending is only one advantage; other factors such as roster construction, development, and health, along with financial efficiency, truly determine how well any team can make payroll win goals a reality.

The outliers of this study were teams with relatively low payrolls such as the Cleveland Guardians or Baltimore Orioles, and those with high payrolls such as the New York Mets they all help to reinforce the idea that efficient team construction helps even small budgets compete with very wealthy franchises. While money can indeed buy talent, it is strategy and efficiency that determine how valuable that talent becomes.

Future studies of this sort could extend well beyond the three seasons assessed to a ten-year window to view how luxury taxes and economic conditions impact long-term strategy. Future studies could also focus upon the percentage of team payroll allocated to each area of the roster and how efficiently those dollars translate to field performance. The incorporation of even more advanced statistics, such as WAR (wins-above-replacement) or pitch-level data, could reveal an even deeper understanding of why pitching performance remains so difficult to purchase directly within the open market.

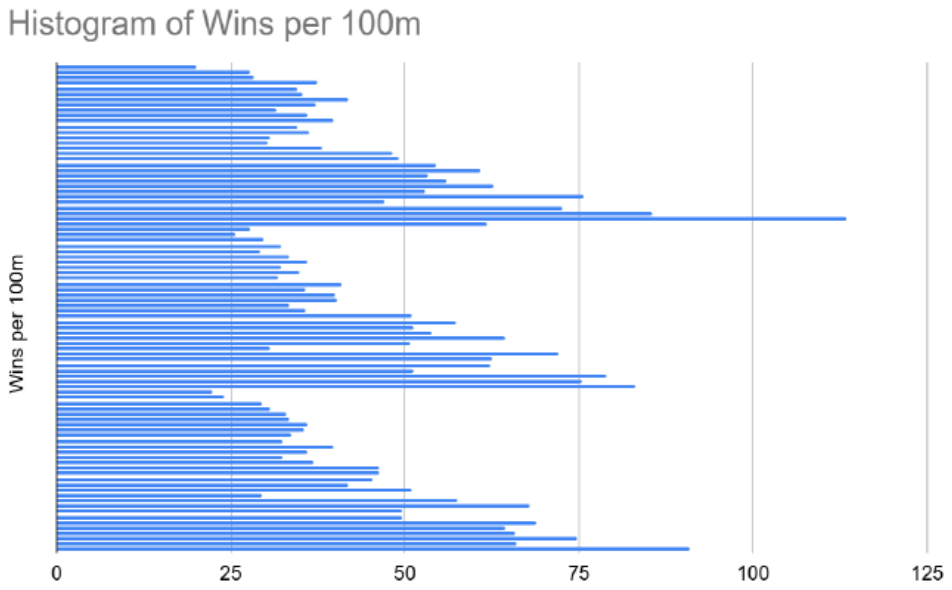


Figure 2: Histogram showing wins per 100m over a 3 year time period.

Works Cited

Spotrac. “MLB Payrolls.” Spotrac. Accessed 15 Mar. 2026. https://www.spotrac.com/mlb/payroll/_/year/2025

Spotrac. “MLB Payrolls – 2024.” Spotrac, Accessed 15 Mar 2026. https://www.spotrac.com/mlb/payroll/_/year/2024.

Spotrac. “MLB Payrolls – 2023.” Spotrac, 15 Mar 2026. https://www.spotrac.com/mlb/payroll/_/year/2023.

“2023 MLB Team Hitting Stat Leaders.” MLB.com, Accessed 25 Mar. 2026, https://www.mlb.com/stats/team/batting-average/2023.

“2024 MLB Team Hitting Stat Leaders.” MLB.com, Accessed 25 Mar. 2026, https://www.mlb.com/stats/team/batting-average/2024.

“2025 MLB Team Hitting Stat Leaders.” MLB.com, Accessed 25 Mar. 2026, https://www.mlb.com/stats/team/batting-average/2025.

“2023 MLB Team Hitting Stat Leaders.” MLB.com, Accessed 25 Mar. 2026, https://www.mlb.com/stats/team/runs/2023.

“2024 MLB Team Hitting Stat Leaders.” MLB.com, Accessed 25 Mar. 2026, https://www.mlb.com/stats/team/runs/2024.

“2025 MLB Team Hitting Stat Leaders.” MLB.com, Accessed 25 Mar. 2026, https://www.mlb.com/stats/team/runs/2025.

“2025 MLB Team Pitching Stat Leaders.” MLB.com, Accessed 31 Mar. 2026, https://www.mlb.com/stats/team/pitching/earned-runs/2025?sortState=asc.

“2024 MLB Team Pitching Stat Leaders.” MLB.com, Accessed 31 Mar. 2026, https://www.mlb.com/stats/team/pitching/earned-runs/2024?sortState=asc.

“2023 MLB Team Pitching Stat Leaders.” MLB.com, Accessed 31 Mar. 2026, https://www.mlb.com/stats/team/pitching/earned-runs/2023?sortState=asc.

Hakes, Jahn, and Raymond D. Sauer. “An Economic Evaluation of the Moneyball Hypothesis.” Journal of Economic Perspectives, vol. 20, no. 3, 2006, pp. 173–186. American Economic Association, Accessed 4 Apr. 2026, https://doi.org/10.1257/jep.20.3.173.

Lu, Xing, Jason Matthews, Miao Wang, and Hong Zhuang. “Team Payroll, Pitcher and Hitter Payrolls and Team Performance: Evidence from the U.S. Major League Baseball.” Economics and Business Letters, vol. 7, no. 2, 2018, pp. 62–69, Accessed 4 Apr. 2026, https://doi.org/10.17811/ebl.7.2.2018.62-69.

Wiseman, Frederick, and Sangit Chatterjee. “Team Payroll and Team Performance in Major League Baseball: 1985–2002.” Economics Bulletin, vol. 1, no. 2, 2003, pp. 1–10, Accessed 4 Apr. 2026, https://www.accessecon.com/pubs/EB/2003/Volume1/EB-03A10003A.pdf.

Anzivino, Ralph C. “The Correlation Between Team Payroll and Competitive Performance in Professional Sports Leagues.” Entertainment & Sports Lawyer, Summer 2004, p. 1, Marquette University Law School, Accessed 4 Apr. 2026, https://scholarship.law.marquette.edu/facpub/303/.

A Fragile Foundation: How Inequality, Financial Collapse, and Government Failure Shaped the Great Depression By Joseph Akilov

The Great Depression was one of the most critical economic crises in modern American history, causing massive unemployment, poverty, and instability across the country and world. Historians continue to debate what caused the Depression to become so severe and why it lasted for so long. Some, such as Robert McElvaine, emphasize the role of economic inequality and the rise of extreme individualism in the 1920s, arguing that these forces weakened consumer demand and concentrated wealth exclusively among the rich. Others, such as Charles Calomiris, emphasize the collapse of the financial system, citing the 1929 stock market crash, bank failures, debt deflation, and a decline in lending. Still, different interpretations point to the failure of government policy, particularly under Herbert Hoover, as a key factor in allowing the crisis to worsen. While each of these explanations captures an important aspect of the Depression, the crisis can be best understood as the result of a mix of socioeconomic inequality and financial instability, in which the unequal distribution of wealth and reliance on credit in the 1920s created a fragile economic system that collapsed under financial strain and was further intensified and prolonged by the inadequate government response.

The Great Depression is most widely seen as an economic crisis at its core. First, the Stock Market Crash of 1929 is virtually inseparable from any discussion of the Great Depression. Stock values experienced unprecedented drops between 1929 and 1932, and the ratio of high to low-grade bonds fell from 2.4 in 1929 to .3 in 1933, causing investors to become extremely afraid of risk(Calomiris 68). By October 29, 1929, stock prices had tumbled from a record high of 381 points in early September to just 198 (Metch). The lending system collapsed during this time period as well, with the percentage of bank deposits loaned out falling from 85% to 58% by 1933, which slowed business activity and constricted the money supply by nearly 25% (Calomiris 69). The financial ramifications cut even deeper still: between 1929 and 1933, industrial production in America plummeted by 47%, GDP declined by 30%, and unemployment skyrocketed upwards of 20% (Duignan). While financial conditions deteriorated, President Herbert Hoover's administration widely rejected the idea of government aid, with the eventual election of Franklin D. Roosevelt marking a shift toward greater federal involvement in economic recovery efforts. The pain caused by the Great Depression lasted for years, and most historians agree on its severity.

However, historians have varying perspectives on the forces that sparked and prolonged the Great Depression. Many view the Depression strictly as a failure of the American financial system. For example, Charles W. Calomiris focuses on the economic mechanisms behind the crash in *Financial Factors in the Great Depression.* Here, Calomiris references John Maynard Keynes, pointing out that Keynes was one of the first economists to argue that when financial markets break down and businesses stop investing, the economy goes into a downturn (Calomiris 71). He goes on to present statistics and information on stocks, bonds, and other metrics in the financial sector. Moreover, raw data from sources like the US Department of Commerce, the

National Bureau of Economic Research, and bank suspension and failure records from 1930 to 1933 further advance this argument, which provides a quantitative look at the Depression's timeline and severity (68-70). Arguments from this perspective are mainly economic, meant to shed light on how financial forces like bank failures, market crashes, and credit breakdowns affected real economic activity like investment, production, and employment. From this perspective, the Great Depression was mainly prolonged by failures with the financial system itself rather than by social inequality or political factors.

Other historians, in contrast, assert that governmental and structural factors were the main drivers of the Great Depression. Robert McElvaine's book *The Great Depression: America, 1929-1941* argues that socioeconomic inequality and President Hoover's governmental response to the Depression were the specific reasons behind the crisis's duration and severity. According to Robert McElvaine, economic growth in the 1920s mostly benefited the wealthy, while wages for most Americans stayed low. "The prosperity of the twenties rested on an increasingly narrow base," he writes in his book, highlighting how only a small group of wealthy Americans supported the economy, and how when that small group failed, the economy struggled (29). He also references the role of the federal government in allowing the Depression to escalate. President Hoover rejected the idea that the government should provide direct help to those in need, and this, McElvaine notes, made people feel abandoned and angry at the government (138). Essentially, historians who side with McElvaine argue that socioeconomic inequality and President Hoover's limited governmental response explain why the Depression lasted for so long and produced such disastrous effects, more than pure financial instability.

While these interpretations emphasize different causes of the Great Depression, these positions can be reconciled to put together a more thorough picture. The unequal distribution of wealth during the 1920s created a shaky and credit-dependent economy that became highly vulnerable to financial disruption once stock prices collapsed and banks began to fail. At the same time, the federal government's limited response under Herbert Hoover allowed these economic problems to intensify during the early years of the crisis, which helps explain the Great Depression's long duration.

Socioeconomic forces like wealth inequality nurtured a climate in which financial instability could spread rapidly under the right conditions. Robert McElvaine traces this inequality in great detail in his book. As mentioned earlier, the economic growth of the early 1920s was very top-heavy. McElvaine notes that this created a boom in credit in the non-elite classes (33). People bought cars, appliances, and goods by borrowing money, even when they couldn't actually afford them, creating the illusion of a healthy economy. Factories kept making products, but regular people didn't have enough money to buy them. This caused overproduction, business losses, and layoffs (36). As the Depression began to get worse, large portions of the population began to lack sufficient income to sustain themselves. Both individuals and businesses became increasingly reliant on borrowing to keep themselves afloat. This dependence on credit expanded rapidly during the late 1920s, including the stock market, where investors frequently purchased stocks on margin, borrowing money in the expectation that

prices would continue to rise (36). Irving Fisher - a famous economist often credited with predicting the 1929 stock market crash and author of *The Debt-Deflation Theory of Great Depressions* - observed this phenomenon during the Depression itself, arguing that "over-indebtedness may lend importance to over- investment or to over-speculation" (Fisher 341). Once stock prices began to fall in 1929, though, the reliance on borrowed money quickly turned into a liability. The "over-investment" Fisher had noted was now collapsing under its own weight. Banks began issuing less credit due to the wave of bank failures in the early 1930s, which resulted in deflation and further hindered the economy's recovery (Carlson and Rose). Debt deflation made debt unpayable, and people and businesses suddenly had to repay loans with money that was worth more than when they borrowed it, which made repayment much harder. John Maynard Keynes also shared this perspective on debt deflation. Charles Calomiris quotes one of Keynes's 1931 essays directly, as Keynes writes, "a decline in the money values so severe as that which we are now experiencing threatens the solidity of the whole financial structure" (Keynes, qtd. in Calomiris 71). Falling asset prices continued to weaken the debt relationships that sustained the banking system, and the Great Depression was further amplified and prolonged as a result. In this way, the financial weaknesses created by socioeconomic inequality -limited consumer demand and heavy reliance on credit - created a fragile financial system that was prone to collapse when conditions changed.

The Federal government's response to the early years of the Great Depression also played a significant role in allowing the crisis to deepen. Under President Hoover, the prevailing belief was that the economy would correct itself without government intervention. This attitude promoted voluntarism, which encouraged businesses and private charities to support those in need rather than relying on federal aid. However, as the economic situation worsened, this approach proved to be insufficient. Millions of Americans lost their jobs, businesses collapsed, and state and local governments quickly ran out of resources to provide assistance. Even when millions were unemployed and starving - with unemployment above 20% in the early 1930s - Hoover did not change his beliefs (Duignan). According to McElvaine, "Hoover's faith in voluntarism remained unshaken even as the economy collapsed" (McElvaine 58). This ideology was rooted in Hoover's fear that government help would weaken individual responsibility and lower the American work ethic. Despite his insistence on voluntarism, cities, states, and charities ran out of money and could not help everyone in need (McElvaine 89). Even the famed economists of the Great Depression era began recognizing the failures of this approach, with many supporting the idea of governmental assistance during times of crisis. Calomiris's study consistently cites Keynes as an advocate for federal aid, and Irving Fisher explicitly states that it would be "silly" and "immoral" to "let nature run unchecked" during the Depression, comparing the lack of aid to a "physician neglecting a case of pneumonia" (Fisher 347). By any measure, the crisis had become too big for private help alone. In a system that had been weakened by inequality and over-reliance on credit, Hoover's response worsened the effects of the financial collapse.

In contrast to Hoover's limited response, Franklin D Roosevelt's administration marked a huge shift in both economic policy and government role in managing the crisis. Upon taking office in 1933, Roosevelt moved quickly to rebuild confidence in the financial system and address the breakdown of credit that had defined the early years of the Depression. One of his first major actions - the national bank holiday - temporarily halted bank runs and allowed the federal government to reorganize and stabilize the banking system (Roosevelt). This was followed by a series of reforms, including the creation of the Federal Deposit Insurance Corporation (FDIC), which guaranteed bank deposits and helped rebuild public trust in financial institutions. FDR also rejected strict laissez-faire ideas - the idea that businesses should be allowed to do what they want without government intervention - and supported government action. As McElvaine argues, Roosevelt understood that "confidence itself was an economic force," and his policies were designed not only to address structural problems but also to restore public faith in the financial system (McElvaine 118). Roosevelt resisted the voluntarist viewpoints of Hoover, and his approach to government action would pave the road for the recovery that would follow.

Although Roosevelt understood the measures necessary to remedy the crisis, his approach was further transformed by popular will. The FDR administration's New Deal consisted of a set of federal policies and programs that were designed to both repair the economic disaster caused by the Depression and to prevent future disasters (McElvaine 164). Although these policies were more impactful than Hoover's, many Americans thought that the government needed to do more. As McElvaine notes, the impetus to change "came not only from Washington, but from below" (231). In other words, it was not a politician's vision but public pressure that impelled FDR to shift towards a more-aggressive version of the New Deal in 1935. This second New Deal promoted long-term economic security through programs like Social Security and labor protections (255). Ultimately, the public's will nudged the US federal government to make "a decisive break with the nation's laissez-faire past" (McElvaine 263) and embrace the notion that its job included ensuring that Americans enjoy basic economic security.

Ultimately, the Depression spread beyond America's borders. A strictly national crisis had suddenly evolved into a global economic disaster. As credit markets tightened and American banks failed, international lending declined sharply, particularly to European economies that had depended heavily on U.S loans during the 1920s. The economic ramifications were massive and long-lasting. For reference, by the Depression's peak in 1932, production in Germany had dropped to only 53% of its 1929 levels, while world global unemployment rates reached nearly 30% ("Money"). Countries such as Germany and Austria, already burdened by war debts and reparations from the First World War, were especially vulnerable to this sudden withdrawal of capital. As financial institutions across Europe began to fail, economic contraction spread rapidly, leading to declines in industrial production, falling trade, and rising unemployment on a global scale. These developments demonstrate that the financial collapse of the early 1930s had escalated into a worldwide crisis. A national event that had been born out of economic inequality

and financial disaster and prolonged by inaction from its own government had triggered a global chain reaction across interconnected economies.

The Great Depression can be best understood as a mix of structural, financial, and political forces. While financial collapse played a central role in triggering and deepening the downturn, it was the underlying inequality of the 1920s that made the economy so dependent on credit and vulnerable to failure, and the federal government's limited response under Herbert Hoover ultimately allowed the crisis to intensify. Eventually, Franklin D. Roosevelt's more interventionist approach helped restore confidence and stabilize the financial system, but the crisis was to cut much deeper and last much longer due to a combination of financial, socioeconomic, and governmental factors behind it. These factors reveal that the severity and duration of the Great Depression were shaped largely by economic conditions, but also by the ideas and policies that governed them. The story of the Great Depression remains relevant to the recessions of recent years, as knowing the forces that sparked and prolonged the crisis in the early 20th Century can help us avoid making the same mistakes in modern times.

Works Cited

Calomiris, Charles W. “Financial Factors in the Great Depression.” Journal of Economic Perspectives, vol. 7, no. 2, 1993, pp. 61–85. JSTOR, http://www.jstor.org/stable/2138200.

Carlson, Mark, and Jonathan Rose. “Credit Availability and the Collapse of the Banking Sector in the 1930s.” FRB: Finance and Economics Discussion Series: Screen Reader Version - Credit Availability and the Collapse of the Banking Sector in the 1930s, 26 Aug. 2011, www.federalreserve.gov/pubs/feds/2011/201138/index.html.

Duignan, Brian. "Causes of the Great Depression". Encyclopedia Britannica, 13 Jun. 2025, https://www.britannica.com/story/causes-of-the-great-depression. Accessed 26 August 2026.

Fisher, Irving. “The Debt-Deflation Theory of Great Depressions.” Econometrica, vol. 1, no. 4, 1933, pp. 337–357. Reprinted in FRASER, Federal Reserve Bank of St. Louis, https://fraser.stlouisfed.org/title/debt-deflation-theory-great-depressions-3596.

McElvaine, Robert S. The Great Depression: America, 1929–1941. Three Rivers Press, 1993.

Metych, Michele. "Black Tuesday and the Crash of 1929". Encyclopedia Britannica, 10 Feb. 2026, https://www.britannica.com/topic/Black-Tuesday-1929. Accessed 26 August 2026.

“Money Matters, an IMF Exhibit -- the Importance of Global Cooperation, Conflict and Cooperation (1871-1944), Part 4 of 6.” International Monetary Fund, www.imf.org /external/np/exr/center/mm/eng/mm_cc_03.htm. Accessed 26 Aug. 2026.

Roosevelt, Franklin D. “Fireside Chat on Banking.” The American Presidency Project, 12 Mar. 1933, https://www.presidency.ucsb.edu/documents/fireside-chat-banking.

Impact of the Russia–Ukraine War on Food Inflation in West Asian Countries: The Roles of Wheat-Import Concentration and Self-Sufficiency

By Shayan Shamsi Pour Siahmazgi and Ariel Goldman

Abstract

This paper examines the relationship between wheat-market characteristics and food inflation in ten West Asian countries following the Russia–Ukraine war. We compare countries' wheat-import concentration, measured using the Herfindahl–Hirschman Index (HHI), and wheat self-sufficiency ratios with their average quarterly food-inflation spikes. We find Pearson correlation coefficients of -0.311 between HHI and the average quarterly food-inflation spike and 0.122 between wheat self-sufficiency and the inflation spike. After excluding Turkey as an outlier, however, the coefficients change to -0.413 and -0.846, respectively. These findings suggest that wheat self-sufficiency may be associated with greater resilience to inflationary pressures from foreign supply-chain disruptions; however, this relationship appeared only after Turkey was excluded and should therefore be interpreted cautiously.

Introduction

The full-scale invasion of Ukraine by Russia began on 24 February 2022, throwing many world markets into turmoil. The Russia–Ukraine war was a global supply-chain shock that disrupted agricultural production, logistics, storage, processing, and trade (Jagtap et al.). Together, Russia and Ukraine account for 30% of global wheat exports (Jagtap et al. 18), with many countries heavily relying on one or both for wheat supplies. We therefore use 24 February 2022 as the dividing point between the pre-shock and post-shock periods.

To measure wheat-import concentration, we calculated the Herfindahl–Hirschman Index (HHI). HHI is a measure of a market's concentration and dependence on its suppliers. For example, an HHI of 1 indicates that all imports come from one supplier, whereas an HHI of 0.500 indicates less supplier concentration (Zhang et al.). Countries differ in both import volume and supplier concentration. HHI allows us to measure how concentrated each country's wheat imports are among its suppliers. Previous research by Zhang et al. used HHI, import quantity, and Russia/Ukraine import share to assess countries' vulnerability to supply disruptions caused by the war. Zhang et al. explain how a higher HHI, higher import share, and larger import quantity imply greater vulnerability to supply disruption. However, Zhang et al. do not examine the correlation between HHI score and food inflation, which is the focus of this paper.

Additionally, we use quarterly food inflation to see changes in food prices throughout the pre- and post-shock periods. A 2022 analysis of food inflation by Kohlscheen shows that domestic food production, imports, exports, and macroeconomic conditions all matter for food-price outcomes. In order to account for this variability, we included wheat self-sufficiency as a second explanatory variable.

In this paper, we use wheat-import concentration (HHI), wheat self-sufficiency ratios, and food-inflation data to examine how ten West Asian countries responded to the Russia–Ukraine

shock. In the full sample, wheat-import concentration had a weak negative correlation with food-inflation spikes ($r = -0.311$), whereas wheat self-sufficiency had a weak positive correlation ($r = 0.122$). After Turkey was excluded, the correlations were $r = -0.413$ for wheat-import concentration and $r = -0.846$ for wheat self-sufficiency.

Methodology

Russia and Ukraine are two of the largest wheat-supplying nations across the globe (Jagtap et al.). Russia's full-scale invasion of Ukraine began on 24 February 2022 and serves as the dividing point between the pre-shock and post-shock periods in this study. The analysis uses three variables: the quarterly food-inflation rate, the Herfindahl–Hirschman Index (HHI), and the wheat self-sufficiency ratio for ten West Asian countries. This study defines West Asia using the United Nations Statistics Division's M49 geographic classification of 18 countries: Armenia, Azerbaijan, Bahrain, Cyprus, Georgia, Iraq, Israel, Jordan, Kuwait, Lebanon, Oman, Qatar, Saudi Arabia, State of Palestine, Syrian Arab Republic, Turkiye, United Arab Emirates, and Yemen.

The sample consists of Armenia, Azerbaijan, Bahrain, Cyprus, Georgia, Israel, Kuwait, Qatar, Saudi Arabia, and Turkey. The remaining eight countries were excluded because usable wheat-import data, food CPI data, or both were unavailable.

We used quarterly food CPI data to increase the precision of the analysis. This approach allowed us to exclude the first quarter of 2022, the quarter in which the full-scale invasion began.

The following data were collected for each of the 10 countries mentioned above:

Table 1
Variable Definitions

Data name	Data description
Pre-Shock CPI Average (2017–2021)	The average of the quarterly food inflation CPI from 2017 through 2021. Calculated from the food inflation index.
Post-Shock CPI Average Excluding the First Quarter of 2022 (2022–2023)	The average of the quarterly food inflation CPI from the second quarter of 2022 through the fourth quarter of 2023, excluding the first quarter of 2022 (the shock period). Calculated from the food inflation index.
Food Inflation Spike	Post-Shock CPI Average – Pre-Shock CPI Average
Wheat Self-Sufficiency Ratio (2021, %)	The percentage of domestic wheat use supplied by domestic wheat production.
HHI Score (2021)	We squared each supplier country's percentage share of wheat imports, summed the squared shares, and divided the result by 10,000 to normalize the HHI to a scale from 0 to 1: $HHI = \frac{\sum_{i=1}^{n} S_i^2}{10000}$ where S_i is supplier i's market share expressed as a whole-number percentage.

From the UN Comtrade Database, we collected 2021 annual wheat-import data under HS code 1001 for each of the ten selected countries. We used these data to calculate each country's HHI.

Using Ha et al.'s World Bank database, we collected the quarterly food CPI index for each country and calculated each quarter-to-quarter food-inflation rate using the equation $(\frac{Q_n}{Q_{n-1}} - 1)\times 100$, where Q_n is the food CPI index in the current quarter and Q_{n-1} is the index in the previous quarter. We then calculated the average quarterly food-inflation rate for 2017–2021 and for 2022–2023, excluding the first quarter of 2022 because the invasion began during that quarter. Finally, we subtracted the pre-shock average from the post-shock average to calculate the post-shock food-inflation spike.

Table 2
Food Inflation Data

Country	**Pre-Shock CPI Average (2017-2021)**	**Post-Shock CPI Average Excluding first quarter of 2022 (2022-2023)**	**Food Inflation Spike**
Armenia	1.18	0.685	-0.495
Azerbaijan	1.729	-4.472	-6.201
Bahrain	0.708	1.548	0.84
Cyprus	0.109	1.649	1.54
Georgia	2.167	1.127	-1.04
Israel	0.27	1.217	0.947
Kuwait	0.876	1.532	0.656
Qatar	0.405	0.709	0.304
Saudi Arabia	1.179	0.575	-0.604
Turkey	4.775	13.892	9.117

Source: Calculations based on quarterly food CPI data from Ha et al.

Additionally, country-level wheat self-sufficiency ratios were obtained from FAOSTAT for Armenia, Azerbaijan, Bahrain, Cyprus, Georgia, Israel, Kuwait, Qatar, Saudi Arabia, and Turkiye. We selected "Self-sufficiency ratio" as the indicator, "Wheat" as the item, and 2021 as the year ("Crops and Livestock").

Table 3
Wheat Self-Sufficiency Ratios in the Ten-Country Sample

Country	**Wheat Self-Sufficiency (%)**
Armenia	26%
Azerbaijan	62%
Bahrain	0%
Cyprus	26%
Georgia	27%

Israel	7%
Kuwait	0%
Qatar	0%
Saudi Arabia	21%
Turkey	67%

Source: Food and Agriculture Organization of the United Nations

We examined the correlation between the HHI score and food inflation spikes, as well as wheat self-sufficiency ratio and food inflation spikes. We compared the correlations between the food-inflation spike and each explanatory variable. The corresponding R^2 values indicate the proportion of variation associated with each bivariate linear model.

Results

The Association between Wheat-Import Concentration and Quarterly Food Inflation

We measured the strength and direction of the relationship between the wheat HHI score and the average quarterly food-inflation spike using Pearson correlation coefficients. We compared the correlations between the food-inflation spike and each explanatory variable. The corresponding R^2 values indicate the proportion of variation associated with each bivariate linear model. The resulting correlation was -0.311. Similarly, the regression line had a negative slope and an R^2 of 0.097, indicating a weaker linear association.

Table 4
HHI Score

Country	**HHI Score (2021)**
Armenia	0.9843
Azerbaijan	0.9105
Bahrain	0.9843
Cyprus	0.1693
Georgia	0.8523
Israel	0.3476
Kuwait	0.7303
Qatar	0.4008
Saudi Arabia	0.1832
Turkey	0.5291

Source: Calculations using UN Comtrade Database import data.

It is worth mentioning that this correlation is between wheat import concentration and food inflation spikes or changes in prices, and not the food prices themselves. This follows previous findings. Metin Çakır et al. examine the relationship between HHI score and food inflation in the United States. The study finds that despite a positive association between market concentration and food inflation, "when comparing the preinflation period with the inflationary phase (June 2021–2022), the association turns negative and significant, indicating that prices in more concentrated markets increased more slowly" (1). Although Çakır et al. examine retail-market concentration rather than import-source concentration, our negative result is consistent with the pattern they report. In our sample, higher wheat-import concentration was associated with smaller food-inflation spikes.

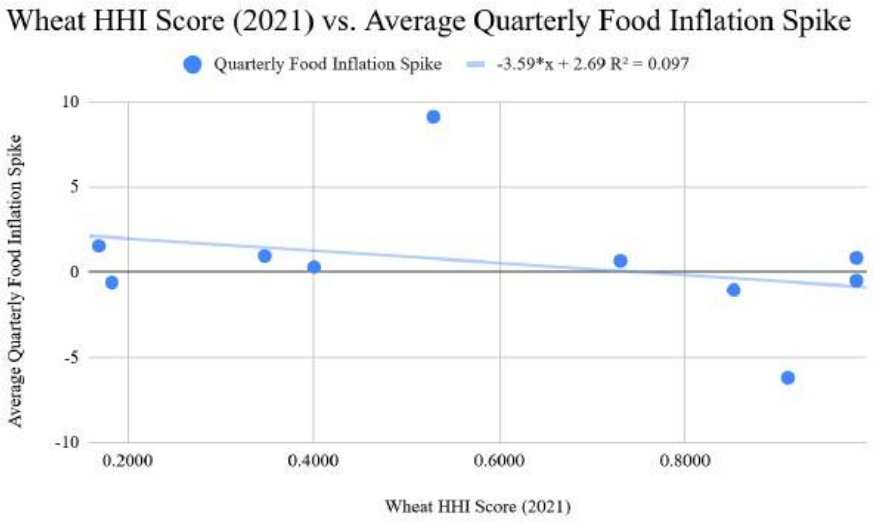


Fig. 1. Wheat import concentration had a weak negative association with food-inflation spikes, with HHI explaining only 9.7% of the variation in the sample. Source: Authors' calculations based on wheat-import data from the UN Comtrade Database and food CPI data from Ha et al.

Turkey, despite having the highest self-sufficiency ratio in our sample (see Table 3), experienced the highest average quarterly food inflation CPI spike (see Table 2). Turkey experienced a significant drought during the 2021/22 growing season. Consequently, domestic wheat production fell by 2 MMT (million metric tons) to 16.25 MMT ("Dry Conditions Continue Affecting Global Wheat Production"). The drought therefore represents a potential confounding factor. Thus, as a robustness check, we ran the correlation again without Turkey. The result was a stronger correlation of -0.413.

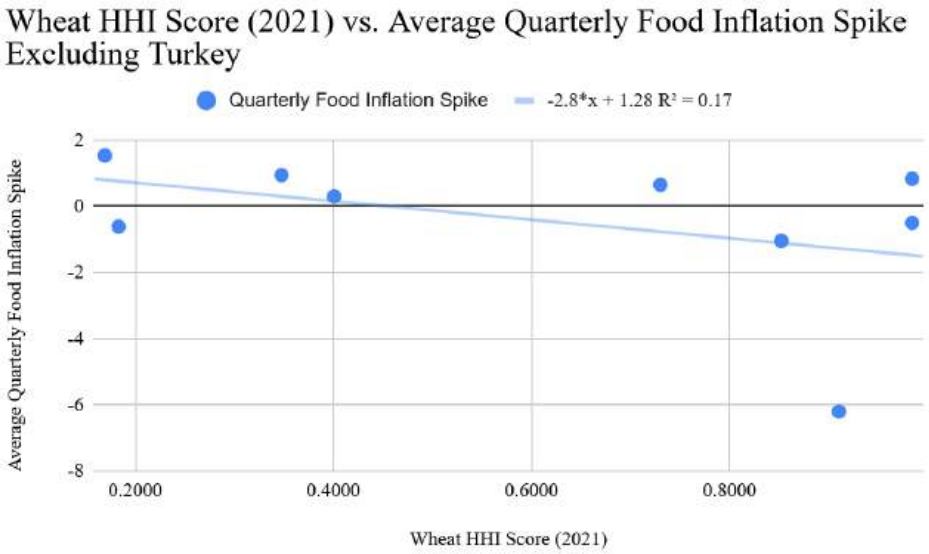


Fig. 2. Wheat import concentration had a somewhat stronger, though still weak, negative association with food-inflation spikes, with HHI explaining only 17% of the variation in the sample. Source: Authors' calculations based on wheat-import data from the UN Comtrade Database and food CPI data from Ha et al.

The Association between Wheat Self-Sufficiency and Quarterly Food Inflation

We also analyzed the relationship between wheat self-sufficiency and the food-inflation spike for the ten countries using Pearson's correlation. The resulting correlation was weak and positive (r = 0.122), contrary to our hypothesis.

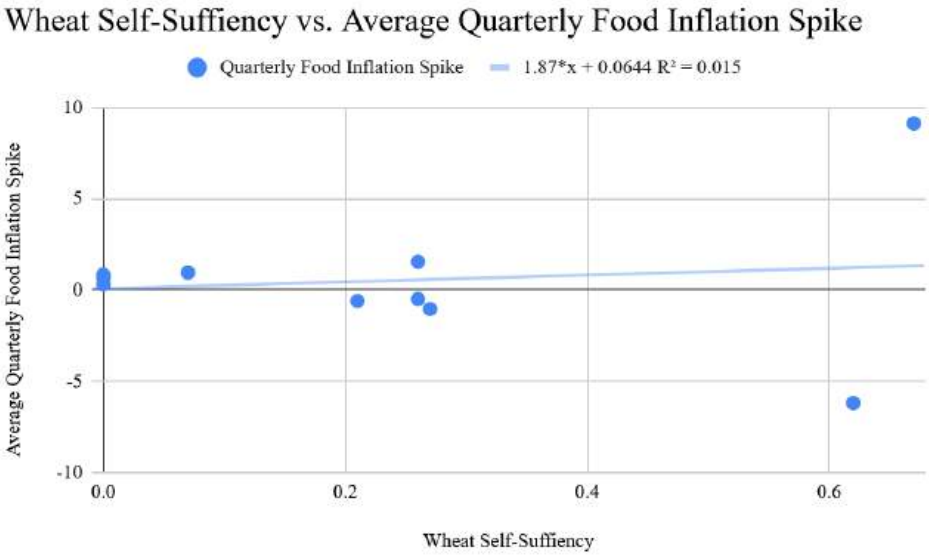


Fig. 3. Wheat self-sufficiency ratio had a weak positive association with food-inflation spikes, with SSR explaining 1.5% of the variation in the sample. Source: Authors' calculations based on FAOSTAT wheat self-sufficiency data and food CPI data from Ha et al.

The weak positive correlation in the full sample is inconsistent with the hypothesis that domestic production reduces exposure to external shocks. Generally, "reducing the foreign dependency on wheat could significantly abate the conveyance of international market shocks to indigenous markets" (Luo and Tanaka 9). As explained earlier, Turkey has the highest quarterly food inflation CPI spike (see Table 2) despite being the most self-sufficient country in our sample in wheat consumption. To assess the sensitivity of this result to Turkey, we recalculated the correlation without Turkey. The resulting correlation was strong and negative (r = -0.846).

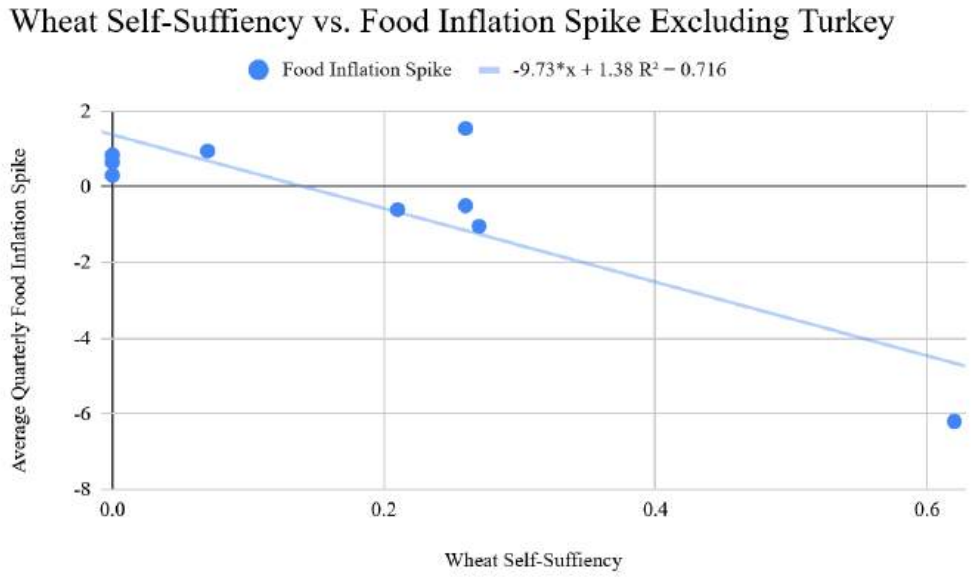


Fig. 4. Wheat self-sufficiency ratio now has a much stronger negative association with average quarterly food inflation spike, with SSR explaining 71.6% of the variation in the sample. Source: Authors' calculations based on FAOSTAT wheat self-sufficiency data and food CPI data from Ha et al.

Without Turkey, 71.6% of the variation in the calculated food-inflation spikes can be explained by the wheat self-sufficiency ratio. In the nine-country sensitivity analysis, wheat self-sufficiency showed a stronger correlation with food-inflation spikes than HHI did.

Limitations and Conclusion

Missing data limited the sample to ten of the 18 West Asian countries in the M49 classification. The results suggest that food security may be relevant to countries' resilience to geopolitical supply-chain disruptions, but they do not establish causation. After Turkey was excluded, wheat self-sufficiency showed a stronger association with food-inflation spikes than HHI; however, this pattern was absent in the full sample and should be interpreted cautiously. The small sample and the sensitivity of the results to Turkey limit the generalizability of the findings. Future research could examine larger samples, additional confounding variables, and whether exporters' dependence on major destination markets affects the prices charged to importing countries.

Works Cited

Çakır, Metin, et al. Concentration in Food Retailing, Prices, and Greedflation, 31 Dec. 2025, https://doi.org/10.2139/ssrn.5995474.

"Crops and Livestock Products Indicators." FAOSTAT, Food and Agriculture Organization of the United Nations, https://www.fao.org/faostat/en/#data/QCL. Accessed 29 July 2026.

"Dry Conditions Continue Affecting Global Wheat Production." U.S. Wheat Associates, https://uswheat.org/trade-activity/dry-conditions-continue-affecting-global-wheat-production/. Accessed 15 June 2026.

Ha, Jongrim, et al. "A Cross-Country Database of Inflation." Version April 2025, World Bank, Apr. 2025, https://thedocs.worldbank.org/en/doc/1ad246272dbbc437c74323719506aa0c-0350012021/original/Inflation-data.xlsx. Microsoft Excel spreadsheet.

Jagtap, Sandeep, et al. "The Russia-Ukraine Conflict: Its Implications for the Global Food Supply Chains." Foods, vol. 11, no. 14, July 2022, article 2098, https://doi.org/10.3390/foods11142098.

Kohlscheen, Emanuel. "Understanding the Food Component of Inflation." BIS Working Papers, no. 1056, Bank for International Settlements, 12 Dec. 2022, www.bis.org/publ/work1056.htm. Accessed 19 June 2026.

Luo, Pengfei, and Tetsuji Tanaka. "Food Import Dependency and National Food Security: A Price Transmission Analysis for the Wheat Sector." Foods, vol. 10, no. 8, July 2021, article 1715, https://doi.org/10.3390/foods10081715.

United Nations Statistics Division. Standard Country or Area Codes for Statistical Use (M49). M49, United Nations, https://unstats.un.org/unsd/methodology/m49. Accessed 16 July 2026.

United Nations Statistics Division. UN Comtrade Database. United Nations, https://comtradeplus.un.org/TradeBalance. Accessed 17 July 2026.

Zhang, Zhengyang, et al. "Countries' Vulnerability to Food Supply Disruptions Caused by the Russia–Ukraine War from a Trade Dependency Perspective." Scientific Reports, vol. 13, no. 1, Oct. 2023, https://doi.org/10.1038/s41598-023-43883-4. Accessed 19 June 2026.

Median Income Regression Model By Siddhant Jha

Abstract

Many people face uncertainty when it comes to choosing a career as though they may know the salary of a specific job they are looking at, it is not so easy to predict the income in a general field. To predict the median income of a person based on their profession, education, and time worked, we use a linear regression model trained with a dataset of 858 examples. With this, we find an equation that can be used to predict the median income. This can be used to make educated decisions about education and career choice.

Introduction

Machine learning is a field of computer science that is trained with data to make decisions or predictions (Bergmann). The model we discuss utilizes supervised learning, a class of ML that is trained off of data with labels, rather than unlabeled data, in order to be able to make predictions when given new data (Belcic and Stryker). Our model will use linear regression to predict someone's income based on factors such as their education, their years since graduation, and their industry.

Other projects like this have been done before. For example, in 2025, Pratiti Soumya published a paper explaining how she used linear regression to predict developer salaries (Soumya). This, however, focuses only on a single profession and cannot make accurate predictions for other occupations. Another study conducted by Michael Xu in 2025 has the same limitations (Xu). Though it takes into factors like gender and geographic location, it does not account for all professions. These studies, though certainly useful for people in these fields, cannot benefit those working in different industries, or those not yet in the workforce.

Data preprocessing and split

The topic we consider here is the median high school graduate income by industry. The dataset in question has 858 people, which we refer to as examples. These are the columns in the dataset: the level of education (High School, Apprenticeship, Associate Certificate, or a Bachelor's or higher), years after graduation, median hours worked, number of records, NAICSCode, NAICSTTL, and median earnings (our label, or what we are trying to predict). These columns are shown in Figure 1 below.

HighestAchievement	NAICSCode	NAICSTTL	YearAfterGrad	MedianEarnings	MedianHours	NumRecords
Associate/Certificate	11	Agriculture, Forestry,	1	11800	766	88
Associate/Certificate	11	Agriculture, Forestry,	2	14600	839	171
Associate/Certificate	11	Agriculture, Forestry,	3	23100	1282	333
Associate/Certificate	11	Agriculture, Forestry,	4	28800	1482	517
Associate/Certificate	11	Agriculture, Forestry,	5	32600	1690.5	642

Figure 1: The dataset has 7 columns some of which require one-hot encoding

The dataset is unclear about what the records in NumRecords were, so we remove this column in order to avoid using unnecessary data. In addition, the North American Industry Classification System (NAICS) is represented in both NAICSCode and NAICSTTL. Since the numerical values in NAICSCode are not ordered in any meaningful way, we remove this column. As the model cannot be trained on the text in NAICSTTL, we practice one-hot encoding. This means that we replace it with a new column for each of the values originally present ("One Hot Encoding in Machine Learning"). Each example gets a 1 in the category it belongs to and a 0 in the others. We use one-hot encoding for the highest achievement column as well.

The dataset is randomly split into two subsets. One has 75% of the examples and is the training set. The other, with 25% of the examples, is the testing set. We need a large set of data to train the model, but a sizable amount of data is also needed to test the model. Too little data in the training set leads to underfitting, while too few in the testing set leads to a potentially misleading value of error.

Linear Regression

After one-hot encoding and removing the MedianEarnings column (as those are the labels, meaning what we are trying to predict), we have 26 columns, which we denote with x_{i1}, x_{i2}, …, x_{i26} for the i^{th} example.

```
YearAfterGrad
MedianHours
Associate_Certificate
Bachelor_Higher
HS_Diploma
Apprentice
NAICSTTL_Accommodation and Food Services
NAICSTTL_Administrative and Support and Waste Management and Remediation Services
NAICSTTL_Agriculture, Forestry, Fishing and Hunting
NAICSTTL_Arts, Entertainment, and Recreation
NAICSTTL_Construction
NAICSTTL_Educational Services
NAICSTTL_Finance and Insurance
NAICSTTL_Health Care and Social Assistance
NAICSTTL_Information
NAICSTTL_Management of Companies and Enterprises
NAICSTTL_Manufacturing
NAICSTTL_Mining, Quarrying, and Oil and Gas Extraction
NAICSTTL_Other Services (except Public Administration)
NAICSTTL_Professional, Scientific, and Technical Services
NAICSTTL_Public Administration
NAICSTTL_Real Estate and Rental and Leasing
NAICSTTL_Retail Trade
NAICSTTL_Transportation and Warehousing
NAICSTTL_Utilities
NAICSTTL_Wholesale Trade
```

We refer to the labels as y, so the label of the i^{th} example is y_i. The labels we are trying to predict are denoted as $\hat{y}$. In our linear regression equation, each coefficient is referred to as w. So, our full equation will be in the form $\hat{y} = w_1x_1 + w_2x_2 + \ldots + w_{26}x_{26} + b$, where b is a constant. This model uses the adam optimizer and a learning rate of 1.0, because lower values would result in a very high error despite using tens of thousands of epochs (where an epoch is one complete look at the dataset) ("Epoch in Machine Learning"). With a learning rate of 1.0, 2950 epochs resulted in the lowest root mean squared error.

Calculating Error

To measure the error with our model, we use the root mean squared error (RMSE). The RMSE is calculated by taking the square root of the arithmetic mean of the squares of the differences between the predictions and real labels. Assuming n is the number of examples,

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}-y)^2}{n}}$$

We use the RMSE because it amplifies larger errors, whereas with the mean absolute error, it would give the average error, placing less of an impact on large errors (*Root Mean Square Error - an Overview | ScienceDirect Topics*). The purpose of this model is to find the parameters that will minimize the error.

Result of Linear Regression

The RMSE was about $7944. The mean absolute error was $5384. The final coefficients are as shown in Table 1.

Table 1

Variable	Coefficient
YearAfterGrad (x_1)	10733.1895
MedianHours (x_2)	4784.9829
Associate Certificate (x_3)	-2831.1191
Bachelor_Higher (x_4)	4581.0933
HS_Diploma (x_5)	-4886.3369
Apprentice (x_6)	4980.1968
NAICSTTL_Accomodation and Food Services (x_7)	-3590.9382
NAICSTTL_Administrative and Support and Waste Management and Remediation Services (x_8)	-1740.8966
NAICSTTL_Agriculture, Forestry, Fishing and Hunting (x_9)	-3001.3340
NAICSTTL_Arts, Entertainment, and Recreation (x_{10})	-3413.3133
NAICSTTL_Construction (x_{11})	1457.8209
NAICSTTL_Educational Services (x_{12})	-1575.4044

Table 1	
NAICSTTL_Finance and Insurance (x_{13})	-292.5255
NAICSTTL_Health Care and Social Assistance (x_{14})	-1600.7770
NAICSTTL_Information (x_{15})	3237.6101
NAICSTTL_Management of Companies and Enterprises (x_{16})	1649.6354
NAICSTTL_Manufacturing (x_{17})	911.7437
NAICSTTL_Mining, Quarrying, and Oil and Gas Extraction (x_{18})	2437.8179
NAICSTTL_Other Services (except Public Administration) (x_{19})	-2143.1526
NAICSTTL_Professional, Scientific, and Technical Services (x_{20})	373.1702
NAICSTTL_Public Administration (x_{21})	470.3937
NAICSTTL_Real Estate and Rental and Leasing (x_{22})	-1091.5970
NAICSTTL_Retail Trade (x_{23})	-2552.6025
NAICSTTL_Transportation and Warehousing (x_{24})	-676.9676
NAICSTTL_Utilities (x_{25})	4824.5674
NAICSTTL_Wholesale Trade (x_{26})	140.8500
Intercept	51564.94

So, our final equation, using the coefficients from table 1, is:

$$\hat{y} = 10733.1895x_1 + 4784.9829x_2 - 2831.1191x_3 + 4581.0933x_4 - 4886.3369x_5 + 4980.1968x_6 - 3590.9382x_7 - 1740.8966x_8 - 3001.3340x_9 - 3414.4133x_{10} + 1457.8209x_{11} - 1575.4044x_{12} - 292.5255x_{13} - 1600.7770x_{14} + 3237.6101x_{15} + 1649.6354x_{16} + 911.7437x_{17} + 2437.8179x_{18} - 2143.1526x_{19} + 373.1702x_{20} + 470.3937x_{21} - 1091.5970x_{22} - 2552.6025x_{23} - 676.9676x_{24} + 4824.5674x_{25} + 140.8500_{26} + 51564.94$$

This is a larger equation that one may expect, but keep in mind that only one of the NAICSTTL features will be 1 and the rest will be 0, so most coefficients are not used for calculations. The same goes for the level of education. So, only five values will be added, including the intercept.

Looking at the coefficients, people who work in utilities will make more money, on average, than other professions. Likewise, people who did an apprenticeship will make more money than other levels of education, on average.

Discussion

This model is trained on many factors like industry and education. However, it would benefit from more factors. Taking gender and location into account would lead to better results that would be more useful to people. With a larger dataset that has more values and factors, the model could produce a far more accurate equation.

Conclusion

In this article, we demonstrated an AI linear regression model using supervised learning that minimizes the root mean square error. We showed how using this model, we can create an equation to predict the income of someone based on factors such as their profession, their education, and the number of years they have worked. While it does have some error, the model is accurate enough that it still provides a reasonable prediction. It can be used to estimate one's future income, whether to plan for retirement, to make a budget, or to help with career selection.

Works Cited

Belcic, Ivan, and Cole Stryker. What Is Supervised Learning? | IBM. 12 Sept. 2025, Accessed July 29, 2026. https://www.ibm.com/think/topics/supervised-learning.

Bergmann, Dave. What Is Machine Learning? | IBM. 18 Aug. 2025, Accessed July 29, 2026. https://www.ibm.com/think/topics/machine-learning.

"Epoch in Machine Learning." GeeksforGeeks, 00:55:17+00:00, Accessed July 29, 2026. https://www.geeksforgeeks.org/machine-learning/epoch-in-machine-learning/.

"One Hot Encoding in Machine Learning." GeeksforGeeks, 11:43:45+00:00, Accessed July 29, 2026. https://www.geeksforgeeks.org/machine-learning/ml-one-hot-encoding/.

Root Mean Square Error - an Overview | ScienceDirect Topics. Accessed July 29, 2026. https://www.sciencedirect.com/topics/engineering/root-mean-square-error.

Soumya, Pratiti. "Predicting Developer Salaries with Machine Learning." Medium, 12 June 2025, Accessed July 29, 2026. https://medium.com/@pratitisoumya/predicting-developer-salaries-with-machine-learning-f2d81579af86.

Xu, Michael. "Salary Prediction Using Machine Learning." Scholarly Review Journal, Summer 2025, no. 13, June 2025. https://doi.org/10.70121/001c.139043.

To What Extent Does Participation in FBLA Correlate with Financial Literacy among High School Students? By Abhyuday Shashi

Abstract

Understanding personal finance is more important than ever, but many high schools struggle to offer formal financial education. This study looks at whether participating in Future Business Leaders of America (FBLA), a popular student-led competitive club helps students become more financially literate. I surveyed high school students (N = 19 analyzed) at Parsippany Hills High School using a validated financial literacy test. Then, I ran a Pearson correlation analysis to see if FBLA experience matched up with higher test scores. The results showed almost no correlation between how long someone was in FBLA and their financial literacy ($r = 0.095$, $p = 0.698$). This suggests that just being in FBLA might not be enough to teach core financial concepts, probably because the club focuses on preparing for competitive exams rather than long-term learning. This paper discusses how we can design better financial education programs based on these findings.

Keywords

Economics, Financial literacy, Competition-based learning, Secondary education, Student motivation

Introduction

These days, knowing how to manage money is a necessary life skill. Young people are entering an economy where they have to make complicated decisions about student loans, credit cards, investing, and eventually retirement (OECD, 2005). But even though everyone agrees financial skills are important, recent international assessments confirm that teenagers are still struggling to grasp the basics (Organisation for Economic Co-operation and Development, 2024). But even though everyone agrees financial skills are important, research has shown for years that teenagers are struggling to grasp the basics, a finding confirmed by recent international assessments (McCormick, 2008; Organisation for Economic Co-operation and Development, 2024). For example, the Jump$tart Coalition found that high school seniors scored an average of only 48.3% on a basic financial literacy test, which was actually a drop from previous years (Mandell, 2008). This shows we really need to find better ways to teach personal finance to teenagers.

"The big question is: what is the best way to teach these skills within the school system (Organisation for Economic Co-operation and Development, 2014)? Structured classroom learning usually works better than just learning from parents (Moscarola & Kalwij, 2021). However, even mandatory classes have mixed results. One study by Mandell and Klein (2009) found that kids who took a required personal finance class didn't actually score any higher on financial literacy tests later on.

Because traditional classes don't always work, educators have started looking at competition-based learning (CBL). Programs like the Stock Market Game, where students

simulate real-world investing, have been shown to boost financial literacy (Hinojosa et al., 2010). This works because games are highly interactive, giving students instant feedback and making them want to learn (Mandell & Klein, 2007). In Malaysia, researchers also found that competition-based learning made university business students feel more confident and capable (Nor et al., 2022).

To understand why some learning methods stick while others don't, we can look at how motivation works. Mandell and Klein (2007) showed that factoring in student motivation helped predict financial literacy scores much better. More recently, Kusurkar et al. (2023) looked at how exams affect student drive. They separated motivation into two types: 'autonomous' motivation (doing something because you genuinely care about it) and 'controlled' motivation (doing something because of external pressure, like grades or winning a trophy). They found that autonomous motivation leads to deep, long-term learning. On the flip side, controlled motivation—which is very common in competitive, high-stakes environments—tends to lead to 'surface-learning' where students quickly memorize facts for a test and then forget them just as fast (Kusurkar et al., 2023).

This issue also ties into school funding. Not every school has the money to train teachers for specialized financial literacy classes (Mancone et al., 2024). This creates a gap between wealthy school districts and underfunded ones, exacerbating existing disparities in financial literacy across socioeconomic groups (Wagner, 2019).

This is where Future Business Leaders of America (FBLA) comes in. FBLA is the largest student business organization in the world (Future Business Leaders of America, 2026). It relies heavily on self-directed study and competitive exams. Unlike the Stock Market Game, FBLA testing is mostly multiple-choice exams on specific business topics, and students have to study for them on their own. Even though FBLA is incredibly popular, nobody has actually studied whether participating in it leads to better overall financial literacy.

My study aims to fill this gap. I wanted to see if being in FBLA correlates with better financial literacy among students at Parsippany Hills High School (PHHS). Unlike simulations or formal classes, FBLA is mostly self-guided test preparation. I measured financial literacy using an 11-question quiz covering interest, savings, loans, investments, and inflation, adapted from Baistaman et al. (2020). Then, I analyzed how these scores related to how long students had been in FBLA. If there is a strong link, it could prove that student-led clubs are a cheap, scalable way to improve financial education everywhere.

Methods

Study Design

I set up a cross-sectional, correlational study to see if there was a relationship between how long students had been in FBLA and their financial literacy scores.

Participants and Sampling

The study focused on students in grades 9 through 12 at Parsippany Hills High School (PHHS). I aimed to get 50 total survey responses, with at least 30 coming from FBLA members. I reached out to students through two main methods:

1. Control group (non-FBLA participants): I asked English teachers at PHHS to share the survey link on their Google Classrooms. I chose English classes because almost everyone has to take English, which helped me avoid surveying only business students.

2. FBLA participant group: I posted the survey on our school's FBLA Google Classroom and emailed members directly.

Taking the survey was completely voluntary and anonymous. I included an informed consent form for both students and parents. I kept the survey open for 10 days.

Instrument

To measure financial literacy, I used an 11-item multiple-choice quiz. It was adapted from a validated test created by Baistaman et al. (2020), which was tested for reliability using factor analysis. The quiz covered five key areas, shown in Table 1:

Table 1. Financial Literacy Domains Covered in the Quiz

Domain	Description
Interest	Calculating simple and compound interest (like on a savings account)
Savings	Knowing healthy saving habits (like how much of your monthly income to save)
Loans	Understanding debt (like limits on monthly loan payments)
Investment	Basic investing rules (like risk vs. reward and diversification)
Inflation	Understanding how inflation affects purchasing power and borrowers

For each question, students chose from four options (A–D). I also added a "Do not know" option to keep people from guessing randomly, which helped make the scores more accurate.

Demographic and Independent Variables

Along with the financial questions, the survey asked for:

In addition to the financial literacy items, the questionnaire collected the following variables:

- What school they attend (to make sure they go to PHHS)
- How many years they have been in FBLA (0 to 4 years)
- If they are currently in FBLA, or when they left (ended 1–3 years ago)
- Grade level (Freshman, Sophomore, Junior, Senior)
- Any business classes they took (Personal Finance, AP Economics, etc.)
- If they were also in DECA (another business club)

Administration Procedures

The test was given using Google Forms. To keep things fair and consistent, I set up a few rules:

- **No time limit:** Students could take as long as they needed.
- **No calculators:** Since students took this remotely, I couldn't make sure everyone was using the same calculator. Instead, I kept the math simple and focused on the concepts.
- **No help:** Students had to complete the quiz on their own.
- **Anonymity:** No names or personal details were collected.

Data Exclusion Criteria

I received 21 total responses. Before analyzing the data, I decided to remove any response with a score of zero (0/11) because it showed a lack of effort (the lowest score from anyone else was 3/11). Two responses met this criteria and were removed. This left me with a final sample size of $N = 19$ (13 FBLA members and 6 non-members).

Statistical Analysis

I organized the data in Google Sheets, grouping students by their grade level, years in FBLA, and how long ago they participated. Then, I used Pearson's correlation coefficient (r) to see if there were any real connections between:

Each response was organized by demographic variables (years of FBLA participation, recency of participation, grade level), total score, and sub-scores across the five financial literacy domains using Google Sheets.

The **Pearson product-moment correlation coefficient (r)** was used to determine the strength and direction of the relationship between:

- Years in FBLA and overall test scores
- Years in FBLA and specific category scores (like interest or inflation)
- Years since leaving FBLA and test scores
- Grade level and test scores

I set the significance level at $\alpha = 0.05$. Because my sample size was relatively small ($N = 19$), I needed a correlation of at least $|r| \geq 0.433$ to call a result statistically significant. I rounded all correlation calculations to three decimal places.

Alignment and Rationale

To my knowledge, no other study has looked directly at how FBLA membership duration relates to general financial literacy. Using an objective multiple-choice test was important to prevent any researcher bias, especially since I serve as the FBLA Vice President at PHHS and obviously wanted to see FBLA members do well.

Ethical Considerations

- Participation was entirely voluntary, and I made sure to get parental consent.
- All data was kept completely anonymous.

- I didn't collect any sensitive personal details, like family income, to protect student privacy and encourage honest answers.

Results and Discussion

Results

After weeding out the two zero-effort responses, I ended up with 19 valid responses ($N = 19$). The group included 13 active or former FBLA members and 6 non-members. The grade distribution was: 4 freshmen, 8 sophomores, 6 juniors, and 1 senior.

Primary Correlation Analysis:

I ran a Pearson correlation to see if students who spent more years in FBLA had better financial literacy scores. The math showed virtually no relationship: $r(19) \approx 0.095, p \approx 0.698$. Since this value is way below our threshold of $|r| \geq 0.433$, the result is not statistically significant. This goes against my original hypothesis that more time in FBLA would equal higher scores.

Segmented Score Analysis

I also checked if FBLA was linked to better scores in specific areas, like saving or investing. The correlations for each category are listed in Table 2.

Table 2. Pearson Correlation Coefficients ($N = 19$)

Score Category	Years of FBLA Participation	Years Ago Participation	Grade Level
Total score	0.095	-0.580^{*}	0.165
Interest questions	-0.043	-0.083	-0.060
Savings questions	-0.124	-0.158	-0.172
Loan questions	0.023	-0.067	-0.054
Investment questions	0.183	-0.518^{*}	0.254
Inflation questions	0.207	-0.808^{*}	0.286

**Statistically significant at* $\alpha = 0.05$ *(*$|r| \geq 0.433$*)*

As Table 2 shows, none of the individual categories showed a significant positive correlation with FBLA participation. This means FBLA membership didn't help students perform better in any single area of personal finance.

Secondary Finding: Time Since Participation

Interestingly, I found strong negative correlations when looking at the "years ago participation" variable (how long ago a student left FBLA). There were significant negative links with total scores ($r \approx -0.580, p \approx 0.009$), investment scores ($r \approx -0.518, p \approx 0.024$), and inflation scores ($r \approx -0.808, p < 0.001$). This seems to show that students who left FBLA in earlier years scored lower over time. However, I have to be careful with this finding because only one student in my sample was an inactive FBLA member, so these correlations are basically driven by a single person's data.

Discussion

Why didn't FBLA experience make a difference in financial literacy? We can explain this using Kusurkar et al.'s (2023) theory on motivation and learning. They argue that the *type* of motivation matters:

FBLA's competitive setup mostly creates "controlled" motivation. The main goal is to study for a specific, high-stakes multiple-choice test to win an award or move on to the next level of competition. This leads to surface-level learning—students memorize facts to pass the test, but they don't actually digest or retain the information.

There is no real room for creativity or practical application in these exams, which further dampens autonomous motivation.

In contrast, other business competitions like the Stock Market Game (SMG) have been shown to actually boost financial literacy (Hinojosa et al., 2010). The difference is in the design. The SMG gives students autonomy (they choose their own stocks), instant feedback (they see their portfolio change in real-time), and has lower stakes, which builds autonomous motivation. Students participate because they are genuinely interested in seeing their virtual money grow, not just because they want to pass a test.

My results match up with what Mandell and Klein (2009) found: just taking a personal finance class in high school doesn't automatically mean you will score higher on financial literacy tests later. Exposure to facts is not enough if the motivation isn't there. Similarly, Kusurkar et al. (2023) pointed out that memorizing facts under high-pressure testing leads to quick knowledge decay.

Even though FBLA does have presentation events that allow for more creativity, these are still high-pressure competitions, and many of them still require taking an objective test first. So, the overall environment remains highly focused on controlled motivation. The negative correlations we saw in students who left FBLA earlier (though based on very limited data) also support the idea of rapid knowledge decay. When you study just to pass a test, you forget the material as soon as the test is over.

Limitations

My study has a few clear limitations:

- **Small sample size:** I only analyzed data from 19 students, which was short of my goal of 50. This makes it harder to find statistically significant trends.

- **Cross-sectional design:** I only looked at a snapshot in time, rather than testing students before and after they joined FBLA.
- **Lack of background controls:** I couldn't control for students' family income or other business classes they might have taken, though surveying English classes helped keep the sample relatively unbiased.

Implications

These results suggest that self-directed studying combined with high-stakes testing might not be the best way to teach financial literacy. If we want kids to actually remember how money works, programs should focus on building autonomous motivation. This means giving students more freedom to explore, keeping the pressure low, using collaborative projects, and providing instant feedback, rather than just relying on competition and multiple-choice tests.

Conclusion

This study looked at how being in Future Business Leaders of America (FBLA) relates to financial literacy among high schoolers at Parsippany Hills High School. Using an 11-question validated test and analyzing responses from 19 students, I found no statistically significant correlation between years in FBLA and overall financial literacy scores ($r(19) \approx 0.095, p \approx 0.698$). I also found no significant link between FBLA participation and any specific financial area, such as interest rates, savings, loans, investing, or inflation.

These findings show that simply being in a business club like FBLA doesn't guarantee a student will actually learn and retain personal finance skills. This aligns with Kusurkar et al.'s (2023) research, which explains that high-stakes, objective-test-based competition creates external, controlled motivation, leading to surface learning that quickly fades.

For future financial literacy efforts to work, they should move away from raw test preparation. Instead, programs should focus on active, low-pressure, and hands-on learning experiences that encourage genuine interest. While my small sample size means we should be careful about generalizing these results, this study highlights the need for more hands-on research and structural changes in how we teach business and finance in high school.

Works Cited

1.Organisation for Economic Co-operation and Development. (2005). Improving financial literacy: Analysis of issues and policies. OECD Publishing. https://doi.org/10.1787/9789264012578-en

2.McCormick, M. H. (2008). The effectiveness of youth financial education: A review of the literature. Journal of Financial Counseling and Planning, 20(1), 70-83.

3.Mandell, L. (2008). The financial literacy of young American adults. Jump$tart Coalition for Personal Financial Literacy.

4.Hinojosa, T., Miller, S., Swanlund, A., Hallberg, K., Brown, M., & O'Brien, B. (2010). The impact of the Stock Market Game on financial literacy and mathematics achievement: Results from a national randomized controlled trial. Society for Research on Educational Effectiveness.

5.Nor, A. M., Ibrahim, S. N., Hashim, N., & Mohamed, M. (2022). Does competition-based learning enhance learning ability? A preliminary study on finance education. International Journal of Academic Research in Business and Social Sciences, 12(10). https://doi.org/10.6007/IJARBSS/v12-i10/14918

6.Future Business Leaders of America. (2026). About FBLA. https://www.fbla.org

7.Mandell, L., & Klein, L. S. (2007). Motivation and financial literacy. Financial Services Review, 16, 105-116.

8.Kusurkar, R. A., Orsini, C., Somra, S., Artino, A. R., Jr., Daelmans, H. E. M., Schoonmade, L. J., & Van der Vleuten, C. P. M. (2023). The effect of assessments on student motivation for learning and its outcomes in health professions education: A review and realist synthesis. Academic Medicine, 98(9), 1083-1092. https://doi.org/10.1097/ACM.0000000000005263

9.Moscarola, F. C., & Kalwij, A. (2021). The effectiveness of a formal financial education program at primary schools and the role of informal financial education. Evaluation Review, 45(3-4), 107-133. https://doi.org/10.1177/0193841X211042515

10.Mancone, S., Tosti, B., Corrado, S., Spica, G., Zanon, A., & Diotaiuti, P. (2024). Youth, money, and behavior: The impact of financial literacy programs. Frontiers in Education, 9. https://doi.org/10.3389/feduc.2024.1397060

11.Organisation for Economic Co-operation and Development. (2014). Financial education for youth: The role of schools. OECD Publishing. https://doi.org/10.1787/9789264174825-en

12.Organisation for Economic Co-operation and Development. (2024). PISA 2022 results (Volume IV): How financially smart are students? OECD Publishing. https://doi.org/10.1787/5a849c2a-en

13.Mandell, L., & Klein, L. S. (2009). The impact of financial literacy education on subsequent financial behavior. Journal of Financial Counseling and Planning, 20(1), 15-24.

14.Baistaman, J., Awang, Z., Afthanorhan, A., & Abdul Rahim, M. Z. (2020). Developing and validating the measurement model for financial literacy construct using confirmatory

factor analysis. Humanities and Social Sciences Reviews, 8(2), 413-422. https://doi.org/10.18510/hssr.2020.8247

15. Wagner, J. (2019). Financial education and financial literacy by income and education groups. Journal of Financial Counseling and Planning, 30(1), 132-141. https://doi.org/10.1891/1052-3073.30.1.132

Author

Abhyuday Shashi is a senior at Parsippany Hills High School enrolled in a STEM pathway, with strong academic and professional interests in business, economics, and leadership. He has demonstrated excellence through academic research, national-level business competitions, and selective entrepreneurship programs at institutions such as Columbia University, Harvard University, and NJIT.

To What Extent Does Impulse Buying and Social Influence Lead to Irrational Spending in the Fashion Industry? By Alexandra Filip

Abstract

The rapid growth of fast fashion, social media and digital marketing has contributed to increasing levels of clothing consumption and raised concerns about the factors that encourage consumers to make unnecessary and irrational purchases. This research investigates the extent to which impulse buying and social influence contribute to irrational spending in the fashion industry, with particular focus on consumer psychology and marketing strategies. Existing academic literature and survey findings were analysed to examine how factors such as social media, influencer marketing, social comparison, fear of missing out (FOMO), personalised recommendation algorithms, repeated exposure, psychological pricing and scarcity marketing influence consumer behaviour. The findings indicate that impulse buying and social influence contribute to irrational spending to a significant extent, as these factors can create emotional, cognitive and social pressures that increase consumers' likelihood of making impulsive purchases. However, the research also demonstrates that these influences do not affect all consumers equally. Financial literacy, self-control, mindfulness, personality traits and individual values can reduce consumers' susceptibility to external and emotional purchasing triggers. Survey findings similarly suggest that social influence has a moderate effect on clothing consumption, while many respondents did not report being influenced by celebrity or influencer marketing. Therefore, the relationship between impulse buying, social influence and irrational spending should not be interpreted as purely causal. Overall, irrational spending in the fashion industry is best understood as the result of an interaction between external, social and marketing influences and individual psychological characteristics, highlighting the importance of self-awareness and self-control in encouraging more conscious consumption.

Introduction

In the past few years, the fashion industry has seen a substantial increase in consumer demand for clothing. This is largely due to an increase in the popularity of fast fashion prompting considerably higher volumes of purchase. Fast fashion refers to a business model in which inexpensive clothing is rapidly produced to keep pace with changing trends. This is becoming an increasingly significant issue as companies produce and innovate new apparel quickly as trends change over time, offering cheap and quick delivery. While this has created profitable opportunities for businesses, and fashion has become more accessible around the world, it has also played a major role in encouraging overconsumption, particularly among young people aged 15 to 29. Fashion overconsumption can be defined as buying fashion items irresponsibly or consuming so many fashion items that it leads to negative consequences (Hakansson, 2014; Pan, 2019). Factors such as social media marketing encourage customers to engage in impulse buying, which reinforces patterns of overconsumption. Impulse buying is an unplanned, spur-of-the-moment decision to purchase a product or service. As a result, high levels of clothing

consumption among young people have attracted increasing attention due to their economic and social implications. Two factors commonly associated with excessive clothing consumption are irrational spending and social influence. In the fashion industry, irrational spending often arises when consumers make financial decisions driven by emotions, cognitive biases, or instant gratification rather than logic. Such behaviour is frequently shaped by social influence, as individuals adjust their attitudes, beliefs, or behaviors to meet the demands of a social environment or the expectations of others.

This paper aims to demonstrate that one of the most significant forms of social influence in today's society is social media, which has been linked to increasing levels of overconsumption. Easy access to the market through e-commerce, digital platforms and social media marketing, motivate consumers toward impulse purchases and emotional buying behaviour. The widespread presence of advertising, for example through ads on social media platforms, can influence consumer behaviour by creating temporary feelings of satisfaction that encourage further purchases, resulting in consumers perceiving their desires as needs. Since individuals spend an average of 5-6 hours per day online (Statista, 2021), they are regularly exposed to marketing content that can shape their purchasing decisions indirectly. This repeated exposure may lead consumers to associate increased consumption with happiness, reinforcing patterns of impulsive and irrational choices, thus leading to overspending, increased debt and unsustainable consumption decisions that may have long term environmental consequences. However, impulse buying and social influence do not always result in irrational spending, as consumers can adopt mindful consumption practices through sustainable fashion choices.

Consumers who are well-informed are more likely to consider the "cause and effect" of purchasing fast fashion and the consequences it may have for society. Hence, sustainable consumption motivates consumers to avoid unnecessary purchases associated with fast fashion (Brydges et al., 2021; Ertekin and Atik, 2020), and make responsible decisions based on rational decision-making rather than peer pressure, social media, and other external influences. This may reduce the prevalence of impulse buying. Sustainable consumption promotes slow fashion and more moderate purchasing habits (Hassan et al., 2022), contradicting fast fashion which encourages irresponsible consumption. A prior study found that consumers who engage in sustainable fashion consumption incline toward satisfying their utilitarian needs instead of hedonistic values (Razzaq et al., 2018).

This paper investigates why excessive clothing consumption is associated with impulse buying and social influence, and how these factors contribute to irrational spending in the fashion industry. This contributes to a greater understanding of environmental challenges and the consequences associated with unsustainable fashion practices. This study is significant because it uncovers the underlying motivations behind irrational spending in the fashion industry, particularly the roles of impulse buying and social influence, including the impact of social media on consumer behaviour. Therefore, this study seeks to examine the extent to which impulse buying and social influence lead to irrational spending in the fashion industry.

Methodology

This chapter aims to summarize a comprehensive body of resources used to support the methodological framework necessary to identify, select and analyse the economical and consumer behaviour literature that depicts the relationship between impulse buying, social influence and irrational spending in the fashion industry. The purpose of this review is to analyse the impact of impulse buying and social influence on consumer behavior in the fashion industry, taking into account their role in shaping purchasing decisions and contributing to irrational spending. The findings support the primary goal of the thesis: impulse buying and social influence play a significant role in shaping consumer behaviour in the fashion industry and may contribute to irrational spending**.** To identify the most relevant and up-to-date evidence, searches were conducted using databases such as Google Scholar, Research Gate, Google Books, Iowa State Digital Press, Cell Press, IJMCER, Norma Group, MDPI, PLOS ONE, and Frontiers. The research period began May 2026 and extended to September 2026. The keywords used during the process, as well as their various combinations, included "impulse buying," "social influence," "irrational spending," "fast fashion" and "sustainable". Boolean operators (AND/OR) were used to narrow or expand search results as the need for more thorough search parameters arose.

In the context of this thesis, the studies spanned between 2012 to 2026 to ensure a comprehensive analysis of the topic and provide credible social and economical evidence on the importance of irrational spending in the fashion industry. Moreover, only articles published in English were considered as it offered a broader library of resources to formulate compelling findings.

The initial search yielded a library of approximately 112 articles. After a thorough review of their titles and abstracts, only 70 articles were selected for in-depth reviews, of which 33 met the date, methodology, and accuracy criteria established. The following variables were decisive in the process of data extraction from the final pool of studies: type of study, publication year, sample size, consumer demographics, social media exposure, and factors influencing irrational spending in the fashion industry.

The economical and social literature reviewed revealed trends consistent with the findings of this thesis. While impulse buying and social influence significantly affect consumer behaviour in the fashion industry, their impact on irrational spending may vary depending on individual preferences and consumption habits.

The Nexus of social Influence, impulse buying & status

According to past surveys, young fashion customers use social media platforms such as Instagram to actively search for fashion inspiration, which influences their purchasing decisions (Rosely, N., & Syed Ali, S. F., 2023). Although the traditional economic theory assumes that individuals make decisions based primarily on their material needs, research shows that social needs, such as desire for social status, can also play a major role in consumption in the fashion industry (Syed Sibghatullah Shah; Zahid Asghar, 2023). People often use material possessions to express their social identity and showcase their achievements. This is explained by Veblen's

theory (1899) as he found that individuals purchase and publicly display luxury goods not only for their practical value but also as a sign of wealth, prestige and social status. For example, owning luxury goods such as a Rolex watch, a Hermès Birkin bag, or a high-end sports car like a Porsche can signal wealth and social image. These possessions are often used to demonstrate financial success and gain recognition or admiration from others. According to Self-Presentation theory (Goffman, 1959) individuals consciously manage the image they present to others in order to create favorable impressions and gain social approval. Displaying luxury goods therefore becomes a way of communicating success, social status, and admiration. This can reinforce self-esteem through positive social feedback and acceptance from others. Over time, these external responses can become incorporated into an individual's self-identity, as people begin to define themselves through the image they project and the possessions they own. Consequently, material consumption becomes not only a means of satisfying functional needs but also a way of maintaining a desired social identity. This psychological mechanism helps explain why some people engage in material consumption more frequently than others, particularly when consumption is linked to their self-identity. This desire for approval and social recognition is frequently reflected in consumers' preference for highly visible or "loud" brands. Such brands are chosen because they are easily recognised by others, and social media platforms provide consumers with an opportunity to display these purchases to a large audience. Consequently, individuals may engage in public consumption in an effort to gain attention and enhance their self-image. Consumers who seek status are particularly drawn to prominent brands, as their visibility can strengthen perceptions of social standing, reinforce personal identity, and increase feelings of self-worth among members of their social group (Rosely, N., & Syed Ali, S. F., 2023). The impact of social influence on purchasing behaviour is further demonstrated by survey findings, which revealed that approximately 45% of the respondents purchased new clothing as a result of social expectations when attending events or interacting with their peers (Dr. Meenakshi Jindal; Rudraveer Suri; Kudratjot kaur; Tanmay Kansra; Aryan Sahoo, 2025). However, these findings also imply that social influence is not universal and that many consumers make purchasing decisions based on factors other than status or peer approval.

The rapid growth of social media and influencer marketing has transformed the way consumers discover, evaluate and purchase fashion products. Platforms such as Instagram and Facebook have repeatedly demonstrated that the content shared on those platforms drives consumers to make frequent impulsive purchases, leading to excessive consumption. For example, Facebook facilitates impulse buying by allowing users to purchase products directly through the platform, while also exposing them to personalized advertisements, making it easier for consumers to succumb to impulse buying. Moreover, Instagram's interactive visual features enhance users' satisfaction while browsing the advertisement. Research has found that these features significantly stimulate impulsive buying by encouraging consumers to skip the evaluation phase of the decision-making process (Rosely, N., & Syed Ali, S. F., 2023). This occurs, because visually appealing content, limited- time offers and personalised advertisements often trigger emotional rather than rational decision making. Consumers rely on cognitive

shortcuts (heuristics), such as assuming the products promoted by the influencer or celebrity are trustworthy, desirable, or of high quality. Consequently, they spend less time evaluating other alternatives or considering the financial consequence of the purchase. This makes consumers potentially more prone to persuasive marketing messages, particularly those delivered by celebrities and social media influencers. Although social media has transformed fashion marketing, consumers have long been influenced by celebrities and public figures when making clothing choices. With the rise of social media, the number of social media influencers has grown alongside the expansion of these platforms. Younger generations have been particularly affected as social media influencers significantly influence consumer behaviour (Shah S, Asghar Z, 2023). Consumers increasingly rely on influencers as trusted and relatable sources of information when making purchasing decisions, contributing to higher levels of impulse buying and excessive fashion consumption. Nevertheless, the influence of social media may not be as significant as previous studies suggest. Findings from a study indicate that 46% of respondents did not believe social media encouraged them to purchase more clothing than necessary, while only 29% agreed that social media trends or influencer endorsements affected their purchasing decisions (Dr. Meenakshi Jindal, Rudraveer Suri, Kudratjot kaur, Tanmay Kansra, Aryan Sahoo, 2025). One possible explanation is that increasing digital literacy has made young consumers more aware of the commercial intentions behind influencer marketing. Instead, social media may influence purchasing behaviour indirectly through mechanisms such as social comparison and fear of missing out rather than through direct exposure to influencer content (Dr. Meenakshi Jindal, Rudraveer Suri, Kudratjot kaur, Tanmay Kansra, Aryan Sahoo, 2025).

Table 4: Social media influence on clothing overconsumption

S.No	Drivers	Item	Strongly agree	Agree	Neutral	Disagree	Strongly disagree
10	Social Media influence	I buy clothes to fit in the latest social media trends or designs .	11.8	27.5	25.5	27.5	7.8
11		I tend to buy clothes that are promoted by celebrities and influencers on social media .	6.9	5.9	26.5	41.2	19.6
12		I wear new clothes to enhance my social media image.	11.8	23.5	22.5	29.4	12.7
	Average		10.17	18.97	24.83	32.7	13.37

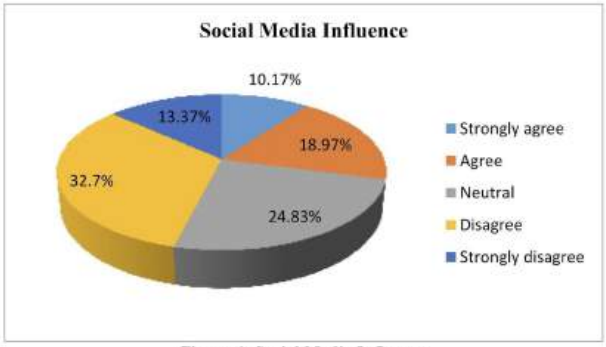


Figure 4: Social Media Influence

Beyond influencers, recommendation algorithms have become another powerful form of digital influence, shaping consumers' purchasing decisions by personalizing the content and product they come across online. The Lyst Index, a quarterly ranking of the world's most popular fashion brands and products, analyses consumer browsing and purchasing behaviour using data such as product views, brand searches, social media mentions, as well as the overall engagement statistics worldwide (Gigauri, Z., 2024). This information enables fashion brands and online retailers to create highly personalized shopping experiences through artificial intelligence and recommendation algorithms. By analyzing consumers' browsing history, preferences and previous interactions, digital platforms can recommend products that closely match their individual interests, making the shopping experience more convenient and personalized. A real-world example of this can be observed on TikTok Shop, where artificial intelligence algorithms analyse users' browsing behaviour, viewing history, likes, searches and interactions with content to identify their interests. The platform then recommends products that closely match these preferences on users' "For You" pages and within TikTok Shop (Minh Thi Hong Le, Hieu Ngoc Minh Pham, Ly Thi Truc Nguyen, Thuy Thi Xuan Le, Suong Thi Mai Nguyen, 2025). By repeatedly exposing consumers to products that align with their existing interests, the recommendations reduce the effort required to search for items and increase the likelihood of spontaneous or impulse purchases. Research suggests that these personalized environments increase the likelihood of impulse buying among Generation Z consumers (Murhundezi, Thierry Junior, 2024). Moreover, these environments appear to have a particularly strong influence on Generation Z, whose spending behaviour has been shaped by the rapid growth of social media and digital innovation. As a result, consumers may rely on emotional responses rather than rational evaluation when making purchasing decisions, increasing the likelihood of irrational spending (Murhundezi, Thierry Junior, 2024). However, the influence of personalised digital environments should not be overemphasized. Studies suggest that individuals are less likely to engage in herding behaviour when preferences are diverse, as they are more capable of making independent and informed decisions despite increasing exposure to fashion trends. Furthermore, increasing awareness of environmental issues has encouraged more individuals to adopt sustainable purchasing practices, demonstrating that greater knowledge and conscious decision-making can reduce consumers' tendency to engage in impulse buying despite personalised marketing (Syed Sibghatullah Shah, Zahid Asghar, 2023). A clear example of this shift is the growing popularity of thrifting. As second-hand fashion becomes increasingly accepted, consumers are beginning to prioritise sustainability over constant consumption of new clothing, demonstrating how greater awareness can encourage more rational purchasing behaviour.

Alongside targeted advertising algorithms, social media influencers have become one of the most influential marketing tools used by fashion brands to shape consumer behaviour. The rapid growth of the creator economy has transformed content creation into a viable career, with influencers generating income through brand partnerships, sponsored advertisements, affiliate marketing and paid memberships. Through platforms such as Instagram, TikTok and YouTube,

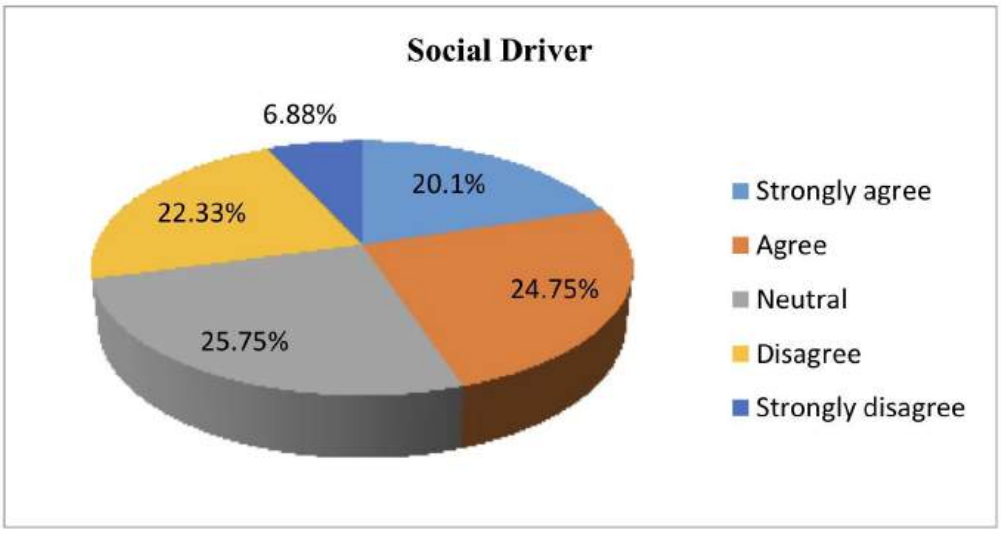
influencers promote products to millions of followers, allowing luxury and fast fashion brands to reach consumers in a more personal and engaging way. As Generation Z is highly influenced by online personalities and celebrity culture, fashion companies increasingly collaborate with influencers to promote their products and strengthen brand visibility (Murhundezi, Thierry Junior, 2024). Therefore, content created by celebrities and social media influencers has been shown to have a significant impact on the impulse purchasing behaviour of this generation's consumers (Rosely, N., & Syed Ali, S. F., 2023). Studies suggest that social media influencers possess considerable influence over their followers and can encourage consumers to purchase products they didn't originally intend to buy. Compared with traditional advertising, influencer recommendations are often perceived as more authentic, relatable, and trustworthy, increasing consumer engagement and strengthening purchase intentions. This influence can be explained by the Affect Infusion Model, which suggests that emotions become incorporated into cognitive processing, meaning that individuals' feelings can influence how they evaluate information and make decisions, particularly when those decisions require personal judgement rather than careful analysis. As influencers are often perceived as authentic and relatable, the positive emotions they evoke can increase the probability of impulse purchases among consumers (Murhundezi, Thierry Junior, 2024). However, the influence of social media influencers may not be as common as the studies suggest. Findings from the present study indicate that 60% of respondents disagreed that they purchased clothing promoted by celebrities or social media influencers. These findings suggest that although influencer marketing affects some consumers, many Generation Z consumers appear to make purchasing decisions independently of celebrity opinions, indicating that influencers are only one of several factors shaping fashion consumption.

While influencers shape purchasing behaviour through emotional appeals and perceived authenticity, their influence is strengthened by social comparison, where individuals compare themselves and their possessions with others. This process is explained by Social Comparison Theory (Festinger, 1954), which proposes that individuals evaluate their own abilities, achievements, and social status by comparing themselves with others, particularly when objective standards are unavailable. As a result, exposure to idealised lifestyles on social media can influence consumers' perceptions of their own social standing and encourage them to purchase products that help them achieve a desired identity. Young people who use social media more frequently are exposed to carefully curated images of fashion, luxury products and lifestyles, making them more sensitive to feelings of envy. Platforms such as Instagram are widely used by young consumers to seek fashion inspiration, influencing their purchasing decisions. At the same time, social media encourages users to display their material possessions, which can evoke envy among viewers, and therefore contribute to social comparison (Rosely, N., & Syed Ali, S. F., 2023). The feelings of envy created through social comparison often develop into a fear of missing out (FOMO), particularly in the fast fashion industry where trends fade rapidly. For example, products that go viral on TikTok, such as the trendy shoes Adidas Sambas or the Rhode Peptide Lip Treatment, often create a sense of urgency as consumers fear they will sell out or become the next "must-have" item. This encourages consumers to purchase quickly

before the trend disappears. Moreover, consumers may experience “a feeling of lacking” if they are unable to purchase the product they desire. This perceived deprivation encourages consumers to buy new styles immediately in an attempt to overcome feelings of anxiety, regret, and sadness. Ultimately, this behaviour is driven by the desire to remain connected with others, and keep up with changing social trends (Rosely, N., & Syed Ali, S. F., 2023). Fear of missing out (FOMO) is closely linked to conformity, as individuals often adjust their behaviour to align with the expectations of those around them. This occurs when the individual wants to be accepted and liked by others and not be excluded by them. Studies have shown that conforming behaviour is common within online communities, where exposure to popular trends and the consumption habits of others can shape individual purchasing decisions (Gigauri, Z., 2024). Therefore, interest in fashion trends is closely associated with impulsive buying behaviour in the fast fashion industry (Barbara Gawior, Michal Polasik and Josep Lluís Del Olmo, 2022). Although social media and peer opinions can encourage conformity, relying on external sources of information may also expose consumers to a wider range of perspectives, enabling more informed and rational purchasing decisions. The findings presented in Table 2 and the Social Driver pie chart indicate that social factors have a moderate influence on clothing overconsumption. Approximately 45% of respondents reported purchasing new clothing because of social expectations. This suggests that social norms continue to shape consumer behaviour. Furthermore, they imply that social influence is often indirect rather than explicit, reinforcing the role of FOMO and the desire for social acceptance in encouraging unnecessary fashion purchases (Dr. Meenakshi Jindal, Rudraveer Suri, Kudratjot kaur, Tanmay Kansra, Aryan Sahoo, 2025).

Table 2: Social Driver of clothing overconsumption

S.No	Drivers	Item	Strongly agree	Agree	Neutral	Disagree	Strongly disagree
4	Social	I am highly influenced by my peers to frequently buy clothes .	9.8	18.6	31.4	29.4	10.8
5		I prefer to wear new or fashionable clothes to impress others.	12.7	22.5	21.6	34.3	8.8
6		When I wear new or fashionable clothes I feel included in my social circle.	10.8	27.5	36.3	18.6	6.9
7		I prefer to wear new clothes for specific events like birthdays, parties or weddings etc.	47.1	30.4	13.7	7.0	1.0
	Average		20.1	24.75	25.75	22.33	6.88



Overall, these findings suggest that awareness of social media marketing and consumer psychology can reduce consumers' likelihood to impulse buying, but it does not completely eliminate the desire to purchase. Emotional motivations, social comparison, and the need for social acceptance continue to influence purchasing behaviour even when individuals recognise the persuasive techniques used by brands and influencers. Therefore, knowledge alone is insufficient to prevent irrational spending, although it can encourage more conscious and informed consumption decisions.

Marketing Strategies that affect consumer psychology

Dupes, short for duplicates, have become a prominent consumer trend over the last few years and refer to products that closely resemble high-end designs while being sold at a significantly lower price (William Kim and Jung E. Ha-Brookshire, 2026). Unlike counterfeit products, dupes imitate the appearance or functionality of popular brands without attempting to deceive consumers into believing they are actually purchasing the original product (Florencia Reagan, Keeya Averielle Tejasukmana, Matthew Toby, 2025). In an era where social media heavily influences spending patterns, dupes serve as a bridge between desire and affordability, allowing many consumers to experience the status and prestige associated with luxury goods in a more realistic way. Furthermore, algorithm-driven recommendation systems on platforms such as TikTok and Instagram have accelerated the popularity of dupe culture by repeatedly exposing users to comparison videos and product recommendations. Short form videos such as TikToks and reels showing side-by-side comparison between the luxury products and the dupes. The rise of dupe culture started around the mid-2010s, when fast fashion and social media started growing. Brands like Zara and Shein normalized dupes in clothing, while content creators amplified the idea of dupes through "hauls" and comparison videos (Meichi Wu, 2026). Consumers often use material possessions to communicate their identity, self-worth and social

status rather than simply satisfy functional needs (Rosely & Syed Ali, 2023). Luxury fashion, in particular, serves as a visible symbol of wealth, status and identity (Gigauri, 2024). Consequently, individuals who can't afford luxury products may turn to dupes as an affordable alternative, enabling them to project a similar image and fit within desired social groups. This reflects Veblen's theory of conspicuous consumption, which argues that consumers purchase goods not only for their functional value but also to display status and gain social recognition. According to Veblen (1918), conspicuous consumption enables individuals to communicate wealth and prestige through visible consumption, suggesting that products often derive value from the social meanings they convey rather than their practical usefulness. Although dupes lack the exclusivity and price associated with luxury brands, they allow consumers to imitate many of the same visual status signals at a fraction of the cost. Similarly, Eastman, Goldsmith and Flynn (1999) describe conspicuous consumption as a motivational process through which individuals seek to improve their social standing by purchasing products that symbolise status. This suggests that consumers may purchase dupes not because they believe they possess superior quality, but because they communicate a desired social image and satisfy the need for social acceptance(Rajasekhara Mouly Potluri, Rizwana Ansari, Siva Kumar Challa, Lavanya Puttam, 2014). This argument is further supported by Hasim et al. (2020), who found that consumers often perceive dupes as a practical alternative to luxury brands, particularly in societies where appearance and social status are highly valued. Their study suggests that affordability is a key determinant of purchasing decisions, as consumers are able to maintain a desired social image through replica products without bearing the financial burden associated with authentic luxury brands (Muhammad Zafran and Aiman Irfan Ahme, 2025). This therefore suggests that the symbolic value attached to luxury goods often outweighs their functional value, as consumers prioritise the social meanings associated with fashion over the products themselves.

Beyond the social influence that normalizes products such as fashion dupes, repeated exposure to products through digital advertising can also shape consumers' perceptions and purchasing decisions. One psychological mechanism that helps explain this effect is the Mere Exposure Effect (ME), which suggests that repeated exposure to a stimulus can increase familiarity and preference. Research suggests that repeated exposure increases perceptual fluency, meaning that familiar products become easier for consumers to perceive, process and recognise than unfamiliar ones. This increased familiarity can therefore contribute to more favourable attitudes towards a product, even when consumers are not consciously aware that repeated exposure has influenced their evaluation. In the context of fashion advertising, this suggests that repeatedly encountering the same product or brand across social media may gradually increase familiarity and preference, potentially increasing consumers' likelihood to impulsive buying (Martina Jakesch, Claus-Christian Carbon, 2012). In the new era of digital marketing, repeated exposure can be facilitated through personalised advertising, allowing brands to target consumers based on their previous online behaviours, such as searches and interests. One form of personalized advertising is online behavioural advertising (OBA). It involves the practice of monitoring people's online behaviour and using collected data to show

them individually targeted advertisements. For example, a consumer who searches for a particular fashion item may later encounter advertisements for the same or similar products on other websites or social media platforms. This creates repeated reminders of products that consumers have already shown an interest in, potentially reinforcing familiarity and increasing their receptiveness to purchasing them. One way in which personalized advertising can be facilitated is through cookies, which collect information about consumers' online activity and allow advertisers to target them with more relevant advertisements (Claire M. Segijna and Iris van Ooijen, 2022). This form of personalized advertising is particularly relevant on platforms such as TikTok, whose recommendation algorithm uses artificial intelligence to provide users with content based on their preferences and interests. Such personalized recommendations can increase the likelihood that consumers repeatedly encounter content related to products that already align with their interests, potentially reinforcing the familiarity described by the Mere Exposure Effect (Minh Thi Hong Le, Hieu Ngoc Minh Pham, Ly Thi Truc Nguyen, Thuy Thi Xuan Le, Suong Thi Mai Nguyen, 2023). This repeated and personalised exposure is particularly significant, because research has identified a positive relationship between social media advertising and consumers' purchase intentions (Sook Fern Yeo, Cheng Ling Tan, Kah Boon Lim, Jia Yin Leong, Yoke Chu Irene Leong, 2020). This suggests that repeated exposure on social media may not only increase familiarity with products but can also strengthen consumers' purchase intentions. However, purchase intention does not necessarily constitute impulse buying, as consumers may intend to purchase a product without making a spontaneous or unplanned decision. Therefore, the extent to which repeated and personalised advertising contributions to irrational spending may depend partly on consumers' ability to recognise and critically analyse persuasive content. Social media platforms such as Instagram and TikTok have introduced disclosure tools, such as **#ad** and **#sponsored**, to indicate when content is paid advertising. This suggests that greater advertising awareness may allow consumers to distinguish genuine product recommendations from promotional content and make more informed purchasing decisions (Ya-En Catherine Lin, Fong-Ching Chang, Tzu-Fu Huang, Shawn C. Chiang, Chiung-Hui Chiu, Ping-Hung Chen, Nae-Fang Miao, Hung-Yi Chuang, 2026).

Thirdly, another marketing strategy that affects irrational spending in the fashion industry is pricing strategies. Research indicates that decision making under uncertainty can be characterized by bias, where individuals make poor judgments. Such judgments can be strongly influenced by the first piece of information provided to the individual in order to make a judgment regarding subsequent information. Kahneman and Tversky (1979) confirmed that the first price listed for a product can influence subsequent price judgments for that same product. Furnham and Boo (2011) argued that the decision bias shown by consumers in such situations can be persistent even when the initial information has no bearing on the value of the product being considered for purchase by the consumer (Yi Zong and Xiaojie Guo, 2022). In the fashion industry for example, a retailer could display the original retail price of an item next to the reduced price for that item (e.g. 'Was £120, Now £69'). In such situations, consumers compare the reduced price with the original price of the item in order to determine the amount of savings

from purchasing the item. Thus the distorted perception of the amount of savings is created by the retailer. Furthermore, by displaying left-digit prices (e.g. £49.99) of products, a retailer creates a strong perception in consumers of a large saving. For example, £49.99 can be perceived by consumers as being 'about forty-something' rather than almost fifty. Anderson and Simester found that left-digit pricing is particularly powerful for new products where consumers have no prior price anchor for that product (Wes Hansen, 2026). The fashion retailer can thus create a perception of value by a simple change in the pricing of the product. In addition to using the anchoring effect to create perceptions of value for consumers, retailers also use the decoy effect in order to influence the purchasing decisions of consumers. According to research, consumers evaluate products such as clothing by comparing them with other similar products in order to make a decision (Bettman, Luce & Payne, 1998). A retailer can thus influence the purchasing decisions of consumers by placing different products in store. Thus by adding or removing products from the store, the retailer can increase or decrease the number of sales of a particular product (Bettman, Luce & Payne, 1998). Ariely (2008) explains that introducing a deliberately less attractive option can make a more expensive product appear better value. Although consumers rarely choose the less attractive option, its presence increases the likelihood that they will purchase the higher-priced alternative. Supporting this theory, hospitality research found that introducing a decoy significantly increased consumers' likelihood of selecting higher-priced bundle offers, particularly when individual component prices were not disclosed. Similar strategies are widely applied within fashion retail through tiered pricing structures that make premium clothing or footwear appear to provide superior value relative to carefully positioned alternatives. Promotional pricing strengthens these cognitive biases further by directing consumers' attention towards perceived gains rather than actual expenditure. Youn and Faber found that price promotions significantly increase impulse buying, while Xu and Huang similarly concluded that discounts increase consumers' intentions to purchase impulsively. Diamond and Sanyal argue that bonus offers are often perceived as gains, whereas Chandran and Morwitz demonstrate that consumers frequently interpret promotional offers as receiving something "free", increasing the perceived value of the purchase. Likewise, Chiang et al. found that online promotional campaigns significantly encourage impulse buying, particularly within digital retail environments. Collectively, these findings suggest that promotional pricing encourages consumers to focus on perceived savings rather than the money actually spent, increasing the likelihood of irrational spending. Nevertheless, these strategies do not influence all consumers equally. Bray et al. (2011) argue that some consumers actively compare prices and place greater importance on ethical considerations than promotional offers, suggesting that price sensitivity and conscious evaluation can reduce the effectiveness of psychological pricing. Therefore, while psychological pricing strategies cannot fully explain irrational spending in the fashion industry, the evidence suggests that they systematically exploit predictable cognitive biases, increasing consumers' susceptibility to impulsive and economically irrational purchasing decisions.

Lastly, a marketing strategy that can strongly influence consumer behaviour is scarcity marketing, in which marketers deliberately restrict the availability of a product either by time or

quantity to create a sense of urgency. Cialdini (2008) distinguishes between two forms of scarcity: limited-time scarcity and limited-quantity scarcity. Under limited-time scarcity, an offer is available only for a specific time period, through messages like “Sale ends Friday”, meaning that the opportunity becomes increasingly restricted as the deadline approaches. In contrast, limited-quantity scarcity restricts the number of products available, with messages such as “Only 100 units available at this price” communicating may become unavailable as the units are purchased (Shipra Gupta, 2012). Gierl, Plantsch and Schweidler (2008) further distinguished the sources of quantitative and time-based scarcity, suggesting that quantitative scarcity can result from changes in supply or demand, whereas time scarcity is determined by a restriction imposed by the supplier. Scarcity tactics can enhance product appeal by creating perceptions of limited availability, therefore stimulating consumers’ purchase desires and sense of urgency. Time-based scarcity can also compress consumers’ decision-making time, increasing the pressure to act quickly. For example, countdown timers have been found to intensify time pressure and encourage impulse buying by reducing the time available for rational decisions. Anqi Lin, Xinran Yu, Geshuo Cao, 2025). This urgency can be explained through loss aversion, a principle derived from Prospect theory developed by Kahneman and Tversky. Prospect theory suggests that consumers evaluate outcomes relative to a reference point and place greater psychological weight on losses than equivalent gains. One example is if a $100 dress drops to $70, you see it as a $30 gain. But if you expected it to cost $50, $70 feels like a $20 loss. Another example is a gain of $100 feels less significant than a loss of $100, illustrating the asymmetry between gains and losses (Frank Weissman, 2025). This asymmetry between gains and losses helps explain why scarcity can create such a strong sense of urgency. When consumers are told that a product is only available for a limited time or in limited quantities, not purchasing the product can be perceived as a potential loss of an opportunity. The possibility of losing access to the product may therefore be psychologically more significant than the potential benefit of simply obtaining it. This can encourage consumers to act immediately and avoid the anticipated loss, reducing the time available for careful consideration and increasing the likelihood of an impulsive purchase. Scarcity can also be explained through psychological reactance theory, which proposes that individuals experience a motivational response when their freedom of choice is threatened. Limited availability can therefore make consumers perceive their ability to obtain a product as restricted. This restriction can actually increase the product’s attractiveness and strengthen the consumer’s desire to obtain it, as they attempt to restore their perceived freedom. An example would be in a situation where an individual can select between Alternative A and Alternative B and that person is told to pick Alternative B (threatening the freedom to choose Alternative A), the individual is more likely to choose Alternative A in order to restore the freedom to have it, and hence Alternative A becomes more desirable (Shipra Gupta, 2012). Moreover, these psychological mechanisms can be observed in the fashion industry, particularly through limited-edition product releases. This can be seen in recent fashion marketing through Gap’s collaboration with Hailey Bieber. In July 2026, the brand released a ’90s-inspired denim collection centred around “The Hailey Jean”, which sold out within minutes of its release despite

customers setting alarms in an attempt to purchase it. The collaboration had also been teased on social media before its release, building anticipation and increasing consumer awareness of the limited opportunity to purchase the product. As a result, consumers were forced to act quickly and impulsively to secure the product before it became unavailable. This demonstrates how scarcity marketing can combine psychological responses with social influences to encourage consumers toward immediate purchasing behaviour (Zixian Tang, 2026).

Limits to Irrational Spending

Although social influence and impulse buying can contribute to irrational spending, financial literacy may reduce the likelihood of such behavior, as individuals who have greater financial knowledge may understand the consequences of impulse buying and can apply that ability when it comes to financial decisions (Wenni Wahyuandari, Latifatul Isro'iyah, Priska Rosita Damayanti, 2024). Supporting this, Aulia et al. (2023), Perry (2011), and Fraczek and Klimontowicz (2015) consistently found that lower financial literacy is associated with greater impulsive or excessive spending. This suggests that greater financial knowledge can help consumers regulate their spending and reduce their susceptibility to irrational purchases. However, even when consumers are exposed to mass marketing, scarcity, social pressure or other impulse triggers, people may not equally be affected. Individuals with higher financial literacy tend to demonstrate greater financial planning and awareness of financial risks (Priskila Panjaitan, Anwar Azazi, Ana Fitriana, M. Ridwan Ristyawan, Anggraini Syahputri, 2026). This knowledge may help individuals reconsider their needs and wants before making a purchase, potentially reducing their tendency to engage in impulse buying. However, financial literacy does not guarantee rational financial behaviour. Huston (2010) argues that factors such as behavioural and cognitive biases, self control issues, and social pressures can influence financial behaviour, meaning that even financially literate individuals may not always make financially beneficial decisions. This suggests that financial knowledge may reduce the likelihood of irrational spending, but it cannot completely prevent consumers from being influenced by emotional or social factors (Nur Aini Anisaa, Syamsul Arifinb, Lis Setyowatic, Nur Hidayahd, Aisyah Darti Megasarie, 2020).

Secondly, self-control is a crucial factor in regulating behavior. It is a personality process that enables individuals to regulate their thoughts, emotions and impulses (Meysam Moayery, Lorea Narvaiza Cantín and Juan José Gibaja Martíns, 2019). Self-control can also be perceived as the determination to resist a specific impulse, such as the urge to make an impulse purchase. As self-control can be viewed as a limited resource, depleted self-control may reduce an individual's ability to resist these urges, making impulse buying more likely. Prior research therefore found a significant relationship between self-control and impulsive buying behavior, with self-control failure increasing the likelihood of impulse purchases (Ying Wang, Jialing Pan, Yizhi Xu, Jianli Luo and Yongjiao Wu, 2022). However, the influences of external pressures, for example FOMO, may not lead to impulse buying in all consumers, as self-control can affect how individuals respond to these pressures. Research has found that self-control significantly

moderates the relationship between FOMO and impulse buying, with individuals with higher self-esteem better able to resist impulsive urges despite experiencing FOMO. Research has found that self-control significantly moderates the relationship between FOMO and impulsive buying, with individuals with higher self-control better able to resist impulsive urges despite experiencing FOMO. In contrast, individuals with lower self-control are more susceptible to making impulsive purchases driven by emotional pressure (Sun et al., 2024). This suggests that although FOMO can contribute to impulsive purchasing, its effect may depend on an individual's ability to regulate their impulses.

Additionally, dual-process theory first developed by Kahneman (2011), proposes that human cognition operates through two systems: System 1, which is fast automatic and driven by emotion, and System 2 which is slow, deliberate and analytically oriented. Under every day conditions System 2 operates when evaluating decisions and can moderate and control System 1's spontaneous impulses. However, when System 1 is strongly activated by affective stimuli, which can be triggers that create emotional responses such as limited-time offers, it can override System 2's processing, causing behaviour to be driven by immediate emotional responses rather than careful evaluation (Allendifa, A., Nasution, M, D, T, P., Fadli., 2026). Thus, when a consumer may feel the urge of making an impulsive purchase (System 1) they may need to engage in a rational evaluation (System 2), (Wijaya, A. G. Wu, W. Widiatmo, G.,2025). Therefore, System 2 allows consumers to reconsider and make a careful decision before acting on an impulse. This demonstrates another limitation to the influence of impulsive and emotional triggers, as System 1 doesn't always lead directly to irrational spending. Consumers have the decision of engaging in evaluation and overriding their initial impulse.

Another factor that may limit irrational spending in the fashion industry is mindful consumption. This encourages consumers to become more aware of their needs and their purchasing decisions. Mindfulness was first defined by Zinn (1994) as "paying attention in a particular way: on purpose, in the present moment, and non-judgmentally". As consumer decisions can often be automatic rather than carefully considered, mindfulness may reduce these automatic responses, as mindfulness is associated with greater self-control (Shahid Nawaz, Yun Jiang, Muhammad Zahid Nawaz, Syeda Farzana Manzoor, and Ruixue Zhang, 2021). This can potentially help consumers make more deliberate purchasing decisions, reducing impulse buying. This is supported by Hua (2022), who found that individuals with higher levels of mindfulness showed lower levels of impulse traits in their buying behaviour. This suggests that mindfulness helps consumers regulate their emotional responses and reduce automatic purchasing decisions, potentially limiting irrational spending (Mr. Jose Mathews, 2025). An alternative approach to traditional fashion consumption is collaborative consumption, otherwise known as CC, where different business models are based on sharing instead of owing. This can involve renting, trading, swapping and lending clothes, as well as purchasing second-hand items (Shahid Nawaz, Yun Jiang, Muhammad Zahid Nawaz, Syeda Farzana Manzoor, and Ruixue Zhang, 2021). This new mindset encouraged consumers to spend less and consider alternatives such as second-hand clothing, rather than repeatedly buying new clothes and throwing old ones out. However,

mindfulness may only be effective depending on the individual's willingness and ability to maintain this awareness when faced with strong emotional or social pressures.

Individual differences can influence consumers' frequency to engage in impulse buying, with personality traits and long-term goals affecting how individuals respond to purchasing situations (Khalid Mahmood, Muhammad Amir Rashid, Ghulam Hussain, 2019). Traits are relatively stable characteristics that distinguish one individual from another (Ceyda Aysuna Turkyilmaza, Sakir Erdema, Aypar Uslu, 2014). For example, Wells et al. (2011) found that personality traits influenced online impulse buying. Individuals who were more outgoing, open to new experiences and agreeable were more likely to engage in impulse buying, while those that were more conscientious were less likely to do so (Ceyda Aysuna Turkyilmaza, Sakir Erdema, Aypar Uslu, 2014). This suggests that consumers' susceptibility to impulse buying varies according to individual characteristics, meaning that exposure to persuasive marketing or social influence does not necessarily result in irrational spending for everyone. Therefore, individual differences in personality, values and long-term goals may act as a limitation on the extent to which social influence and impulse-buying triggers contribute to irrational spending.

Discussion

The findings of this research suggest that impulse buying and social influence contribute to irrational spending in the fashion industry to a great extent, but their influence is not the same for every consumer. Given all the factors that were analyzed, such as social media, influencer marketing, social comparison, FOMO and psychological marketing strategies were repeatedly associated with impulsive purchasing and excessive clothing consumption. However, the evidence also demonstrates that consumers do not respond equally to these influences. Therefore, the relationship is dependent on the individual consumer's ability to regulate their behaviour. The findings particularly highlight the importance of social influence in shaping fashion consumption. Consumers may purchase clothing not only to satisfy functional needs but also to maintain social status, gain acceptance and construct a desirable identity. Veblen's theory of conspicuous consumption and Goffman's Self-Presentation Theory help explain why visible fashion products can have social value beyond their practical purpose. This is further supported by the survey findings, where approximately 45% of respondents indicated that social expectations, such as attending events or interacting with peers, contributed to purchasing new clothing. However, my findings contradict that social influence affects all consumers equally. This is supported by a finding shown that 60% of respondents disagreed that they purchased clothing promoted by celebrities or influencers. This contrasts with research suggesting that influencer marketing has a strong influence on Generation Z and indicates that consumers may increasingly recognise the persuasive intentions behind social media marketing.

The research also demonstrates that social influence and impulse buying are often interconnected. Social comparison can create feelings of envy, while FOMO and conformity can create pressure to purchase products before trends disappear. These psychological processes may reduce the time consumers spend evaluating alternatives and decide on financial consequences.

Marketing strategies can intensify these effects, using personalized algorithms repeatedly, exposing consumers to products that match their interests and potentially increasing familiarity through the Mere Exposure Effect. Similarly, anchoring, promotional pricing, decoy options and scarcity marketing can exploit predictable cognitive biases by directing consumers' attention towards perceived savings, urgency or the possibility of missing an opportunity. These strategies therefore demonstrate how businesses can transform existing psychological and social motivations into purchasing behaviour.

Nevertheless, the findings suggest that exposure to these strategies does not automatically result in irrational spending. Financial literacy can help consumers understand the consequences of unnecessary purchases, while self-control can reduce the likelihood of acting on emotional impulses. This is particularly important in relation to FOMO, as consumers with stronger self-control may experience the same social pressure without responding through impulse buying. Similarly, Kahneman's dual-process theory suggests that although System 1 may generate an immediate emotional desire to purchase, System 2 can intervene by encouraging consumers to evaluate whether the purchase is actually necessary. Mindful consumption can operate in a similar way by encouraging individuals to become more aware of their needs and purchasing decisions. Sustainable and collaborative consumption also provide alternatives to repeatedly purchasing new clothing, suggesting that greater awareness can shift consumers towards more eco-friendly forms of consumption.

Individual differences further explain why the influence of impulse buying and social influence varies between consumers. For example, research by Wells et al. (2011) suggests that individuals who are more outgoing, open to new experiences and agreeable may be more susceptible to online impulse buying, whereas reserved individuals are less likely to engage in such behaviour. This reinforces the argument that marketing and social influence should not be viewed as direct causes of irrational spending, but rather as factors that increase the probability of such behaviour among susceptible consumers.

Overall, the evidence provides support for the research question, but indicates that impulse buying and social influence lead to irrational spending to a significant extent. However, these relationships should not be interpreted as evidence of causation, as correlation does not necessarily imply causation. Furthermore, individual differences mean that consumers do not all respond to these influences in the same way. Therefore, irrational spending in the fashion industry is best understood as the result of an interaction between external influences and individual psychological characteristics. This suggests that reducing irrational spending in the fashion industry requires not only greater awareness of persuasive marketing techniques, but also the development of stronger self-awareness and self-control to resist external and emotional awareness.

Works Cited

Rosely, N. and Syed Ali, S. F. (2023) ‘A Challenge Towards Sustainable Fashion Consumption: Fast Fashion and Impulsive Purchase Behaviour’.

Shah, S. S. and Asghar, Z. (2023) ‘Dynamics of Social Influence on Consumption Choices: A Social Network Representation’.

Jindal, M., Suri, R., Kaur, K., Kansra, T. and Sahoo, A. (2025) ‘Understanding the Drivers of Clothing Overconsumption among Teenagers in the Era of Fast Fashion’.

Gigauri, Z. (2024) ‘Marketing and Social Influence of Fashion on Clothing Purchase Behavior of Consumers’.

Murhundezi, T. J. (2024) ‘Irrational vs Rational Spending Within the Fashion Industry: Are Generation Z Spending Irrationally on Fashion Products?’

Wang, Y., Pan, J., Xu, Y., Luo, J. and Wu, Y. (2022) ‘The Determinants of Impulsive Buying Behavior in Electronic Commerce’.

Le, M. T. H., Pham, H. N. M., Nguyen, L. T. T., Le, T. T. X. and Nguyen, S. T. M. (2024) ‘Augmenting TikTok Shop Adoption: The Mediating Role of Key Opinion Consumers (KOCs) in Live Streaming, Video Content, and Artificial Intelligence (AI) Recommendations’.

Kim, W. and Ha-Brookshire, J. E. (2026) ‘What Makes It a Dupe? Exploring Consumer Understanding of Legally Copied Products in the Fashion Industry’.

Reagan, F., Tejasukmana, K. A. and Toby, M. (2025) ‘The Fine Line Between Inspiration and Imitation: Delving into the Intersection of Fashion Dupes, Law and Society’.

Zafran, M. and Irfan, A. (2025) ‘If I Can’t Afford It, I Will Buy Dupes: Influence of Dupe Culture on Consumer Perceived Value of Luxury Products’.

Yeo, S. F., Lim, K. B., Leong, Y. J., Tan, C. L. and Leong, Y. C. I. (2020) ‘Effects of Social Media Advertising on Consumers’ Online Purchase Intentions’.

Wu, M. (2026) ‘Luxury or Dupe: How Social Media Shapes International Students’ Perceptions of Luxury and Dupe Culture’.

Jakesch, M. and Carbon, C. C. (2012) ‘The Mere Exposure Effect in the Domain of Haptics’.

Yeo, S. F., Lim, K. B., Leong, Y. J., Tan, C. L. and Leong, Y. C. I. (2020) ‘Effects of Social Media Advertising on Consumers’ Online Purchase Intentions’.

Segijn, C. M. and van Ooijen, I. (2022) ‘Differences in Consumer Knowledge and Perceptions of Personalized Advertising: Comparing Online Behavioural Advertising and Synced Advertising’.

Lin, Y. E. C., Chang, F. C., Huang, T. F., Chiang, S. C., Chiu, C. H., Chen, P. H., Miao, N. F. and Chuang, H. Y. (2026) ‘Effects of Social Media Sponsorship Disclosure on Adolescents’ Advertising Literacy and Purchase Intention’.

Zong, Y. and Guo, X. (2022) ‘An Experimental Study on Anchoring Effect of Consumers’ Price Judgment Based on Consumers’ Experiencing Scenes’.

The Launchpad Incubator (2025) ‘Psychological Pricing’.

The Launchpad Incubator (2025) ‘Pricing Strategy’.

Weissman, F. (2025) ‘Loss Aversion in Marketing and Commercials: A Multifaceted Analysis’.
Gupta, S. (2020) ‘The psychological effects of perceived scarcity in a retail setting and its impact on consumer buyer behavior’.
Zixian Tang (2026) ‘Scarcity Appeals in Advertising: Psychological and Social Mechanism of Consumer Buying Behaviour’.
Anisa, N. A., Arifin, S., Setyowati, L., Hidayah, N. and Megasari, A. D. (2020) ‘Financial Literacy on Impulsive Buying Behavior in Y Generation’.
Panjaitan, P., Azazi, A., Fitriana, A., Ristyawan, M. R. and Syahputri, A. (2026) ‘Financial Literacy as a Moderator of the Effects of Fintech Payment, Income, and Hedonic Lifestyle on Impulse Buying among Generation Z’.
Wenni Wahyuandari, Latifatul Isro’iyah, Priska Rosita Damayanti (2024) ‘How financial literacy and lifestyle shape impulsive buying habits in e-commerce’.
Wang, Y., Pan, J., Xu, Y., Luo, J. and Wu, Y. (2022) ‘The Determinants of Impulsive Buying Behavior in Electronic Commerce’.
Moayery, M., Narvaiza Cantín, L. and Gibaja Martíns, J. J. (2017) ‘How Does Self-Control Operate? A Focus on Impulse Buying’.
Wijaya, A. G., Wu, W. and Widiatmo, G. (2025) ‘Determinants of Consumer Purchase Intentions on Video Commerce: A Dual-Process Perspective’.
Nawaz, S., Jiang, Y., Nawaz, M. Z., Manzoor, S. F. and Zhang, R. (2021) ‘Mindful Consumption, Ego-Involvement, and Social Norms Impact on Buying SHC: Role of Platform Trust and Impulsive Buying Tendency’.
Mathews, J. (2025) ‘Mediational Mechanisms of Mindful Consumption: A Conceptual Analysis’.
Mahmood, K., Rashid, M. A. and Hussain, G. (2019) ‘Personality and Post-Purchase Consumer Regret Experienced After Impulse Buying: A Cross-Theoretical Approach with Individual Differences Moderator’.
Ceyda Aysuna Turkyilmaza, Sakir Erdema, Aypar Uslua (2015) ‘The Effects of Personality Traits and Website Quality on Online Impulse Buying’.
Allendifa, A., Nasution, M, D, T, P., Fadli. (2026). Pleasure-Seeking Or Fear-Driven? A Conceptual Examination Of Hedonic Motivation And FOMO As Catalysts Of Impulsive Buying In E-Commerce

Supply Chain Risk Management: Case Studies and Strategies for Resilience
By Aarav Agarwal

Introduction

Modern supply chains are highly interconnected, meaning a disruption in one area can quickly affect the rest of the network. Risks involving suppliers, technology, and geopolitics can create major delays, shortages, and financial losses. This paper uses the SCOR model and three real-world case studies to examine how these risks develop and interact. It then applies those lessons to identify ways Unilever can improve the resilience of Dove's palm-oil supply chain.

The SCOR Model

"The SCOR model was developed by the Supply Chain Council (SCC) in 1996 but is now maintained by APICS."[54] "The SCOR model is designed to evaluate your supply chain for effectiveness and efficiency of sales and operational planning."[1] It streamlines the language used to describe supply chain[1] into 5 key aspects: plan, procure, make, distribute and support. Within a SCOR based evaluation, organizations can determine the costs for each phase of the supply chain and assess operational efficiency while also identifying potential areas of improvement.[54] Conducting this analysis also clarifies how the different aspects of the supply chain are interconnected, ultimately providing a stronger business strategy.[54]

The planning stage focuses on balancing supply chain resources with efficiency and business goals.[54] This includes forecasting demand, managing inventory levels, coordinating transportation, maintaining assets, and ensuring regulatory compliance.[54] Strong planning ensures companies have enough resources to meet demand without overproducing or wasting inventory.[54] It also establishes communication systems and policies that guide overall supply chain performance.[1] [54]

Procurement is the process of obtaining the materials and resources to be able to manufacture products.[55] This includes managing supplier networks, negotiating contracts, evaluating supplier performance[54], and ensuring purchases follow the "Five Rights of Procurement"[56] (right quality, quantity, condition, place, and time). Sourcing is a major risk as many materials are concentrated in certain areas of the world.[57] For example, China controls a large percentage of the world's raw materials and geopolitical issues like Trump's increased tariffs can heavily impact, by decreasing, the amount China exports to the US.

The make stage focuses on transforming the raw materials into the actual products and making sure they are ready to be sold.[55] This includes production scheduling, equipment and facility management[1], quality control, and packaging. Efficient manufacturing processes are critical because disruptions such as labor shortages, equipment failures, or production delays can impact the entire supply chain.

Delivery involves all logistics required to move finished goods to customers.[55] This includes order processing, warehousing, transportation, and distribution management.[54] Products may move between distribution centers (DC to DC) or directly to customers (DC to Home). This

stage also includes inventory handling, quality inspections, customs processes, and documentation such as invoices and shipping records.[54] Because delivery relies heavily on transportation networks, it is vulnerable to delays caused by traffic or infrastructure issues.

Finally, the support stage focuses on tracking and tracing the movement of freight. This includes tracking referring to where a shipment currently is, and tracing referring to where it originated. This stage also handles customer requests and returns, ensuring the supply chain continues to function smoothly even after goods reach their destination.[1]

Supply Chain Risk

Supply chain risk refers to anything that can disrupt the flow of materials, products, information, or money throughout a supply chain [7]. These disruptions can come from inside a company or from external factors that are outside of the company's control [5]. Oftentimes the reasons for such risks are posed by natural disasters, political conflicts, etc… [2] However, no matter where the problem starts, the effects can spread throughout the supply chain causing delays, increasing costs and making it harder for companies to meet consumer demand [7, 5].

The goal of supply chain risk management is to identify potential problems before they become major disruptions and to reduce their impact if they occur [2]. The process generally involves four key steps: identification, assessment, mitigation, and monitoring [2]. Identification focuses on recognizing possible risks, assessment is where the chance of the risks occurring and their impact, mitigation involves creating plans to reduce or prevent those risks, and monitoring ensures that any issues are addressed and do not continue to affect operations [2]. Together, these steps help companies respond more effectively when disruptions occur.

A strong risk management strategy helps businesses become more resilient and better prepared for unexpected challenges [2]. “With proactive preparedness, companies can avoid or minimize disruptions, reduce costs, improve quality and enhance customer satisfaction” [2]. As supply chains continue to become more global and interconnected, managing risk has become increasingly important for long-term success [2].

Types of Supply Chain Risk

Supply chain risks can come from many different sources. The 5 major categories of risk are operational, geopolitical, environmental, technology, and supplier risks [4].
Geopolitical risk involves disruptions caused by political events and changing government regulations [5]. These risks can have a major impact on companies that rely on international suppliers and manufacturers [4]. Sanctions can suddenly limit access to important suppliers or markets, forcing businesses to find alternatives on short notice [5]. One example is the U.S.-China trade war, which caused many companies to rethink where they sourced their products to avoid higher costs from tariffs [4].

Environmental risk comes from natural disasters and extreme weather events that are outside a company's control [4]. Disasters can damage facilities, delay transportation, and interrupt production [4]. “The 2011 earthquake and tsunami in Japan caused severe disruptions in the

automotive and electronics industries" [4]. As severe weather events become more common, environmental risk has become an even greater concern for businesses.

Technology risk has become increasingly important as supply chains rely more heavily on digital systems [7]. Cyberattacks and software failures can stop production, delay shipments, and expose sensitive company data [2]. Even a temporary outage in a tracking or logistics system can disrupt operations and create delays throughout the supply chain [6]. As businesses continue to adopt new technologies, protecting digital infrastructure has become a key part of risk management [7].

Supplier risk occurs when problems at a supplier affect the rest of the supply chain [4]. Financial instability, production issues, quality concerns, or supplier bankruptcy can all create major disruptions [4, 5]. In some cases, a company may depend on a supplier for a critical component with no immediate alternative available [6]. This risk becomes even more serious when suppliers are concentrated in a single region, making companies vulnerable if a local disruption occurs [6, 5].

Although these categories are often discussed separately, real-world supply chain disruptions usually involve several risks at once [3]. A natural disaster may impact suppliers [5], a geopolitical conflict may create transportation delays [6], or a cyberattack may shut down critical operations [2]. The case studies that follow demonstrate how these different risks can combine to create large-scale disruptions across global supply chains.

Case Study 1: Technology Risk

On June 27, 2017, Maersk, the world's largest container shipping company, experienced one of the most destructive cyberattacks ever [8]. Maersk is responsible for about 20% of global container shipping so when its system started to fail, the effects were seen not only in the shipping process, but the entire global supply chain [9]. The attack was caused by malware NotPetya which entered their network through a computer in Odessa, Ukraine. [8] A Maersk finance executive requested that the company's IT staff install M.E.Doc which is an accounting software widely used in Ukraine for tax filing and business transactions [8]. The software became compromised by hackers and the malware was spread to companies who utilized the software [8]. The malware used a hacking tool called EternalBlue which had originally been developed by the US National Security Agency and then later leaked online [8]. The malware also stole login credentials from computers and used them to move through the network [8]. Since Maersk's systems were highly interconnected, the attack spread across the company in just a few hours [8]. In total, 45,000 computers and 4,000 servers were fully infected [10]. This included all 150 of the company's domain controllers which are critical systems that manage access to networks and this basically shut down Maersk's entire global IT infrastructure [8]. Maersk operates 76 port terminals and 17 of these could not process shipments and had to stop operations as the booking and tracking systems were offline [8, 10]. With such a large share of the global shipping market, the disruption created serious delays across international trade [8].The company estimated losses of $200-$300 millions due to the attack [12] and reported $264 million total consolidated quarterly

loss [11]. The disruptions lasted roughly two weeks while the company rebuilt its entire IT system [11].

The recovery process became extremely difficult because the company did not have a backup of its domain controllers, yet one server in a Ghana office had been offline due to a power outage and therefore still contained a clean copy of the company's domain controller data [8]. But the damages were not limited to Maersk. We saw how FedEx's TNT Express division lost about $400 million, while pharmaceutical company Merck lost around $870 million [8]. In total, global damages from the attack were estimated to be around $10 billion [8]. This incident highlights several important lessons about technology risk. First, third-party software can create major vulnerabilities. Even if a company has strong internal security, it can still be exposed through software supplied by outside vendors. Second, Maersk's network structure made the situation worse because its systems were not segmented [8]. This meant that once the malware entered one computer, it could spread across the entire global network. Another lesson is the importance of backup systems and disaster recovery planning [9]. Companies must prepare for worst-case scenarios where entire systems fail at the same time. After the attack, Maersk rebuilt its cybersecurity strategy and invested heavily in stronger defenses, including multi-factor authentication, improved monitoring systems, and stricter access controls [8].

Overall, the Maersk case shows how technology risk can disrupt not just a single company but the entire global supply chain. A cyberattack that began with one piece of compromised software ended up affecting international trade across more than 130 countries, demonstrating how dependent modern logistics systems are on secure digital infrastructure.

Case Study 2: Supplier Risk

"During 2020 to 2023, the world faced a major semiconductor shortage that affected 169 industries, ranging from automobiles to electronics" [20]. The shortage began during the COVID-19 pandemic as the key chip manufacturing plants in Asia and Europe were disrupted alongside the rapid demand increases for electronics [17]. As people shifted to working from home and entertaining themselves at home, the demand for laptops, routers and gaming systems grew largely [20]. However, the semiconductor plants were either operating at a reduced capacity or they were completely shut down [17]. This created a huge imbalance between the effective supply of semiconductors and the vast demand [18]. At the same time, Automakers expected their demand to significantly drop and therefore ordered less semiconductors [14]. But owners of automakers saw the demand for vehicles to recover much more quickly than expected [14]. The owners decided to reorder chips, but the factories had already shifted their production capacity to focus on consumer electronics as they were much more profitable [15]. This left owners waiting for a supply and therefore resulting in a mismatch between the large demand and little supply. The limited production triggered a global shortage that lasted for several years [20]. One of the reasons for the shortage becoming so severe is that many of the manufacturing plants were centered within specific regions or countries [14]. For example, about 75% of global chip production takes place in East Asia, especially in Taiwan, South Korea, and China [18]. A single company, Taiwan

Semiconductor Manufacturing Company (TSMC), accounted for more than half of global foundry revenue in 2020 [21]. Since many major US technology companies depended heavily on these manufacturers, a very small number of suppliers controlled a large portion of the global chip supply [14, 21]. This creates a major risk because if even one of these suppliers faces disruption, there are very few alternatives available. The economic impact of the shortage was nothing short of vast. The automotive industry alone lost about $210 billion in revenue in 2021 because companies were forced to slow down or stop production due to a lack of chips [17]. Across 169 industries, the shortage disrupted production and left consumers waiting longer for products such as gaming consoles [16].

To combat the shortage, the CHIPS and Science Act of 2022 provided about $52 billion in funding to increase domestic chip production [15]. This led to major investments from companies like TSMC, Samsung, and Intel to build new manufacturing facilities [13]. Europe and Japan also pursued initiatives to expand semiconductor manufacturing capacity and strengthen their domestic supply chains [20]. This shortage highlights several important lessons about supplier risk in global supply chains. Firstly, relying too heavily on a small number of suppliers and especially when they are concentrated in one region can create serious vulnerabilities. Second, many companies did not fully understand their supply chains beyond their immediate suppliers, which made it difficult to respond quickly when problems arose. As a result, businesses are now working to map out their full supply networks and identify potential risks earlier. For example, companies are looking past their direct suppliers and looking into similar dependencies farther upstream. Since ASML makes the lithography machines that are required to produce advanced semiconductors, a lot of chipmakers depend on this equipment to keep production running and any disruption in the supply of these machines can still slow or stop production [14, 20]. By recognizing these dependencies, businesses can better identify possible risks and prepare for them in advance. Another key lesson is the importance of inventory strategy. Many companies in the automotive industry used just-in-time systems to reduce costs, but this approach failed when supply became limited and lead times increased [14]. JIT (just-in-time) systems are when companies receive materials and produce goods only when they are needed rather than storing large inventories [14]. Keeping some level of backup inventory for critical components can help prevent major disruptions [14]. Yet with this comes a major tradeoff between keeping inventory low to save costs and holding extra buffer stock to make sure operations can continue during uncertain times.

Overall, the semiconductor shortage demonstrates how supplier risk can impact not just individual companies but the entire global economy. A combination of high demand, limited supply, and overdependence on a small group of manufacturers led to widespread disruptions across multiple industries. This event showed that supply chains must be designed not only for efficiency, but also for resilience in the face of unexpected shocks.

Case Study 3: Geopolitical Risk

On February 24, 2022, Russia invaded Ukraine and triggered one of the worst disruptions

the world has seen to supply chains [23]. The conflict did not stay confined to the two countries, but affected many more [23] as Russia and Ukraine are major exporters of food [22], energy [26] and industrial materials [29]. “Russia and Ukraine together accounted for 30 per cent of wheat exports, 25 per cent of barley exports and 15 per cent of maize exports globally” [22]. When the fighting broke out, millions of tons of grains were simply left stranded and left to rot [23]. Alongside this, the USA and EU set sanctions on Russia that restricted Russia’s access to parts of the global financial system and made some international transactions more difficult [26]. Russia responded to these sanctions by restricting some of its own exports, such as energy [26]. The first major impact was on global food markets. Ukraine had supplied food to about 400 million people worldwide [23] and the blocked ports resulted in the price of wheat spiking sharply [22]. However in July 2022, the UN and Turkey helped broker the Black Sea Grain Initiative which allowed Ukrainian exports to resume through safe shipping corridors [24]. Over the next year, about 33 million tons of grains and other foods were exported through this initiative to about 45 countries [24]. Then again, Russia withdrew from the agreement [25] and took the world back to square 1. The second major impact was on energy supply. Many European countries depended on Russian oil and natural gas [26]. When sanctions and supply cuts were put into place, these cities had to seek energy from other suppliers such as the US, Norway and Qatar [27]. This however led to higher costs and long transportation routes [26]. The third impact was on industrial supply chains. Ukraine produced about 45 to 54 percent of the world's semiconductor grade neon through 2 companies [28]. Due to the proximity of the war, both of these companies had to shut down and the world saw a huge decrease in the available semiconductor grade neon [28]. At the same time, Russia produced 40% of the global palladium and also exported a lot of neon gas [29]. Neon was used in the lasers in chip lithography and palladium was used in sensor and memory chips [29]. So the shortage of such rare metals and gasses forced chipmakers to search for alternative sources [29].

This case highlighted the issue with relying on a small number of suppliers, especially when they are concentrated in one region. Further, manufacturers should look beyond suppliers and focus on the transportation routes that could possibly be affected by conflicts. Third, sanctions can disrupt trade for goods that are not even directly due to restrictions on banking and shipping that can make it harder to complete transactions. Now, many companies are trying to shift production and sourcing to more stable regions even if it costs more. Overall, the Russia-Ukraine war shows how geopolitical risk can have immediate and widespread effects on global supply chains. A conflict between two countries was able to affect food markets, energy prices and semiconductor production all at once. This just demonstrates how interconnected supply chains and global politics are.

How Companies Manage Risk

Unilever's Dove Soap Supply Chain

Plan: The planning stage is where Unilever predicts the customer demand and uses that to plan its production. Since Unilever processes over 25 million customer orders every year, it

needs a strong planning system to keep everything running smoothly across its 280 factories, 500 warehouses, and more than 52,000 suppliers in over 150 countries [35]. Its systems are heavily interconnected so that the company can easily and better manage inventory and operations [35]. For Dove soap, the four factories that produce Dove bars share real-time data that is used to adjust the production quickly based on the demand [32].

Procure: For Dove soap, the most important raw material is palm oil [30]. Unilever buys around 1 million tons of palm oil and related products each year which makes them one of the largest buyers in the world. Most of this palm oil comes from plantations in Indonesia and Malaysia where it becomes ready to ship to the Dove factories [30, 34]. To have more control over its supply, Unilever invested €280 million into its own processing plant in Indonesia [30]. Before, Unilever would have had to have bought from some middlemen but now they can buy directly from the mills [30]. For example, it allows them to track the nearby farmers [30].

Make: Once the raw materials arrive at the factory, they are turned into soap through a process called saponification [34]. Saponification is where the oils and fats react with an alkali and allow the resources to be turned into soap [34]. After that, the soap goes through quality testing, packaging, and preparation for shipping [32, 34]. Dove soap is made in several countries, including the United States, Canada, Germany, India, Brazil, Indonesia, and Thailand. Even though production takes place in different locations, Unilever follows the same quality standards at every factory to maintain consistency [32].

Distribute: After production, Dove soap is shipped through Unilever's distribution network before reaching stores [31]. Products are usually transported by truck from factories to distribution centers [31]. Here they are stored and prepared for retailers. Unilever operates about 500 warehouses and distribution centers around the world [33]. The company has also expanded its direct dispatch system so that products can go straight from factories to retailers without stopping at a warehouse [31]. This method increased from 8% of deliveries in 2018 to 16% today, with a goal of reaching over 25% by 2026 [31]. One of Unilever's largest distribution centers in Mannheim, Germany, can store over 36,000 pallets and has the capacity to ship over 60 million units annually [32].

Support: The support stage includes customer service, returns, and tracking products throughout the supply chain. Unilever works closely with retailers to understand customer demand and keep Dove products in stock

Managing Supplier Risk

The most significant supplier risk that Unilever faces in its Dove supply chain is its dependence on Palm oil . Currently they source it mainly from Indonesia and Malaysia which total up to account for over 85% of the world’s palm oil production [36]. To manage this, Unilever can diversify their suppliers. The most structured approach would be implementing a supplier scoring system [40]. This risk assessment model ranks each supplier by categories (say attributes) such as quality, financial stability, delivery reliability, country risk, sustainability, traceability and capacity [41]. Based on their overall score, suppliers could then be grouped into different risk levels

[41]. Low-risk suppliers would continue to be monitored regularly [41], medium-risk suppliers would work with Unilever on improvement plans, and high-risk suppliers would be considered for replacement or reduced sourcing. Further, the weights given to each attribute could be dynamically adjusted based on the current situation, and supplier scores could be updated at a regular frequency. For example, the country risk could be weighted more as political instability and natural disaster rate can severely affect the supply chain and the different weights would be updated each quarter. This allows Unilever to identify potential problems early and make sourcing decisions before disruptions affect Dove production. For example, in the current situation, Unilever could start ordering more from the suppliers in Thailand and Papua New Guinea [37]. In addition, strategic relationships could be developed with suppliers in places like Nigeria and Columbia [37].

A second approach would be finding strong substitutes for palm oil [39]. In terms of substitutes, end product quality and cost would have to be considered [39]. "Hard butters like mango butter, shea butter and cocoa butter can be used to replace palm oil but they add hardness and moisturizing properties to the bars" [38]. While alternatives like soybean or sunflower are viable, palm oil remains as the most land efficient oil crop as producing the same amount of plant-derived oil from soybean or sunflower would require roughly four to ten times more land [36]. These factors would have to be considered against the risk of major disruption of palm oil.

Managing Technological Risk

While Unilever faces several technology risks across its Dove supply chain, the most critical is data vulnerability. Unilever's supply chain runs on a heavily digital focused infrastructure so that means data flowing through this system is crucial for both operations and attackers [42]. Every component and raw material is tracked digitally from the source to the customer so this same depth of data is what makes it a lucrative target for hackers [42]. The very first step in protecting the data is understanding where it lives and how it moves as much of the risk profile varies depending on the current status [44].

The first method of protection is data encryption. Encryption works by scrambling the sensitive information into unreadable coded language and can only be uncoded by authorized users who have the key to access it [47]. This type of encryption is often applied to data that is at rest [43]. For data that is in transit, Unilever must use end-to-end encryption which is similar to encryption where the data is scrambled and can only be accessed using keys. However, with E2EE, the key lies only with the sender and the receiver of the data whereas in encryption anyone in the middle, such as the IT team, could hold the key [42]. They can also use technologies such as TLS, HTTPS, VPNs to ensure that even if a hacker were to intercept the data or gain access to the systems, they would not be able to read or use it [44].

The second method is data replication and backups. From our Maersk case study, we saw the only reason Maersk was able to rebuild was because one server in Ghana appeared to be offline during the attack and that servers still had a clean copy of the company's data. Backups need to be stored across multiple separate locations around the world so a single attack can't

wipe everything at once [45]. A clear Recovery Time Objective should also be set which would define exactly how long recovery can take before it causes serious damage [46]. To be even more secure, regular simulations should be held to make sure the backup infrastructure can actually hit the time target [46].

The third method is strengthening the overall infrastructure. Encryption and backups help contain damage once an attack gets through, but the main goal is to even stop it from getting in at all. This could mean locking down third-party vendor access as any weak link in a supplier's system can be treated as an entry point into Unilever's network. Also, any vendor with access to Unilever's systems needs to meet the same cybersecurity standards Unilever holds internally. With these numerous methods and efficient use of systems, Unilever can advantage their technology to become a strength rather than a weakness.

Managing Geopolitical Risk

The geopolitical risk in Unilever's dove soap supply chain is closely related to where the palm oil is sourced from [49]. Palm oil is primarily transported in liquid bulk form using specialized chemical or food-grade tankers [48]. It usually departs from ports in Sumatra and Java and moves through major shipping lanes into European and North American markets. This then leads to 2 major types of possible disruptions: policy decisions by the governments on foreign countries and physical disruptions to the shipping routes [49, 51].

The first strategy is scenario planning [53]. Unilever should model different scenarios in advance and build a response plan for each. This planning could include securing extra suppliers and balancing inventory [52]. The ability to switch routes and suppliers quickly, and to anticipate disruptions is what can create a resilient supply chain [52, 53]. Further, if political issues from the government were to arise, Unilever should watch for any signs of initial trouble and see how that would affect the suppliers, shipments and materials [51]. Then they should identify the most vulnerable links, and then turn to alternatives [51]. The biggest difference between reactive and resilient companies is how quickly they can put their plans into action. That's why scenario plans shouldn't just sit, they need to be constantly updated and revised [50].

The second strategy would be to understand that map the existing distribution routes and then identify alternatives before disruptions arise [53]. Palm oil requires a coordinated logistics network to move from production facilities to markets around the world [48]. Musim Mas, for example, manages multiple stages of this process and distributes its products through more than 200 ports worldwide [48]. Unilever should map exactly which ports its palm oil flows through and which alternative ports exist. Unilever also needs to understand the entire transportation network which includes the freight forwarders, shipping lanes, and bottlenecks [48, 53]. During the 2024 Red Sea crisis, for example, Home Depot worked with its logistics carriers to reroute shipments and reduce delays [51]. That kind of response is only possible if alternative routes have already been identified before a disruption occurs. Unilever also needs to be cautious that their alternatives don't result in a financial problem as well [52].

Conclusion

Supply chains are largely interconnected systems so that when a supplier, technology platform or region is disrupted, the effects quickly spread throughout the entire business. The Maersk cyberattack, semiconductor shortage and Russia-Ukraine war each show how technology, supplier and geopolitical risks can impact companies across multiple industries and countries. These case studies highlight why businesses need to understand not only how every part of their supply chain works, but also where their biggest vulnerabilities are and how disruptions can affect the entire network.

For Unilever, keeping the Dove soap supply chain running smoothly means being prepared before problems happen. Resilience is not achieved through a single back up supplier or cybersecurity control. Supplier risk can be reduced by evaluating suppliers regularly and avoiding dependence on a single source. Technology risk can be minimized through stronger cybersecurity measures, encryption, and reliable data backups. Geopolitical risk can be managed by planning for different scenarios and using alternative suppliers or transportation routes. While companies cannot predict every disruption, they can build supply chains that identify risks early, respond quickly, and recover more effectively. In today's global economy, the companies that balance efficiency with resilience will be the ones best equipped to handle future challenges.

Works Cited

1. CIO. “What Is SCOR? A Model for Improving Supply Chain Management.” https://www.cio.com/article/222381/what-is-scor-a-model-for-improving-supply-chain-management.html Accessed 24 July 2026.
2. IBM. “What Is Supply Chain Risk Management?.” https://www.ibm.com/think/topics/supply-chain-risk-management Accessed 24 July 2026.
3. ASCM. “Supply Chain Risk Management.” https://www.ascm.org/topics/supply-chain-risk-management/ Accessed 24 July 2026.
4. Inbound Logistics. “Supply Chain Risk Management.” https://www.inboundlogistics.com/articles/supply-chain-risk-management/ Accessed 24 July 2026.
5. Precoro. “7 Basic Types of Supply Chain Risks.” https://precoro.com/blog/7-basic-types-of-supply-chain-risks/ Accessed 24 July 2026.
6. Rasmussen University. “What Is Supply Chain Risk Management?.” https://www.rasmussen.edu/degrees/business/blog/what-is-supply-chain-risk-management/ Accessed 24 July 2026.
7. NetSuite. “Supply Chain Risks.” https://www.netsuite.com/portal/resource/articles/inventory-management/supply-chain-risks.shtml Accessed 24 July 2026.
8. Wired. “The Untold Story of NotPetya, the Most Devastating Cyberattack in History.” https://www.wired.com/story/notpetya-cyberattack-ukraine-russia-code-crashed-the-world/ Accessed 24 July 2026.
9. SOS Intelligence. “Maersk's Response to NotPetya: How Cybersecurity Best Practices Mitigated a Major Cyberattack.” https://sosintel.co.uk/case-study-maersks-response-to-notpetya-how-cybersecurity-best-practices-mitigated-a-major-cyberattack/ Accessed 24 July 2026.
10. LRQA. “NotPetya Ransomware Attack on Maersk: Key Learnings.” https://www.lrqa.com/en/insights/articles/notpetya-ransomware-attack-on-maersk-key-learnings/ Accessed 24 July 2026.
11. Los Angeles Times. “Maersk Cyberattack.” https://www.latimes.com/business/la-fi-maersk-cyberattack-20170817-story.html Accessed 24 July 2026.
12. Maersk. “Investor Reports.” https://www.maersk.com/investors/reports Accessed 24 July 2026.
13. Substack. “An Investment Boom Triggered.” https://barberd.substack.com/p/an-investment-boom-triggered-bb7 Accessed 24 July 2026.

14. Interface. “Understanding the Global Chip Shortages.” https://www.interface-eu.org/storage/archive/files/understanding_the_global_chip_shortages.pdf Accessed 24 July 2026.
15. Congressional Research Service. “CRS Report R47523.” https://www.congress.gov/crs-product/R47523 Accessed 24 July 2026.
16. Yahoo Finance. “These Industries Are Hit Hardest by the Global Chip Shortage.” https://finance.yahoo.com/news/these-industries-are-hit-hardest-by-the-global-chip-shortage-122854251.html Accessed 24 July 2026.
17. Schneider Downs. “Semiconductor Chip Shortages: Continual Impact.” https://schneiderdowns.com/our-thoughts-on/semiconductor-chip-shortages-continual-impact/ Accessed 24 July 2026.
18. Semiconductor Industry Association. “2021 State of the U.S. Semiconductor Industry.” https://www.semiconductors.org/wp-content/uploads/2021/09/2021-SIA-State-of-the-Industry-Report.pdf Accessed 24 July 2026.
19. Journal of Global Business and Community. “Addressing America's Semiconductor Dependency on East Asia.” https://jgbc.scholasticahq.com/article/134052-addressing-america-s-semiconductor-dependency-on-east-asia Accessed 24 July 2026.
20. Wikipedia. “2020-2023 Global Chip Shortage.” https://en.wikipedia.org/wiki/2020%E2%80%932023_global_chip_shortage Accessed 24 July 2026.
21. Wccftech. “TSMC Earned $1,634 Revenue per Wafer in 2020 with a 54% Global Market Share.” https://wccftech.com/tsmc-earned-1634-revenue-wafer-in-2020-with-a-54-global-market-share/ Accessed 24 July 2026.
22. ZOiS. “The International Grain Trade and the War.” https://www.zois-berlin.de/en/publications/zois-spotlight/the-international-grain-trade-and-the-war Accessed 24 July 2026.
23. World Food Programme. “War in Ukraine: How a Humanitarian Tragedy Fed a Global Hunger Crisis.” https://www.wfp.org/stories/war-ukraine-how-humanitarian-tragedy-fed-global-hunger-crisis Accessed 24 July 2026.
24. United Nations. “Black Sea Grain Initiative: Achievements.” https://www.un.org/en/black-sea-grain-initiative/achievements Accessed 24 July 2026.
25. Council of the European Union. “Statement on Russia's Termination of the Black Sea Grain Initiative.” https://www.consilium.europa.eu/en/press/press-releases/2023/07/17/black-sea-grain-initiative-statement-by-the-high-representative-on-behalf-of-the-european-union-on-russia-s-termination-of-the-black-sea-grain-initiative/ Accessed 24 July 2026.

26. Federal Reserve Bank of St. Louis. “Reviewing the Impact of Energy Sanctions on Russia.” https://www.stlouisfed.org/on-the-economy/2023/feb/reviewing-impact-energy-sanctions-russia Accessed 24 July 2026.
27. Natural Gas Intelligence. “EU Seeks to End Russian Gas by 2027 with U.S. LNG Demand Cuts.” https://naturalgasintel.com/news/eu-seeks-to-end-russian-gas-by-2027-with-us-lng-demand-cuts/ Accessed 24 July 2026.
28. AOL / Reuters. “Ukraine Halts Half the World's Neon Output for Chips.” https://www.aol.com/news/exclusive-ukraine-halts-half-worlds-110634039.html Accessed 24 July 2026.
29. Resilinc. “Russia-Ukraine War: Semiconductor Value Chain Impact.” https://resilinc.ai/blog/russia-ukraine-war-semiconductor-value-chain-impact/ Accessed 24 July 2026.
30. Unilever. “Our Palm Oil Story.” https://www.unilever.com/sustainability/nature/addressing-deforestation-risk/palm-oil-story/ Accessed 24 July 2026.
31. Unilever. “Leading the Next Era of Logistics: Direct Dispatch for Efficiency.” https://www.unilever.com/news/news-search/2024/leading-the-next-era-of-logistics-direct-dispatch-for-efficiency/ Accessed 24 July 2026.
32. Unilever. “How We're Building a Leaner, More Agile Supply Chain.” https://www.unilever.com/news/news-search/2025/how-were-building-a-leaner-more-agile-supply-chain/ Accessed 24 July 2026.
33. SCM Insight. “Supply Chain Analysis of Unilever.” https://scminsight.com/supply-chain-analysis-of-unilever/ Accessed 24 July 2026.
34. Design Life-Cycle. “Dove Bar Soap.” https://www.designlife-cycle.com/dove-bar-soap Accessed 24 July 2026.
35. RFgen. “Unilever Supply Chain: Lessons Learned on Digital Transformation.” https://www.rfgen.com/blog/unilever-supply-chain-lessons-learned-on-digital-transformation/ Accessed 24 July 2026.
36. WWF UK. “8 Things to Know About Palm Oil.” https://www.wwf.org.uk/updates/8-things-know-about-palm-oil Accessed 24 July 2026.
37. TraceX. “Top Global Palm Oil Exporters 2025.” https://tracextech.com/top-global-palm-oil-exporters-2025/ Accessed 24 July 2026.
38. HBNO Bulk. “Palm Oil for Soap Making: Benefits, Substitutes, and Where to Buy.” https://www.hbnobulk.com/blogs/natural-essential-oils/palm-oil-for-soap-making-benefits-substitutes-and-where-to-buy?srsltid=AfmBOooB0cIafRLmGT58nSZYHTmwhRR46Mzz5UN93aldWOAvjuJYeqJg Accessed 24 July 2026.
39. Covalo. “6 Sustainable Alternatives to Palm Oil in Personal Care.” https://blog.covalo.com/personal-care/6-sustainable-alternatives-to-palm-oil-in-personal-care Accessed 24 July 2026.

40. Risk Ledger. “Supplier Risk Assessment Process.” https://riskledger.com/resources/supplier-risk-assessment-process Accessed 24 July 2026.
41. TUV SUD. “Guide to Supplier Risk Assessment Templates and Best Practices.” https://www.tuvsud.com/en-us/resource-centre/blogs/guide-to-supplier-risk-assessment-templates-and-best-practices Accessed 24 July 2026.
42. Manufacturing.net. “Securing Supply Chain Data with Advanced Encryption.” https://www.manufacturing.net/cybersecurity/blog/22912541/securing-supply-chain-data-with-advanced-encryption Accessed 24 July 2026.
43. Fit Small Business. “Encryption at Rest and in Transit.” https://fitsmallbusiness.com/encryption-at-rest-and-in-transit/ Accessed 24 July 2026.
44. TechTarget. “Best Practices to Secure Data at Rest, in Use and in Motion.” https://www.techtarget.com/searchsecurity/feature/Best-practices-to-secure-data-at-rest-in-use-and-in-motion Accessed 24 July 2026.
45. OpenText. “What Is Data Backup and Recovery?.” https://www.opentext.com/what-is/data-backup-and-recovery Accessed 24 July 2026.
46. Greater Birmingham Chambers. “Disaster Recovery in the Cyber Supply Chain: Is Your Business Prepared?.” https://www.greaterbirminghamchambers.com/resource/disaster-recovery-in-the-cyber-supply-chain-is-your-business-prepared.html Accessed 24 July 2026.
47. CISA. “Encrypt Business Data.” https://www.cisa.gov/audiences/small-and-medium-businesses/secure-your-business/encrypt-business-data Accessed 24 July 2026.
48. Musim Mas. “How Is Palm Oil Transported from Plantation to Supermarket Shelf?.” https://www.musimmas.com/resources/blogs/how-is-palm-oil-transported-from-plantation-to-supermarket-shelf/ Accessed 24 July 2026.
49. ASEAN Briefing. “Indonesia Bans the Export of Palm Oil, Impacting Global Food Prices.” https://www.aseanbriefing.com/news/indonesia-bans-the-export-of-palm-oil-impacting-global-food-prices/ Accessed 24 July 2026.
50. PwC. “Managing Geopolitical Risk.” https://www.pwc.com/us/en/america-in-motion/managing-geopolitical-risk.html Accessed 24 July 2026.
51. Supply Chain Digital. “Trade Tensions: How Geopolitics Rewires Supply Chains.” https://supplychaindigital.com/news/trade-tensions-how-geopolitics-rewires-supply-chains Accessed 24 July 2026.
52. IJIRCT. “Developing Comprehensive Contingency Plans for Supply Chains.” https://www.ijirct.org/viewPaper.php?paperId=2412013 Accessed 24 July 2026.
53. UST. “Mastering Uncertainty: Building Resilient Supply Chains in a Geopolitical World.” https://www.ust.com/en/insights/mastering-uncertainty-building-resilient-supply-chains-in-a-geopolitical-world Accessed 24 July 2026.

54. CIPS. “SCOR Model.” https://www.cips.org/intelligence-hub/procurement/kpis/scor-model Accessed 24 July 2026.
55. Wikipedia. “Supply Chain Operations Reference.” https://en.wikipedia.org/wiki/Supply_chain_operations_reference Accessed 24 July 2026.
56. Smart Procurement. “The Five Rights of Procurement.” https://smartprocurement.co.za/the-five-rights-of-procurement/ Accessed 24 July 2026.
57. Visual Capitalist. “How Much Control China Has Over the World’s Critical Minerals.” https://www.visualcapitalist.com/how-much-control-china-has-over-the-worlds-critical-minerals/ Accessed 24 July 2026.
58. Fortune. “China’s Rare Earth Processing Dominance / Tariff Retaliation.” https://fortune.com/2026/03/11/china-us-rare-earth-processing-critical-minerals/ Accessed 24 July 2026.

The Effect of the COVID-19 Pandemic on U.S Financial Companies By Pitt Sinonpat

Abstract

The study examines the effect of the COVID-19 pandemic on the stock returns and volatility of six major U.S financial institutions and the S&P 500 index. Using daily data from January 2015 to April 2026, the sample is divided into three periods: Pre-COVID-19, COVID-19, and Post-COVID-19 periods. The analysis uses two sample t-tests, regression analysis, and Granger causality test. The results show that the stock volatility increased significantly for all six financial institutions and the S&P 500 during the pandemic. Wells Fargo and Citibank experienced the largest increase in volatility. Volatility went down after the pandemic but it remained a bit higher than before the pandemic. In contrast, the effect of COVID-19 on stock returns was limited, with most financial institutions showing no significant change in average returns. The results also show that volatility spillover between financial institutions became weaker during and after the pandemic. However, financial firms remained as important sources of overall market volatility.

Keywords: Volatility, COVID-19, volatility spillover, financial companies, S&P 500.

Introduction

The COVID-19 pandemic created major uncertainty in financial markets. Business closure, travel restrictions, changes in consumer behavior, and concerns about the global economy caused large fluctuations in stock prices. Financial companies were especially important during the period because banks and other financial institutions are closely connected to global economic activities, investor confidence, and the overall stability of financial markets.

Previous studies generally show that the COVID-19 pandemic increased stock market volatility and reduced financial performance of many companies. However, the size of these effects differed across countries, sectors, and firms. Some studies found that government policies and lockdown measures helped rebuild investor confidence, while others showed that fears about COVID-19 infection, deaths, and economic conditions increased uncertainty and market volatility. The COVID-19 pandemic also affected the relationship between financial institutions, as changes in one company or sector could influence other parts of the market.

This study examines the effects of the COVID-19 pandemic on six major financial institutions in the United States: Bank of America, Citibank, Goldman Sachs, JP Morgan, Morgan Stanley, and Wells Fargo. The S&P 500 index is also included as a proxy for the overall US stock market. This study focuses on stock returns and volatility across three periods: Pre-COVID-19, COVID-19, and Post-COVID-19. Comparing these periods helps determine whether the effects of the COVID-19 pandemic on financial markets were temporary or continued after the pandemic period.

The study uses the two sample t-test and regression analysis to examine changes in mean volatility and stock returns. In addition, the Granger causality test is used to examine volatility

transmission between individual financial institutions and the S&P 500. This approach provides a broader understanding of how the pandemic affected major U.S. financial institutions and their relationship with the overall stock market.

This study contributes to the existing literature in several ways. First, unlike many previous studies that focus on either stock returns or volatility in one specific financial market, this paper examines both returns and volatility of six major US financial institutions with the S&P 500 index, providing a more comprehensive assessment of the pandemic's effects on the financial sector. This paper allows comparison of financial market behavior before and during the pandemic.

The remainder of this paper is organised as follows. Section 2 reviews previous studies on the effects of COVID-19 on financial markets and companies. Section 3 describes data and methods used in the analysis. Section 4 presents empirical findings on stock volatility, returns and volatility spillovers. Section 5 concludes the paper.

Literature review

Different methodologies have been used to examine the effect of the COVID-19 pandemic on financial markets and companies. These include OLS (Rababah et al., 2020), regression analysis (Puławska.,2021), panel data analysis (Achim et al., 2022; Utomo and Hanggraeni., 2021; Anh and Gan., 2020), ARCH model (Chaudhary et al., 2020), GARCH model (Khan et al., 2023; Endri et al., 2021; Onali., 2020), DCC-GARCH model (Batten et al., 2023), GJR-GARCH model (Priya and Sharma., 2023), MS-AR model (Baek et al., 2020), DY-spillover index (Aslam et al., 2021), Wavelet coherence analysis (Choi., 2020), Quantile regression analysis (Ullah et al., 2023), Time-series regression (Demir et al., 2021), TGARCH model (Lúcio and Caiado., 2022), Var model (Chowdhury et al., 2022), and Skewed t distribution of Hansen (Basuony et al., 2022).

The variables used in previous studies also varied widely. These include confirmed recoveries (Basuony et al., 2022), daily closing price of individual US company (Chowdhury et al., 2022), VIX (Onali., 2020), volatility of industry indices (Lúcio and Caiado., 2022; Priya and Sharma., 2023), volatility of Datastream Energy indices (Demir et al., 2021), Human Development Index (Ullah et al., 2023), Economic Policy Uncertainty index (Choi., 2020), volatility of European stock indexes (Aslam et al., 2021), Macroeconomic indicators (Baek et al., 2020), aggregate stock market index (Endri et al., 2021), volatility of Deutsche Bank (Batten et al., 2023), volatility of gold (Khan et al., 2023), volatility of SSE composite (Chaudhary et al., 2020), Market capitalisation (Anh and Gan., 2020), market-to-book ratio (Utomo and Hanggraeni., 2021), cash ratio (Achim et al., 2022), FinLeverage (Puławska., 2021), and return on assets (Rababah et al., 2020).

Previous studies have focused on different countries and regions such as China (Rababah et al., 2020), Europe (Puławska., 2021; Batten et al., 2023; Aslam et al., 2021), Romania (Achim et al, 2022), Indonesia (Utomo and Hanggraeni., 2021; Endri et al., 2021), Vietnam (Anh and Gan., 2020), developed countries (Chaudhary et al., 2020; Basuony et al., 2021), Global (Khan et

al., 2023; Ullah et al., 2023; Demir et al., 2021), and U.S. (Baek et al., 2020; Choi., 2020;Lúcio and Caiado., 2022; Onali., 2020; Chowdhury et al., 2022).

The COVID-19 pandemic became an important topic in finance literature because of its unexpected effects on financial markets and corporate performance. Many studies found that the pandemic significantly increased financial market volatility. For example, Chaudhary et al.,(2020) found a significant increase in volatility in major international stock markets during the early stages of the pandemic. Khan et al (2023) also reported increases in volatility across global financial assets, such as equity, commodities, and cryptocurrency. Similarly, Onali (2020) reported a significant increase in market volatility and uncertainty. Baek et al (2020) similarly reported a significant increase in market volatility across all sectors in the USA. Choi (2020) found that market uncertainty caused by the COVID-19 pandemic increased market volatility across all sectors in the USA.

Several studies also found that pandemic had a negative impact on firm performance and profitability. Rababah et al (2020) shows that the COVID-19 pandemic negatively impacted the financial performance of Chinese companies. Puławsk, (2021) similarly shows that the COVID-19 pandemic reduced the profitability and solvency of European insurance firms. Achim et al (2022) also shows reduced profitability in many firms, with the exception of firms with strong finances, which actually performed better during the pandemic. Batten et al (2023) shows that rising market volatility and uncertainty hurt the stocks of European banks worse than in the 2008 financial crisis.

Government policies also played an important role in stock market performance. Utomo and Hanggraeni (2021) found that in Indonesia, rising numbers of COVID-19 cases and deaths negatively impacted stock market returns, while lockdown policies improved stock market performance and investor confidence. Anh and Gan (2020) similarly found that in Vietnam, an increasing number of COVID-19 cases negatively impacted stock market performance, with lockdown policies improving stock market performance and investor confidence. Priya and Sharma (2023) found that stricter policies actually significantly increased market volatility initially, gradually decreasing except for in the healthcare and pharma sectors.

The effects of the pandemic were also different across sectors. Lúcio and Caiado (2022) found that in the U.S, industries such as apparel, hotels, airlines, and automobiles had experienced the largest increase in volatility. Priya and Sharma (2023) found that in India, pharma and healthcare sectors had the largest increase in volatility.

The interconnectedness in the financial market during the COVID-19 pandemic is also an important feature. Batten et al (2023) showed that the stock volatility of European banks during the pandemic was highly interconnected. Aslam et al (2021) found that market volatility of different European countries were highly interconnected as well.

After the COVID-19 pandemic, financial markets gradually stabilised. Demir et al (2021) found that vaccinations reduced the volatility of the global energy sector, helping to stabilise it.

Overall, the existing literature suggests that the COVID-19 pandemic increased financial market volatility, reduced firm performance in many cases, and affected the level of interconnectedness

between financial markets. However, the size of these effects differ across different countries and industries, effectiveness of government policy and strength of financials.

Table 1. An extensive review of literature on volatility characteristics of financial companies during the COVID-19 pandemic.

Authors	Year	Key variable	Country	Method	Findings	Time Period
Rababah et al.	2020	ROA, Size, LEV, Growth, REV	China	OLS	COVID-19 pandemic affected the financial performance of companies.	2013-2020
Puławska	2021	ROA, Z-score, LEV, Liquidity	Europe	Regression	COVID-19 reduced profitability and weakened solvency.	2010-2020
Achim et al.	2022	Working Capital, Assets, DebtEquity and Cash Ratio	Romania	Panel data	COVID-19 reduced the profitability of many Romanian companies, while financially strong firms performed better.	2019-2020
Utomo & Hanggraeni	2021	MB, Market Capitalisation	Indonesia	Panel data	COVID-19 reduced stock returns, while lockdown policies improved market performance by increasing investor confidence.	2020
Anh & Gan	2020	Market capitalization, MB, ROE	Vietnam	Panel data	COVID-19 negatively affected stock return especially before lockdowns. However, the lockdown period improved stock market performance	2020
Chaudhary et al.	2020	Volatility of stock markets	Developed countries	ARCH	COVID-19 increased volatility and initially reduced stock returns in major global markets.	2019-2020
Khan et al.	2023	Gold, Bitcoin, EUR, S&P 500, oil	Global	GARCH	COVID-19 significantly increased volatility across major financial markets gold, bitcoin, and oil.	2018-2021
Batten et al.	2023	Deutsche Bank, Ubs, Credit suisse group, VIX, Natixis	Europe	DCC-GARCH	Rising market fear (VIX) hurt European bank stocks more during COVID-19 than during the Global Financial Crisis.	2002-2021
Priya & Sharma	2023	12 Nifty sectorial indices	India	GJR-GARCH	Stricter COVID-19 measures caused a sharp rise in market volatility. FMCG, IT, media, and financial services reacted more quickly.	2020-2021
Endri et al.	2021	LQ45, JCI	Indonesia	GARCH	COVID-19 caused a sharp decline in stock prices and negative abnormal	2020

					returns as investors sold their shares.	
Baek et al.	2020	Stock market indexes and macroeconomic indicators	US	MS-AR	Market volatility was driven more by COVID-19 news than economic indicators	2020
Aslam et al.	2021	European stock indexes	Europe	DY-spillover index	During COVID-19, European stock markets were highly interconnected.	2019-2020
Choi	2020	S&P 500 and EPU	US	Wavelet coherence	COVID-19-related economic uncertainty strongly increased volatility across all U.S. stock market sectors, with a greater impact than the GFC.	2008-2020
Ullah et al.	2023	Market volatility, HDI, Life expectancy index	Global	Quantile regression	COVID-19 cases and deaths significantly raised stock market volatility.	2020
Demir et al.	2021	Datastream Energy indices	Global	Time-series regression.	Vaccination programs significantly reduced the volatility of energy companies worldwide.	2020-2021
Lúcio & Caiado.	2022	Volatility of S&P 500 industry indices	US	TGARCH	COVID-19 significantly increased volatility across most U.S. stock market industries.	2016-2020
Onali	2020	VIX, S&P 500 and DJIA.	US	GARCH	COVID-19 mainly increased market volatility rather than affecting returns.	2019-2020
Chowdhury et al.	2022	S&P 500, DJIA, Microsoft. VIX	US	VAR	COVID-19 increased volatility and reduced stock market performance.	2020
Basuony et al.	2021	Indice returns, cases, deaths, recoveries	US, Italy, Spain, UK, Germany, China, Brazil, Russia, India	Skewed t distribution	COVID-19 negatively affected major global stock markets by increasing volatility and the likelihood of market downturns.	2013-2020

Empirical finding

Data

The study examines the impact of the COVID-19 pandemic on the financial performance of six major financial institutions and the S&P 500 index. The data set is derived from Google Finance and contains approximately 2800 daily observations for each variable, covering the period from January 2nd, 2015 to April 8th, 2026.

The data covers six major financial institutions as Bank of America (BAC), Citibank (C), Goldman Sachs (GS), Morgan Stanley (MS), Wells Fargo (WFC), together with the S&P 500 index (SP).

To investigate the effect of COVID-19 on stock market volatility, the full sample is divided into three periods using dummy variables, as Pre-COVID-19 (From January 2nd, 2015 to March 13th, 2020.), COVID-19 (From March 16th, 2020 to January 3rd, 2022) , and Post-COVID-19 (January 4th, 2022 to April 8th, 2026).

Descriptive statistics

Table 2. Descriptive statistics of stock volatility for 6 major financial institutions and S&P 500.

Variable	Time period	Observation	Mean	Standard deviation	Minium	Maximum
Volatility_BAC	**Pre-COVID-19**	1279	.0155	.005	.007	.0605
	COVID-19	454	.0246	.0157	.0095	.0794
	Post-COVID-19	1064	.0165	.0048	.0087	.0347
Volatility_C	**Pre-COVID-19**	1279	.0151	.0056	.0066	.072
	Covid 19	454	.0271	.0192	.0108	.0916
	Post-COVID-19	1062	.0178	.0057	.0086	.0383
Volatility_GS	**Pre-COVID-19**	1279	.0145	.0045	.0075	.0548
	COVID-19	454	.0225	.0139	.0114	.0735
	Post-COVID-19	1064	.017	.005	.0088	.0383
Volatility_JPM	**Pre-COVID-19**	1279	.0128	.0045	.0051	.0583
	COVID-19	454	.022	.0152	.0102	.0762
	Post-COVID-19	1064	.015	.0051	.0062	.0313
Volatility_MS	**Pre-COVID-19**	1279	.0161	.0052	.0088	.0624
	COVID-19	454	.0238	.0155	.0095	.081
	Post-COVID-19	1064	.0175	.0053	.0075	.0398
Volatility_WFC	**Pre-COVID-19**	1279	.013	.0042	.0069	.0591
	COVID-19	454	.0286	.0146	.0114	.0794
	Post-COVID-19	1064	.0182	.0052	.0084	.0333
Volatility_SP	**Pre-COVID-19**	1275	.0079	.004	.0022	.0421
	COVID-19	453	.0129	.0108	.0047	.0533
	Post-COVID-19	1063	.0102	.0045	.0042	.0263

Table 2 presents the descriptive statistics of stock volatility for six major financial institutions and the S&P 500 across the Pre-COVID-19, COVID-19, and Post-COVID-19 periods. The results show that volatility increased substantially during the COVID-19 period, but declined and returned close to its Pre-COVID-19 levels during the Post-COVID-19 period.

Model

To investigate the impact of the COVID-19 pandemic on the volatility of major financial institutions, this study employs a two-sample t test. The test examines whether mean volatility differs significantly between the Pre-COVID-19 and COVID-19 period, and the COVID-19 and Post-COVID-19 period.

This study also uses a regression analysis to examine whether the COVID-19 period has a statistically significant effect on volatility. The model is presented as:

$$Y_t = \beta_0 + \beta_1 COVIDDummy_t + \acute{\varepsilon}_t$$

where Y_t represents the volatility of a financial institution, β_0 represents the average volatility during non-COVID-19 periods, while β_1 captures the change in volatility during the COVID-19 period. $COVIDDummy_t$ equals 1 for the COVID-19 period and 0 otherwise. $\acute{\varepsilon}_t$ is the error term.

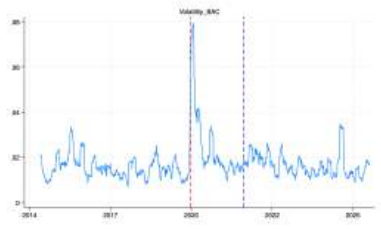
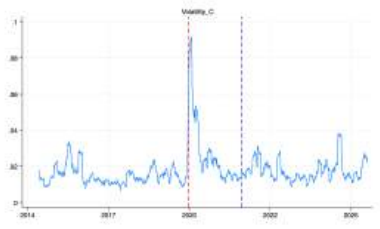
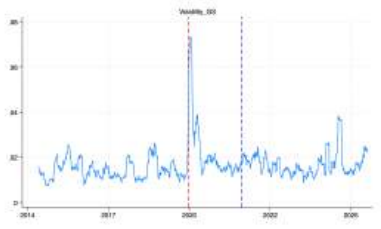
To investigate transmission of volatility among major financial institutions, this study employs Granger causality test. The test examines whether past values of one institution's volatility can predict another institution's volatility. The following bivaried model is estimated:

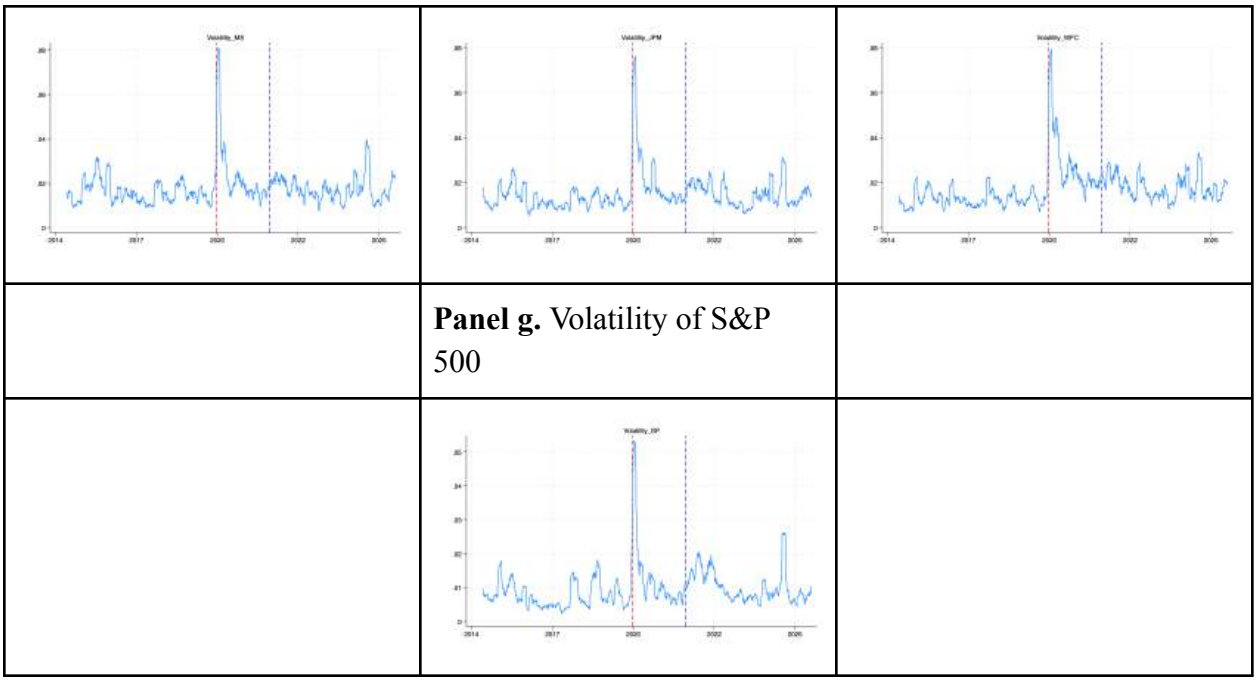
$$Y_t = \alpha + \square_1 Y_{t-1} + \square_2 Y_{t-2} + \square_3 Y_{t-3} + \gamma_1 X_{t-1} + \gamma_2 X_{t-2} + \acute{\varepsilon}_t$$

where Y_t represents the volatility of one financial institution, and X_t represents the volatility of another financial institution at time t. $\acute{\varepsilon}_t$ is the error term.
This model tests whether the past volatility of X provides useful information for predicting the current volatility of Y.

Findings

Figure 1. Volatility graphics of 6 financial companies and the S&P 500

Panel a. Volatility of BAC	**Panel b.** Volatility of C	**Panel c.** Volatility of GS
		
Panel d. Volatility of MS	**Panel e.** Volatility of JPM	**Panel f.** Volatility of WFC



Panel g. Volatility of S&P 500

Figure 1 illustrates the volatility of six major financial institutions and the S&P 500 index over the sample period. All seven series show a sharp increase in volatility during the COVID-19 pandemic, indicating a substantial rise in market uncertainty. Volatility reached its highest level during the pandemic and gradually declined afterwards. However, for most financial institutions, post-pandemic volatility remained slightly higher than pre-pandemic period. Overall, Figure 1 provides evidence that the COVID-19 pandemic significantly increased volatility in the financial sector.

Table 3, Two-sample t test with equal variances

Variable	**Time**	**Mean 1**	**Mean 2**	**dif**	**St_Err**	**P value**
Volatility_C	*Pre- COVID-19/COVID-19*	.015	.027	-.012	.0005	0.000***
	COVID-19/Post-COVID-19	.027	.018	.0095	.0005	0.000***
Volatility_BAC	*Pre- COVID-19/COVID-19*	.0155	.0245	-.009	.0005	0.000***
	COVID-19/Post-COVID-19	.0245	.0165	.008	.0005	0.000***
Volatility_GS	*Pre- COVID-19/COVID-19*	.0145	.0225	-.008	.0005	0.000***
	COVID-19/Post-COVID-19	.0225	.017	.0055	.0005	0.000***
Volatility_JPM	*Pre- COVID-19/COVID-19*	.013	.022	-.009	.0005	0.000***
	COVID-19/Post-COVID-19	.022	.015	.007	.005	0.000***
Volatility_MS	*Pre- COVID-19/COVID-19*	.016	.024	-.0075	.0005	0.000***

	COVID-19/Post-COVID-19	.024	.0175	.006	.005	0.000***
Volatility_WFC	*Pre- COVID-19/COVID-19*	.013	.0285	-.0155	.0005	0.000***
	COVID-19/Post-COVID-19	.0285	.018	.0105	.0005	0.000***
Volatility_SP	*Pre- COVID-19/COVID-19*	.008	.013	-.005	.0005	0.000***
	COVID-19/Post-COVID-19	.013	.01	.0025	.0005	0.000***

Table 3 reports the results of two sample t tests comparing the average volatility between Pre-COVID-19 and COVID-19 period, and between COVID-19 and Post-COVID-19 period. This result shows statistically significant differences at the 1% level for all six financial institutions and the S&P 500. The negative mean difference between the Pre-COVID-19 and COVID-19 periods indicate the volatility increased significantly during the pandemic. In contrast, the positive mean differences between the COVID-19 and Post-COVID-19 period show that volatility declined after the pandemic. However, Post-COVID-19 volatility generally remained slightly above Pre-COVID-19 levels. Overall, the findings suggest that the COVID-19 pandemic caused a significant increase in financial market volatility, followed by a gradual stabilisation.

Table 4. Regression analysis for volatility of Bank of America, Citibank,Goldman Sachs, and JP Morgan.

	Model 1	Model 2	Model 3	Model 4
VARIABLES	Volatility_BAC	Volatility_C	Volatility_GS	Volatility_JPM
dummy	0.0086***	0.0108***	0.0069***	0.0082***
	(0.0007)	(0.0009)	(0.0007)	(0.0007)
Constant	0.0160***	0.0163***	0.0156***	0.0138***
	(0.0001)	(0.0001)	(0.0001)	(0.0001)
Observations	2,797	2,795	2,797	2,797
R-squared	0.1428	0.1534	0.1106	0.1370

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 4 presents the regression results evaluating how the COVID-19 pandemic impacted the stock volatility of four major financial institutions. Table 4 reported the result of four different regression models, where the dependent variables are Volatility_BAC (Bank of America), Volatility_C (Citibank), Volatility_GS (Goldman Sachs), and Volatility_JPM (JP Morgan). The independent variable is the COVID-19 dummy, which takes 1 during the COVID-19 period, and 0 otherwise. We indicated the statistical significance of the results at the 1 percentage level (***), the 5 percentage level (**), and the 10 percentage level (*). The sign in front of the dummy variable suggests whether the COVID-19 pandemic has statistically a

positive or negative impact on the volatility of six major financial institutions "-" represents negative impact, while "+" represents positive impact.

The coefficient for the COVID-19 dummy is positive and statistically significant at 1% level ($p<0.01$) for all four financial institutions. This indicates that COVID-19 significantly increased stock volatility and market uncertainty in the financial sector. For example, Citibank experienced the largest increase in volatility, with a coefficient of 0.0108. Bank of America and JP Morgan followed with a coefficient of 0.0086 and 0.0082, respectively. Goldman Sachs showed the smallest increase with a coefficient of 0.0069.

Table 5. Regression analysis for volatility of Morgan Stanley, Wells Fargo, and S&P 500.

	Model 5	Model 6	Model 7
VARIABLES	Volatility_MS	Volatility_WFC	Volatility_SP
dummy	0.0070***	0.0132***	0.0040***
	(0.0007)	(0.0007)	(0.0005)
Constant	0.0167***	0.0154***	0.0089***
	(0.0001)	(0.0001)	(0.0001)
Observations	2,797	2,797	2,791
R-squared	0.0962	0.2904	0.0576

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 5 presents the regression results examining the impact of the COVID-19 pandemic on the stock volatility of Morgan Stanley, Wells Fargo, and the S&P 500 index. The dependent variables are Volatility_MS, Volatility_WFC, and Volatility_SP, while the independent variable is the COVID-19 dummy.

The COVID-19 dummy coefficient is positive and statistically significant at 1% level ($p<0.01$) in all three models. This indicates that the COVID-19 pandemic caused a sharp rise in the market uncertainty for the financial sector and the overall stock market. For example, Wells Fargo experienced the largest increase in volatility, with a coefficient of 0.0132, followed by Morgan Stanley with a coefficient of 0.0070. The S&P 500 showed the smallest increase with a coefficient of 0.0040.

Table 6. Regression analysis for return of Bank of America, Citibank,Goldman Sachs, and JP Morgan.

	(1)	(2)	(3)	(4)
VARIABLES	Return_BAC	Return_C	Return_GS	Return_JPM
dummy	0.0016	0.0007	0.0017	0.0009
	(0.0012)	(0.0014)	(0.0011)	(0.0011)
Constant	0.0001	0.0002	0.0002	0.0004
	(0.0004)	(0.0004)	(0.0004)	(0.0003)
Observations	2,827	2,825	2,827	2,827
R-squared	0.0009	0.0001	0.0012	0.0004

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 6 demonstrates the regression results examining the impact of the COVID-19 pandemic stock return of four major financial institutions: Bank of America, Citibank, Goldman Sachs, and JP Morgan. The dependent variables are Return_BAC , Return_C, Return_GS, and Return_JPM, while the independent variable in table 6 is the COVID-19 dummy.

Unlike the volatility results, the COVID-19 dummy coefficient is statistically insignificant for all four financial institutions. This suggests that the COVID-19 pandemic did not have a significant effect on their average returns.

Table 7. Regression analysis for return of Morgan Stanley, Wells Fargo, and S&P 500.

	(1)	(2)	(3)
VARIABLES	Return_MS	Return_WFC	Return_SP
dummy	0.0024**	0.0014	0.0013**
	(0.0012)	(0.0014)	(0.0007)
Constant	0.0001	-0.0001	0.0002
	(0.0004)	(0.0004)	(0.0002)
Observations	2,827	2,827	2,821
R-squared	0.0019	0.0007	0.0019

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 7 demonstrates the regression results evaluating how the COVID-19 pandemic impacted the stock return of Morgan Stanley, Wells Fargo, and the S&P 500 index. Table 7 reported the result of 3 different regression models, where the dependent variables are Return_MS (Morgan Stanley), Return_WFC (Wells Fargo), Return_SP (S&P 500). The independent variable is the COVID-19 dummy.

Unlike the results in Table 6, the COVID-19 dummy has a positive and statistically significant effect at 5% level (p<0.05) for both Morgan Stanley, with a coefficient of 0.0024, and the S&P 500 index, with a coefficient of 0.0013. However, the effect on Wells Fargo's stock returns is not statistically significant.

Table 8. Volatility spillover map during Pre-COVID-19 period. (Prob >Chi2)

From	*BAC*	*C*	*GS*	*JPM*	*MS*	*WFC*	*SP*
BAC		**0.000**	**0.000**	**0.009**	**0.000**	**0.000**	**0.000**
C	**0.0014**		0.136	**0.009**	**0.098**	**0.000**	**0.000**
GS	**0.071**	**0.000**		**0.002**	**0.001**	**0.000**	**0.000**

JPM	**0.002**	**0.000**	**0.000**		**0.000**	**0.000**	**0.000**
MS	**0.077**	**0.000**	**0.002**	**0.000**		**0.000**	**0.00**
WFC	0.485	**0.001**	**0.000**	0.273	**0.007**		**0.000**
SP	0.680	0.315	0.166	0.253	0.224	**0.036**	

Table 8 presents the volatility spillover map during the Pre-COVID-19 period. The findings indicate several statistically significant volatility transmission channels amongst US financial institutions. For example, Bank of America receives significant volatility spillover from Wells Fargo, and the S&P 500. Overall, the results suggested that volatility transmission before the pandemic was relatively structured and mainly driven by market-wide and sector-wide information rather than by individual financial institutions.

Table 9. Volatility spillover map during COVID-19 period. (Prob >Chi2)

From	*BAC*	*C*	*GS*	*JPM*	*MS*	*WFC*	*SP*
BAC		0.276	**0.017**	0.582	**0.036**	0.137	**0.098**
C	0.417		**0.017**	**0.055**	0.109	0.603	0.778
GS	0.137	0.547		0.126	0.928	0.257	0.717
JPM	0.363	0.215	0.025		**0.007**	**0.089**	**0.022**
MS	**0.034**	0.321	0.417	**0.032**		0.637	0.725
WFC	0.257	**0.018**	**0.030**	0.302	0.050		0.149
SP	**0.001**	0.221	**0.000**	**0.000**	**0.000**	0.528	

Table 9 presents the Granger causality results for volatility spillovers during the COVID-19 period. Compared with the Pre-COVID-19 period, the number of significant volatility spillovers decreases considerably. Although the COVID-19 pandemic increased volatility across all financial institutions, most Granger causality relationships became insignificant. Interestingly,
more significant spillovers are observed from individual financial institutions to the S&P 500 than in the opposite direction. These findings suggest that the financial sector became an important source of market-wide uncertainty during the pandemic. One possible explanation is that firm-specific information became more important than sector-wide information in deriving volatility.

Table 10. Volatility spillover map during Post-COVID-19 period. (Prob >Chi2)

From	*BAC*	*C*	*GS*	*JPM*	*MS*	*WFC*	*SP*
BAC		**0.089**	0.387	0.524	0.295	0.581	0.889
C	0.887		0.941	0.792	0.555	0.894	0.192
GS	0.166	**0.013**		0.647	0.193	0.710	0.813
JPM	0.294	**0.052**	0.268		0.239	0.673	0.671
MS	**0.044**	**0.002**	0.308	0.480		**0.014**	0.896
WFC	**0.655**	**0.028**	0.487	0.881	0.974		0.460
SP	**0.000**	**0.000**	**0.034**	**0.011**	**0.019**	**0.024**	

Table 10 presents the Granger causality results for volatility spillovers during the Post-COVID-19 period. Compared with the COVID-19 period, the number of significant spillovers remains relatively low, suggesting that volatility transmission between financial institutions weakened as markets gradually stabilised. However, the financial sector remained an important source of market wide volatility. The findings also suggest that firm-specific information continued to play an important role rather than sector-wide information in deriving volatility during the Post-COVID-19 period.

Conclusion

The study investigates the impact of the COVID-19 pandemic on the volatility of major financial institutions and examines whether volatility was transmitted between these institutions. The two-sample t-tests show statistically significant negative differences in volatility for all six financial companies between the Pre-COVID-19 and COVID-19 periods. This indicates that volatility increased across all firms during the COVID-19 pandemic. In contrast, there were positive differences throughout all financial institutions between COVID-19 and Post-COVID period, showing that volatility decreased after the COVID-19 pandemic period. Citibank and Wells Fargo experienced the largest changes, with mean differences of -0.012 and -0.0155 respectively, between the Pre-COVID-19 and COVID-19 period, and 0.0095 and 0.0105, respectively, between the COVID-19 and Post-COVID-19 periods.

The regression results indicate that the COVID-19 dummy variable is positive and statistically significant at the 1% level for all six financial companies and the S&P 500 index, indicating a significant increase in volatility during the pandemic. Wells Fargo and Citibank experienced the largest increases with coefficients of 0.0132 and 0.0108 respectively. However, the effect of the COVID-19 pandemic on returns was inconsistent. Only Morgan Stanley and the S&P 500 experienced statistically significant coefficients, at the 5% level, with coefficients of 0.0024 and 0.0013 respectively. All other financial firms had a statistically insignificant coefficient, indicating that the COVID-19 pandemic had a limited effect on their daily returns.

The Granger causality results indicated that there were several significant volatility transmission channels throughout the six firms during the Pre-COVID-19 period. This suggests that volatility transmission was relatively strong before the COVID-19 pandemic and was influenced by sector-wide information. All six firms also had significant transmission channels to the S&P 500, indicating that the financial sector was a major driving force for market-wide volatility.

During the COVID-19 period, the number of significant volatility transmission channels between the six firms decreased. However the number of significant volatility transmission channels from individual financial firms to the S&P 500 increased. This suggests that individual financial institutions became more important sources of market volatility during the pandemic. The decline in spillovers between the six financial companies may also indicate that firm-specific conditions became much more important rather than sector-wide conditions to investors.

During the Post-COVID-19 period, the number of significant volatility transmission channels remained relatively low. This may suggest a longer lasting change in investors valuing individual firm financials over sector-wide sentiments and performance also seen during the COVID-19 pandemic. At the same time, the financial sector remained an important source of market-wide volatility and continued to influence overall market sentiment.

This study further strengthens the connection between the COVID-19 pandemic and an increase in volatility and explores the relationship between the COVID-19 pandemic and daily returns of the stock of these firms. By identifying the effects of the COVID-19 pandemic on the volatility and returns of financial companies, a better framework for portfolio managers can be built to better manage the money of everyday citizens to find ways to minimize volatility and risk during pandemic times, possibly preventing millions from losing their retirement funds or their children's university funds.

Limitations

This study has several limitations. First, only six financial companies were used in the study. Although these are amongst the largest financial firms in the US, smaller or more local firms are not included. This could affect volatility as these large firms operate differently from smaller or more local firms. These firms also have a 'too big to fail' reputation, possibly affecting consumer confidence and thus affecting volatility. Although small, this sample size includes the largest and most influential firms in the sector, containing a large part of the total market capitalisation, thus being one of the best proxies for sector wide performance.

Second, this study does not control for other economic or financial sector conditions that may have affected volatility. Such factors could have influenced the observed mean differences.

Works Cited

Achim, M. V., Safta, I. L., Văidean, V. L., Mureșan, G. M., & Borlea, N. S. (2022). The impact of covid-19 on financial management: evidence from Romania. Economic Research-Ekonomska Istraživanja, 35(1), 1807-1832.

Anh, D. L. T., & Gan, C. (2021). The impact of the COVID-19 lockdown on stock market performance: evidence from Vietnam. Journal of Economic studies, 48(4), 836-851.

Aslam, F., Ferreira, P., Mughal, K. S., & Bashir, B. (2021). Intraday volatility spillovers among European financial markets during COVID-19. International Journal of Financial Studies, 9(1), 5.

Baek, S., Mohanty, S. K., & Glambosky, M. (2020). COVID-19 and stock market volatility: An industry level analysis. Finance research letters, 37, 101748.

Basuony, M. A., Bouaddi, M., Ali, H., & EmadEldeen, R. (2022). The effect of COVID‐19 pandemic on global stock markets: Return, volatility, and bad state probability dynamics. Journal of Public Affairs, 22, e2761.

Batten, J. A., Choudhury, T., Kinateder, H., & Wagner, N. F. (2023). Volatility impacts on the European banking sector: GFC and COVID-19. Annals of Operations Research, 330(1), 335-360.

Chaudhary, R., Bakhshi, P., & Gupta, H. (2020). Volatility in international stock markets: An empirical study during COVID-19. Journal of Risk and Financial Management, 13(9), 208.

Choi, S. Y. (2020). Industry volatility and economic uncertainty due to the COVID-19 pandemic: Evidence from wavelet coherence analysis. Finance research letters, 37, 101783.

Chowdhury, E. K., Dhar, B. K., & Stasi, A. (2022). Volatility of the US stock market and business strategy during COVID‐19. Business Strategy & Development, 5(4), 350-360.

Demir, E., Kizys, R., Rouatbi, W., & Zaremba, A. (2021). COVID-19 vaccinations and the volatility of energy companies in international markets. Journal of Risk and Financial Management, 14(12), 611.

Endri, E., Aipama, W., Razak, A., Sari, L., & Septiano, R. (2021). Stock price volatility during the COVID-19 pandemic: The GARCH model. Investment Management & Financial Innovations, 18(4), 12.

Khan, M., Kayani, U. N., Khan, M., Mughal, K. S., & Haseeb, M. (2023). COVID-19 pandemic & financial market volatility; evidence from GARCH models. Journal of Risk and Financial Management, 16(1), 50.

Lúcio, F., & Caiado, J. (2022). COVID-19 and stock market volatility: A clustering approach for S&P 500 industry indices. Finance Research Letters, 49, 103141.

Onali, E. (2020). Covid-19 and stock market volatility. Available at SSRN 3571453.

Priya, P., & Sharma, C. (2023). COVID-19 related stringencies and financial market volatility: Sectoral evidence from India. Journal of Financial Economic Policy, 15(1), 16-34.

Puławska, K. (2021). Financial stability of European insurance companies during the COVID-19 pandemic. Journal of Risk and Financial Management, 14(6), 266.

Rababah, A., Al‐Haddad, L., Sial, M. S., Chunmei, Z., & Cherian, J. (2020). Analyzing the effects of COVID‐19 pandemic on the financial performance of Chinese listed companies. Journal of Public Affairs, 20(4), e2440.

Ullah, S., Khan, S., Hashmi, N. I., & Alam, M. S. (2023). COVID-19 pandemic and financial market volatility: A quantile regression approach. Heliyon, 9(10).

Utomo, C. D., & Hanggraeni, D. (2021). The impact of COVID-19 pandemic on stock market performance in Indonesia. The Journal of Asian Finance, Economics and Business (JAFEB), 8(5), 777-784.

Appendix

Appendix 1. Data definition

Variable name	**Ticker symbol**	**Market capitalisation**	**Source**
Bank of America	BAC	447.54 billion USD	Google Finance
Citibank	C	227.75 billion USD	Google Finance
Goldman Sachs	GS	301.19 billion USD	Google Finance
JP Morgan	JPM	962.37 billion USD	Google Finance
Morgan Stanley	MS	337.38 billion USD	Google Finance
Wells Fargo	WFC	264.45 billion USD	Google Finance
S&P 500	.INX	-	Google Finance

Is It Possible and Desirable to Raise Taxation on the Top 1%? By Jayden Tran

Intro

The issue of increasing taxes on the top 1% consists of two separate questions that are often treated as one. First, can the government actually implement a higher tax on this group while also being able to collect significantly more revenue, given the risk of capital flight and offshoring. Second, even if it were possible to implement, does taxing the top 1%[1] benefit society when comparing jobs generated, entrepreneurship, and innovation against the redistribution of wealth to the public? This essay argues that raising taxes on the top 1% is possible given the government's tools and power, but should not be implemented because it may generate more societal costs than benefits.

Terms

For the essay, the top 1% refers to individuals or households in the top 1% of adjusted gross income when reported to tax authorities. This group's income threshold is around $435,000 annually (Committee for a Responsible Federal Budget). "Possible" in this essay refers to a legal and administrative mechanism to raise taxes on this group, while also generating revenue rather than leaving it flat or negative once the response of the group is accounted for. "Desirable" is a broader welfare question — a tax increase is desirable only if the social benefits of the revenue raised, through redistribution and public services, exceed the costs imposed through reduced labour supply, avoidance, capital flight, and dampened innovation. This essay treats these as separate hurdles, and argues that the top 1% clears the first while failing the second.

Why It Is Possible

There is no structural barrier preventing governments from raising tax rates on the top 1%, closing loopholes, or changing estate taxes. Proposals such as Elizabeth Warren's Ultra-Millionaire Tax, Bernie Sanders' wealth tax, and Spain's Solidarity Tax on Large Fortunes (applied to net wealth above €3 million at rates of 1.7% to 3.5%) show that such taxes can be designed, legislated, and — in Spain's case — sustained in practice (Tax Cuts Calculator). The mechanism is not in question; only its consequences are.

The top 1% earns roughly 13% of after-tax national income (Federal Reserve Bank of Philadelphia), and with typical earnings above $435,000, even a moderate rate increase touches a large base. Reducing the gap between what the law says people should pay and what they actually pay (mostly due to deductions and lower taxes on investment income) could bring in around $560 billion in one year, mostly from the highest income groups (Yale Budget Lab). Restoring the estate tax exemption closer to its inflation-adjusted year-2000 level, alongside a higher top rate, could add roughly $75 billion annually (Economic Policy Institute). These

[1] This essay will focus solely on the United States. While arguments around optimal taxation for the 1% are relevant globally and can be transferable between countries, given the space constraints of this essay, and to ensure consistency, the arguments and data presented here will concern solely the U.S.

numbers demonstrate that the current tax base is large enough for an increase to be "possible" in the sense of generating positive revenue.

Moreover, public support for taxing the ultrarich has been consistent for years (Economic Policy Institute), which removes another practical barrier: this is not a policy that requires overcoming sustained public opposition to implement.

However, a meaningful share of top-earner income is difficult to observe and tax accurately. Roughly 8% of top 1% income and 12% of top 0.1% income may go unrecorded, with gifts and inheritances particularly poorly tracked (Delestre et al.). Compounding this, much of the income at the very top arrives through business partnership structures taxed as low as 27% instead of a nominal 49%, and capital gains held until death can escape taxation entirely (Delestre et al.). A rate increase layered on top of a base that is improperly documented doesn't produce proportionate revenue; it primarily raises the incentive to exploit the very loopholes that skew the tax base, without the enforcement and base-broadening reform needed to close them.

Taken together, the legal mechanism, the size of the taxable base, and the political appetite all confirm that raising taxes on the top 1% is possible, but when accounting for the amount of income and wealth being difficult to track, the effects of a tax increase may be reduced.

Why It Is Not Desirable

The case for possibility rests heavily on static figures that assume the top 1% will not change their behavior in response to higher rates. This assumption doesn't hold. The empirical literature on the elasticity of taxable income (notably the Saez–Slemrod–Giertz body of work) finds that this group responds to higher rates by working less, shifting income into tax-favored forms, or relocating. These responses are large enough to significantly erode the revenue gains that a static calculation predicts. The high revenue figures that are used to justify a tax increase are a "best-case" scenario, and not a realistic forecast.

In an open global economy, the top 1% is disproportionately mobile, and the historical record of tax flight is not a hypothetical risk but a repeated, observed outcome. More than 10,000 high-net-worth individuals left France when their government implemented a wealth tax. As a result, the French Senate estimated roughly €200 billion in assets moved abroad over two decades (Lviv Herald). Sweden abolished its wealth tax in 2007 after its tax agency reported "significant capital flight" (Lviv Herald). The UK's 2017 tightening of non-domiciled rules saw roughly 5,000 non-domiciled leave or change status, with HMRC reporting a 20% drop the following year (The Telegraph). Even within a single country, the pattern is reoccuring: California lost 27,000 high-income taxpayers in 2021–2022, mostly to Texas and Florida (Tax Cuts Calculator). Each case shows a tax policy that looked desirable on paper, but produced disappointing, and sometimes negative, net revenue once mobility is accounted for. Each departure also removed that individual's future tax contributions entirely, not just the incremental amount from the rate increase.

The top 1% includes entrepreneurs and risk-takers whose activity generates spillovers such as new firms, jobs, and innovation that benefit people far outside this income bracket. Raising taxes on this group takes away the incentive to take on the kind of risk that produces those spillovers. This is where the concept of possibility and desirability converge: the costs of a tax increase don't only apply to the top 1% themselves but extend to workers and consumers who benefit indirectly from the economic activity that higher taxation discourages. A policy can raise revenue from a taxed group while still leaving society worse off if it suppresses activity with positive spillovers for everyone else.

Empirical work on this question largely supports that concern, though with an important qualification. Using data on essentially every U.S. patent-holder since 1920, Akcigit, Grigsby, Nicholas, and Stantcheva find that higher state-level income tax rates are associated with fewer inventors and lower rates of patenting, and that top inventors relocate toward lower-tax states (Stantcheva). Moretti and Wilson similarly find that "star scientists" are measurably responsive to state tax differentials, moving toward states that tax them less (Moretti and Wilson). A separate study of corporate taxation finds that tax increases reduce R&D spending, patenting, and new-product introduction, in part because inventors leave firms following a tax increase (ScienceDirect). Akcigit and Stantcheva further note that this relationship is asymmetric: tax increases suppress innovation more than tax cuts stimulate it, meaning the downside risk of raising taxes on this group is larger than the potential upside of lowering them (Stantcheva). Together, this evidence supports the essay's claim that taxing the top 1% carries real costs to the spillovers they generate.

The arguments for raising taxes on the top 1% establish possibility, not desirability. Large revenue estimates, a workable legal mechanism, and public support all show that the policy can be enacted and can raise money, despite some income and wealth being difficult to track. But every one of the desirability-side arguments points to the same underlying problem: the top 1% is precisely the population most able to respond to a tax increase by leaving, restructuring, or scaling back the activity that generates spillovers, and the empirical record of tax flight in France, Sweden, the UK, and even between US states confirms that this is not a theoretical risk. Once those responses are accounted for, the realistic revenue gain shrinks substantially while the costs are lost entrepreneurial activity, an eroded tax base, and intensified international tax competition. Widening inequality is a genuine problem, but a tax increase whose actual gains are far below its advertised gains, and which comes with real costs to investment and innovation, is a poor way of addressing it compared with alternatives such as closing loopholes at existing rates, improving enforcement, or pursuing internationally coordinated reform that removes the incentive to relocate in the first place.

Work Cited

Committee for a Responsible Federal Budget. “Can We Fix the Debt Solely by Taxing the Top 1 Percent?” https://www.crfb.org/blogs/can-we-fix-debt-solely-taxing-top-1-percent

Delestre, Isaac, Wojciech Kopczuk, Helen Miller, and Kate Smith. “Top Income Inequality and Tax Policy.” Oxford Open Economics, vol. 3, Supplement_1, 2024, pp. i1086–i1112. https://academic.oup.com/ooec/article/3/Supplement_1/i1086/7708125

Diamond, Peter, and Emmanuel Saez. “The Case for a Progressive Tax: From Basic Research to Policy Recommendations.” Journal of Economic Perspectives, vol. 25, no. 4, 2011, pp. 165–90. https://eml.berkeley.edu/~saez/diamond-saezJEP11opttax.pdf

Federal Reserve Bank of Philadelphia. “Taxing the 1 Percent.” https://www.philadelphiafed.org/the-economy/macroeconomics/taxing-the-1-percent

Kleven, Henrik. “Externalities and the Taxation of Top Earners.” NBER Working Paper No. 33345, National Bureau of Economic Research, Jan. 2025, revised Jan. 2026. https://doi.org/10.3386/w33345

Lviv Herald. “The Wealth Tax Is Back in Fashion, But Europe Has Been Here Before.” https://www.lvivherald.com/post/the-wealth-tax-is-back-in-fashion-but-europe-has-been-here-before

Moretti, Enrico, and Daniel Wilson. “The Effect of State Taxes on the Geographical Location of Top Earners: Evidence from Star Scientists.” American Economic Review, vol. 107, no. 7, 2017. https://www.aeaweb.org/articles?id=10.1257/aer.20150508

ScienceDirect. “Do Corporate Taxes Hinder Innovation?” Journal of Financial Economics. https://www.sciencedirect.com/science/article/abs/pii/S0304405X17300041

Saez, Emmanuel, Joel Slemrod, and Seth H. Giertz. “The Elasticity of Taxable Income with Respect to Marginal Tax Rates: A Critical Review.” Journal of Economic Literature, vol. 50, no. 1, 2012, pp. 3–50. https://eml.berkeley.edu/~saez/saez-slemrod-giertzJEL12.pdf

Stantcheva, Stefanie. “The Effects of Taxes on Innovation: Theory and Empirical Evidence.” NBER Working Paper No. 29359, National Bureau of Economic Research, 2021. https://www.nber.org/papers/w29359

Tax Cuts Calculator. “Wealth Tax Impact on Tax Planning: Lessons from European Implementation.” https://taxcutscalculator.com/blog/wealth-tax-impact-on-tax-planning-lessons-from-european-implementation

The Telegraph. “UK Wealth Tax Is ‘Inevitable,’ Says Billionaires’ Nemesis.” Feb. 5, 2026. https://www.telegraph.co.uk/business/2026/02/05/uk-wealth-tax-is-inevitable-says-billionaires-nemesis

Yale Budget Lab. “Who Is Paying Their Fair Share of Taxes? New Analysis and Interactive Tool.” https://budgetlab.yale.edu/research/who-paying-their-fair-share-taxes-new-analysis-and-interactive-tool

Economic Aspects of Sustainable Solutions In Data Centres By Isabel Shiju

Abstract

Data centres are becoming increasingly resource-intensive as demand for artificial intelligence and machine learning expands. A single data centre can consume millions of gallons of water per day, highlighting the growing environmental costs associated with data-centre expansion (Olson et al.). This paper assesses the economic implications of sustainable data-centre solutions focusing on liquid cooling, renewable energy integration, and AI-driven energy optimisation. Drawing on secondary data from academic literature, policy reports, and corporate sustainability reports from hyperscale data-centre operators, the study evaluates the costs and potential financial benefits associated with these practices. The analysis further considers the role of policy in encouraging economically viable adoption of sustainable infrastructure and proposes a policy mechanism to support this transition. The findings suggest that, although sustainable technologies may require higher initial capital investment, reductions in operational costs and resource consumption can contribute to their long-term economic viability.

Introduction

With the increasing demand for computing power driven by technological innovation and artificial intelligence, data centres have become critical infrastructure in the modern economy. As of 2025, the global data centre market is projected to generate approximately $452.53 billion in revenue ("Understanding Modern Data Center Data"). At the same time, global data centre power demand is expected to reach 219 GW over the next five years ("Data Center Demands"). These trends highlight the growing economic importance of data centres and the scale of the environmental costs associated with their expansions.

The urgency of the decisions to implement sustainability solutions increases when one looks at the long-lived nature of data centre infrastructure. Investments in facilities, energy systems, and cooling technologies can shape operating costs and environmental impacts for decades. Decisions made today about how data centres are designed and powered may influence both their financial performance and environmental footprint over the next 20–30 years.

Although the environmental impacts of data centres have been widely studied, their economic implications have received comparatively less attention. Major technology companies, including Google and Amazon, have invested in renewable energy, energy-efficient systems, and other strategies to reduce the environmental footprint of their data centres. However, the economic viability of these investments remains an important question: can sustainable infrastructure reduce environmental impacts while remaining financially viable or maybe even beneficial?

The analysis finds that although liquid cooling, AI optimisation, and renewable energy require heavy upfront investment, they can generate significant long-term savings through reduced energy and operating costs. These investments also create wider societal benefits by

reducing carbon emissions, which I quantify through the social cost of carbon. However, despite these potential long-term gains, high initial capital requirements may discourage firms from investing, particularly when they face liquidity constraints. This creates a role for government policy to reduce the financial barriers to adopting sustainable infrastructure.

Major Environmental Impacts of Data Centres

Data centres have a large environmental impact due to their constant operational needs. To understand the importance of sustainable data centres and why they are the need of the hour, it is crucial to break down the reasons for high usage of power and resources.

Data centres power every aspect of the digital world, whether it's the online banking website, AI platforms, or cloud storage. Behind the scenes, however, is a system that consumes vast amounts of energy. It projects that electricity demand from data centres worldwide is set to more than double by 2030 to around 945 terawatt-hours (TWh), slightly more than the entire electricity consumption of Japan today ("Executive Summary"). This section will look at the infrastructure, operational cycles, and cooling mechanisms that contribute to this massive energy consumption.

Massive energy consumption

Modern data centres operate continuously, and they consume large amounts of electricity. Key causes of this energy use include: continuous 24/7 uptime, high-performance computing tasks such as AI training, and redundant systems for backup and failover

Energy is one of the largest operational expenses for data centres. As prices for energy rise, operational costs increase. Without energy-efficient systems, hyperscale data centres face significant economic strain.

Cooling methodologies

Cooling systems are one of the largest energy consumers. In some facilities, they account for nearly 30-40% of total energy consumption ("How Much Electricity"). Traditional air-based cooling is particularly inefficient in hot or humid climates and often leads to overcooling, poor airflow, and energy waste.

Chillers and other mechanical systems are not only energy-intensive but also space-consuming and high-maintenance. Liquid cooling and other techniques offer improved performance but also require upgrades.

Inefficient cooling increases electricity bills and infrastructure costs. Direct liquid cooling can reduce cooling-related energy consumption by up to 72.4% compared with conventional cooling approaches (Al Kez et al.).

Water consumption

Water is used in many data centres, especially in evaporative cooling systems, cooling towers, child water systems. In these systems, water is circulated through the cooling process,

but a portion is lost through evaporation and must subsequently be replaced. Sometimes water is used indirectly through electricity generation. Lately, AI servers/GPU intensive workloads have changed cooling requirements compared with traditional computing. They produce more heat and require more intense cooling methods and hence more water. Data centres located in water-stressed regions may compete with residential, agricultural, and industrial users for limited freshwater resources.

The more pressing issue is that data centres are consuming water, and this affects water-stressed regions. Water is becoming a major concern environmentally and economically. Excessive water use leads to higher direct costs (procurement, treatment, disposal) and potential indirect costs from regulations, penalties, or reputational damage. Companies operating in water-stressed regions risk losing permits or facing additional fees. Water should not be viewed solely as an environmental input but also as a potential operational and financial risk.

Sustainable solutions and how much they cost

In the previous section, we discussed the major environmental impacts of data centers. Below, I analyze different sustainability solutions to address these concerns, such as water consumption, high energy usage, and the nature of conventional cooling methodologies.

Liquid cooling

Liquid cooling is becoming an important change in the way modern data centres are managed. Instead of relying only on fans and air conditioning, these systems use water or other cooling liquids to absorb heat directly from servers. This method is much more efficient, especially for AI and high-performance computing workloads that generate large amounts of heat. Installing liquid cooling systems can cost around 20–30% more than traditional cooling infrastructure (Miller), making the initial investment expensive. However, lower electricity use and reduced cooling costs can lead to significant long-term savings, and many companies recover their investment within a few years (Berteletti et al.). Microsoft and Meta have already introduced large-scale liquid cooling systems to support growing AI demand. As energy costs continue to rise, many companies view liquid cooling as both an environmental and economic advantage.

Using the Social Cost of Carbon (SCC), we can translate these emission reductions into broader social and economic benefits over a data center's lifetime. For instance, if a liquid cooling system cuts emissions by thousands of metric tons of CO_2 over several years, the avoided environmental and public health damages could lead to significant long-term savings for society, along with direct operational cost reductions. This supports the idea that liquid cooling not only helps the environment but also offers economic benefits.

Studies suggest that liquid cooling can reduce cooling-related electricity consumption by approximately 20–40% compared to traditional air-cooling systems. Assuming a medium-sized data centre consumes around 10 million kWh of electricity annually, this would result in electricity savings of approximately 2–4 million kWh (2,000–4,000 MWh)each year. Using

India's grid emission factor of 0.757 tCO_2/MWh, these energy savings correspond to an annual reduction of approximately 1,514–3,028 metric tons of CO_2. Applying India's Social Cost of Carbon (SCC) estimate of approximately US$86 per metric ton of CO_2 (Ricke et al. 897), these avoided emissions represent around US$130,000–260,000 in avoided environmental damages each year. Assuming an average industrial electricity tariff of approximately ₹10 per kWh in India, these electricity savings would reduce annual operating costs by approximately US$210,000–420,000, further strengthening the long-term economic case for investing in liquid cooling. These benefits extend beyond lower electricity bills, reflecting the wider costs of climate change that society would otherwise bear. This suggests that investing in liquid cooling benefits not only data centre operators through reduced operating costs but also society by lowering the environmental costs associated with carbon emissions.

However, this analysis has some limitations. The calculations assume that the electricity grid's fossil fuel composition stays relatively constant over time. In reality, many countries are gradually increasing their use of renewable energy, which could lessen the future carbon reduction impact of liquid cooling systems. Additionally, factors like changes in electricity prices, technological progress, regional energy policies, and differences in data center size may affect the actual long-term economic and environmental benefits.

AI-powered optimisation

AI optimisation is also changing how data centres operate. Instead of relying only on human monitoring, companies now use artificial intelligence to manage cooling systems, energy usage, and server workloads in real time. AI systems can identify when energy is being wasted and automatically adjust operations to improve efficiency. Setting up these systems can be costly because companies need advanced software and specialised infrastructure. However, the reduction in electricity consumption and maintenance costs can lead to significant savings over time. Google has reported that AI-based cooling systems helped reduce energy used for cooling in some facilities (Evans and Gao). As demand for AI services grows, many companies see AI optimization as a practical way to lower costs while improving performance.

Studies suggest that AI-powered optimisation can reduce overall data centre energy consumption by approximately 10–30%, depending on the facility and the extent of AI integration (Li et al. 108). For a medium-sized data centre consuming around 10 million kWh of electricity annually, this would represent electricity savings of roughly 1–3 million kWh (1,000–3,000 MWh) each year. In regions where electricity generation relies heavily on fossil fuels, such as India, these energy savings could reduce annual carbon emissions by approximately 800–2,400 metric tons of CO_2. Applying India's Social Cost of Carbon (SCC) estimate of approximately US$86 per metric ton of CO_2 (Ricke et al. 897), these avoided emissions represent annual environmental benefits of roughly US$68,800–206,400, in addition to lower electricity costs. This suggests that AI optimisation not only improves operational efficiency but also helps reduce the environmental costs associated with growing data centre demand.

However, the effectiveness of AI optimization depends on several factors. Energy savings vary according to data centre size, workload intensity, hardware compatibility, and the quality of the AI algorithms used. In addition, implementing AI optimisation requires reliable data collection systems, specialised expertise, and continuous monitoring, which may increase initial costs. As electricity grids transition towards renewable energy, the carbon reduction benefits may also decline over time, although operational cost savings are likely to remain significant.

Renewable Energy

Another major change in data centres is the adoption of **o**n-site solar photovoltaic (PV) systems. Unlike conventional data centres that rely entirely on grid electricity generated largely from fossil fuels, on-site solar installations allow operators to generate a portion of their electricity directly. Although installing solar panels requires substantial upfront capital investment, generating electricity on-site reduces long-term electricity purchases and provides greater protection against rising energy prices (*Renewable Power Generation Costs*). Many hyperscale companies, including Google and Amazon, have invested in renewable energy to reduce the carbon emissions associated with their operations and electricity consumption ("Achieving Our 100% Renewable Energy"; "Sustainability in the Cloud").

Assume a medium-sized data centre consumes approximately 10 million kWh of electricity annually. If an on-site solar PV installation supplies 50% of this demand (5 million kWh), it replaces electricity that would otherwise be purchased from India's electricity grid. Using India's average grid emission factor of 0.757 tCO_2/MWh (Central Electricity Authority), replacing 5,000 MWh of grid electricity avoids approximately 3,785 metric tons of CO_2 emissions annually. Applying India's Social Cost of Carbon estimate of US$86 per metric ton (Ricke et al. 897), these avoided emissions correspond to approximately US$325,000 in avoided environmental damages each year. In addition to reducing emissions, generating electricity on-site lowers long-term electricity expenditure and reduces exposure to fluctuations in grid electricity prices.

However, the economic benefits depend on the initial cost of installation, amount of electricity generated by the solar installation, local weather conditions, and the availability of battery storage for periods of low solar generation. Despite these limitations, on-site solar remains an effective long-term investment for reducing both operating costs and carbon emissions. Installation does require high upfront cost, however, by generating electricity on site the data centre has to purchase less electricity from the grid which reduces costs.

Comparison

When comparing these three approaches, each offers different economic advantages for data centres. Liquid cooling primarily reduces electricity and cooling costs by improving heat management, making it particularly valuable for AI and high-performance computing workloads. On-site renewable energy lowers long-term electricity expenditure and protects operators from fluctuations in grid electricity prices, although it requires substantial upfront investment in

infrastructure. AI optimisation, in contrast, improves operational efficiency by reducing unnecessary energy use and maintenance costs through real-time monitoring and automated adjustments. While liquid cooling and renewable energy involve significant initial capital investment, AI optimisation mainly requires investment in software, monitoring systems, and implementation. As these technologies address different aspects of data centre operations, many hyperscale operators are increasingly adopting them together to maximise energy efficiency, reduce operating costs, and improve long-term financial sustainability.

Policies and Incentives

This section discusses policies and initiatives that have already been established in several countries. These policies either enforce the use of sustainable solutions or provide incentives to companies that require it.

European Union – Data Centre Sustainability Reporting

The European Union aims to triple its data centre capacity by 2035 ("Energy Performance of Data Centres"). To measure energy consumption and environmental damage, more transparency from data centres is essential. A European database collects and publishes data relevant to energy performance and water consumption. The reporting applies to data centres with a power demand of at least 500kW.

This could include cloud service providers, enterprise data centres, and AI facilities. Operators must submit information on sustainability indicators, including Power Usage Effectiveness (PUE), which measures how efficiently a data centre uses electricity by dividing the total energy by equipment energy. A lower PUE means less electricity is wasted on cooling, lighting, and other supporting systems. Next is Water Usage Effectiveness (WUE), which reports how much water is consumed for cooling relative to IT usage. Next is Renewable energy use, where operators report the percentage of renewable energy used, the energy sourcing and renewable procurement methods. This encourages companies to increase renewable use over time. Waste Heat Recovery encourages operators to report whether waste heat is recovered and supplied to nearby homes or offices or heating networks. Lastly, energy consumption requires operators to disclose annual electrical use, helping policymakers understand the electricity demand and better plan the future grid.

Singapore Green Plan

Singapore is one of Asia's largest data centre hubs but has limited land and natural resources. The government developed the Green Data Centre Roadmap to ensure that future growth is both economically and environmentally sustainable (*Green Data Centre Roadmap*).

Between 2019 and 2022, the Singapore government had paused the construction of new data centres mainly due to the increasing electricity demand, the lack of land resources and rising carbon emissions. In 2022, this pause was lifted, but the government came up with a roadmap (Barrowclough). It asked operators to apply in order to get permission to set up a data centre.

The operator would only get the permission if they provided a sustainable plan. Companies had to compete for approval by demonstrating that their proposed facilities would be among the most sustainable and energy-efficient. The proposals were evaluated on their innovative cooling technologies, energy efficiency, renewable energy integration, land efficiency, and carbon reduction (*Green Data Centre Roadmap*). One of the biggest barriers to sustainable data centres is the high upfront cost of technologies such as liquid cooling, AI-based energy management, and high-efficiency power systems. Singapore's policy changes the investment landscape by making sustainability a condition for market entry. Companies that want to build new data centres must invest in efficient technologies from the outset if they wish to obtain approval.

Ireland's Data Centre Grid Connection Policy

Ireland has become one of Europe's largest data centre hubs due to its cool climate, reliable infrastructure, and presence of major technology companies such as Google, Microsoft, Amazon, and Meta (Department of Enterprise, Tourism and Employment). However, the rapid growth of data centres has placed significant pressure on the country's electricity grid. To address these challenges, the Commission for Regulation of Utilities (CRU) issued a policy in 2021 that allows Ireland's electricity transmission operator, EirGrid, to assess new grid connection applications from data centres based on their impact on the electricity system (*CRU Direction to the System Operators*). EirGrid evaluates applications using criteria given by the CRU. The criteria include showcasing the renewable energy contribution. Some proposed data centres are expected to reduce electricity consumption during periods of peak demand or operate backup generation when requested to help stabilize the grid. The policy considers whether approving a new data centre would compromise the country’s ability to provide electricity to households and essential services (*CRU Direction to the System Operators*).

Policy Recommendation To Drive Change

While sustainable technologies can substantially reduce the environmental impact of data centres, their high upfront costs remain a major barrier to adoption, particularly for small and medium-sized operators. Different sustainability measures require different policy instruments. Governments should introduce a Green Infrastructure Loan Scheme to support capital-intensive investments such as liquid cooling systems and on-site renewable energy installations. Data centre operators would submit a sustainability investment plan outlining the proposed technologies, expected improvements in energy efficiency, and projected reductions in carbon emissions and water consumption. Projects that meet predefined sustainability criteria would receive low-interest green loans and expedited approval, reducing the financial burden of initial investment.

Following implementation, operators should report key sustainability indicators such as Power Usage Effectiveness (PUE), renewable energy usage, carbon emissions, and water consumption to demonstrate that the proposed improvements have been achieved. This reporting

process ensures transparency and accountability while allowing governments to monitor the effectiveness of funded projects.

In addition to investment support, broader market-based policies such as carbon pricing or emissions trading systems can encourage greater adoption of renewable electricity by increasing the cost of carbon-intensive energy. Together, targeted financial support for infrastructure investments and market-wide incentives can accelerate the transition to sustainable data centres while maintaining long-term economic viability.

Conclusion

This paper demonstrates that sustainable data centre practices are not only environmentally beneficial but also economically viable. Cutting energy waste, switching to renewables, and improving cooling systems all reduce long-term operating costs. Additionally, several government policies are pushing companies in this direction, so the shift will become unavoidable. The real challenge is scaling these solutions to meet the rising global demand. If companies combine innovation with strong environmental planning, data centres can become much more efficient and still support the growth of AI and global digital services. Sustainable design is no longer optional; it's turning into the foundation of future data-centre development.

Works Cited

"Achieving Our 100% Renewable Energy Purchasing Goal." Google, 2024, sustainability.google/reports/achieving-100-renewable-energy-purchasing-goal/.

Al Kez, Dlzar, et al. "AI-Driven Cooling Technologies for High-Performance Data Centres: State-of-the-Art Review and Future Directions." Sustainable Energy Technologies and Assessments, vol. 82, 2025, no. 104511, https://doi.org/10.1016/j.seta.2025.104511.

Barrowclough, Nicholas. "Singapore Lifting the Ban for New Data Centers: What's the Catch?" AKCP, 24 Feb. 2022, www.akcp.com/2022/02/24/singapore-lifting-the-ban-for-new-data-centers-whats-the-catch/.

Berteletti, Elia, et al. "Keeping Cool in the Data Age." McKinsey & Company, 24 Sept. 2025, www.mckinsey.com/capabilities/operations/our-insights/operations-blog/keeping-cool-in-the-data-age.

Central Electricity Authority. CO_2 Baseline Database for the Indian Power Sector: User Guide. Version 20.0, Ministry of Power, Government of India, 2024, cea.nic.in/wp-content/uploads/baseline/2025/12/User_Guide_V_21.0.pdf.

CRU Direction to the System Operators Related to Data Centre Grid Connection Processing. CRU/21/124, Commission for Regulation of Utilities, 23 Nov. 2021, cruie-live-96ca64acab2247eca8a850a7e54b-5b34f62.divio-media.com/documents/CRU2025236_Large_Energy_User_connection_policy_decision_paper.pdf.

"Data Center Demands." McKinsey & Company, 20 May 2025, www.mckinsey.com/industries/public-sector/our-insights/the-data-center-balance-how-us-states-can-navigate-the-opportunities-and-challenges.

Department of Enterprise, Tourism and Employment. The Value of Data Centres to Ireland. Government of Ireland, 2026, enterprise.gov.ie/en/publications/publication-files/the-value-of-data-centres-to-ireland.pdf

"Energy Performance of Data Centres." European Commission, energy.ec.europa.eu/topics/energy-efficiency/energy-efficiency-targets-directive-and-rules/energy-efficiency-directive/energy-performance-data-centres_en.

Evans, Rich, and Jim Gao. "DeepMind AI Reduces Energy Used for Cooling Google Data Centers by 40%." Google, 20 July 2016, blog.google/company-news/outreach-and-initiatives/environment/deepmind-ai-reduces-energy-used-for/.

"Executive Summary." Energy and AI, International Energy Agency, 10 Apr. 2025, www.iea.org/reports/energy-and-ai/executive-summary.

Green Data Centre Roadmap. Infocomm Media Development Authority, 2024, www.imda.gov.sg/-/media/imda/files/how-we-can-help/green-dc-roadmap/green-dc-roadmap.pdf.

"How Much Electricity Does a Data Center Use? Complete 2025 Analysis." IAEI Magazine, 1 Jan. 2026,

iaeimagazine.org/electrical-fundamentals/how-much-electricity-does-a-data-center-use-complete-2025-analysis/.

Li, Xiyong, et al. "A Review on AI-Driven Optimization of Data Center Energy Efficiency and Thermal Management." International Journal of Applied Science, vol. 8, no. 3, 2025, p. 108, https://doi.org/10.30560/ijas.v8n3p108.

"Liquid Cooling in Data Centres." IEA 4E Energy Efficient End-use Equipment, 22 June 2026, www.iea-4e.org/edna/publications/liquid-cooling-in-data-centres/.

Miller, Garth. "The Data Center Cooling Revolution: Comparing Liquid Cooling Technologies." Data Centre Digest, 26 Apr. 2026, datacentredigest.com/the-data-center-cooling-revolution-comparing-liquid-cooling-technologies/.

Olson, Eric, et al. "Data Centers Are Draining Resources in Water-Stressed Communities." The Dallas Morning News, 6 May 2024, www.dallasnews.com/opinion/commentary/2024/05/06/data-centers-are-draining-resources-in-water-stressed-communities/.

Renewable Power Generation Costs in 2022. International Renewable Energy Agency, 2023, www.irena.org/Publications/2023/Aug/Renewable-Power-Generation-Costs-in-2022.

Ricke, Katharine, et al. "Country-Level Social Cost of Carbon." Nature Climate Change, vol. 8, no. 10, 2018, pp. 895–900, https://doi.org/10.1038/s41558-018-0282-y.

"Sustainability in the Cloud." Amazon Web Services, 2026, docs.aws.amazon.com/wellarchitected/latest/sustainability-pillar/best-practices-for-sustainability-in-the-cloud.html.

Perceived AI-Likelihood, Disclosure, and Consumer Responses to AI-Generated Fashion Imagery: A Between-Subjects Pilot Study By Gabriella Messina

Abstract

As generative AI becomes increasingly common in advertising, brands face questions about its effects on consumer perceptions and the ethics of AI disclosure. This study examined how consumers respond to AI-generated versus human-created fashion imagery, focusing on perceived AI-likelihood, emotional response, perceived authenticity, brand trust, purchase intention, and brand evaluation following AI disclosure. Fifty participants were randomly divided into two groups and shown either human-created or AI-generated fashion images featuring the same products. Results showed differences in participants' perceived AI-likelihood ratings between the AI-generated and human-created image conditions, although the magnitude of the overall difference was limited. Gender and age were also associated with differences in perceived AI-likelihood ratings. Participants demonstrated greater uneasiness and discomfort toward AI-generated images, although this difference was not statistically significant. AI disclosure produced different changes in brand evaluation depending on participant characteristics, particularly age and AI tool familiarity. Mediation analyses further suggested that emotional responses were related to purchase intention through brand trust, while AI familiarity helped explain changes in brand evaluation following disclosure. These findings suggest that the effects of AI-generated advertising extend beyond consumers' perceptions of whether an image was created using AI and highlight the importance of transparency and responsible AI use in fashion advertising. The small convenience sample and pilot design limit the generalizability and precision of these findings.

Keywords

Generative AI; AI-generated imagery; fashion advertising; consumer behavior; AI disclosure; brand perception; authenticity

Introduction

In only a few years, generative AI has evolved from producing obviously artificial images to creating visuals that consumers may struggle to distinguish from human-created content. Many brands have taken advantage of this, using AI for more efficient advertising, cheaper product imaging, and quicker results (Reid). However, detection accuracy is not the only factor shaping consumer responses to AI images. If consumers are unaware of an image's origins, their emotional responses, perceptions of authenticity, and evaluations of the brand may be affected. This is especially relevant in fashion, where consumers rely heavily on visual advertising to evaluate products and where the appearance of models can influence both product perceptions and personal standards of appearance.

Previous research has examined several aspects of how consumers respond to AI-generated imagery, but the findings remain mixed, and many factors have been studied

separately. Research on AI detection ability suggests that consumers are generally not highly accurate at distinguishing AI-generated images from human-created ones, although some studies have found differences based on characteristics such as age and gender. Other research has focused on the emotional effects of AI imagery, finding that consumers may experience greater uneasiness or discomfort when viewing AI-generated content compared with human-created content. Studies have also examined authenticity and trust, suggesting that perceptions of whether an image feels genuine can influence how consumers evaluate the brands and products associated with it. In fashion specifically, research has raised additional concerns about the effects of AI-generated models on consumers' perceptions of appearance. For example, Alleva et al. found that exposure to idealized AI-generated models could negatively affect aspects of women's body image, demonstrating that the effects of AI-generated fashion imagery can extend beyond the advertisement itself. AI disclosure creates another important issue. Although disclosure can provide consumers with greater transparency and allow them to make more informed judgments, it may also change how consumers evaluate a brand. Existing research therefore suggests that AI-generated advertising can affect consumers' emotional, psychological, and behavioral responses, but less is known about how these responses interact with one another. In particular, limited research has examined detection ability, emotional response, authenticity, brand trust, purchase intention, and reactions to AI disclosure together within the context of fashion advertising. Additionally, less is known about how individual differences, including age, gender, and familiarity with AI technology, shape these responses.

These gaps are increasingly important as brands gain access to more realistic and accessible AI-generated imagery. If consumers cannot reliably identify AI-generated content, brands may have greater responsibility for deciding whether and how to disclose its use. This creates an important tension between the potential marketing benefits of AI and the ethical considerations surrounding transparency and accurate representation. This study therefore asks: “How does exposure to AI-generated versus human-created product images influence consumers' emotional responses, purchase intentions, and brand perception, and does consumers' ability to identify AI-generated imagery relate to these responses?” To answer this question, this study compares consumer responses to AI-generated and human-created fashion imagery while examining detection ability, emotional response, perceived authenticity, brand trust, purchase intention, and changes in brand evaluation following AI disclosure. It also examines whether demographic characteristics and AI familiarity influence these responses.

Literature Review

The increasing use of generative AI in advertising and product imaging has raised questions about how consumers perceive and respond to AI-generated visual content, especially as generative imaging capabilities improve. Existing research suggests that consumer responses to AI-generated images are complex and may depend on factors such as AI disclosure, perceived authenticity, emotional response, and the type of product being advertised. While some studies indicate that consumers may respond similarly to AI-generated and human-created content when

AI use is undisclosed (Lee), other research has found that disclosure can increase perceptions of artificiality, eeriness, and manipulation (Rapp et al.). Together, these findings suggest that the effects of AI-generated imagery on consumer behavior are not straightforward and warrant further investigation.

AI Marketing Compared to Traditional Marketing

As AI advances, it has become increasingly convenient for advertisers to incorporate AI product imaging and customization into their marketing strategies. Many advertisers are turning towards AI for its ability to create specific, targeted advertisements (Ford et al.). Additionally, AI advertising is often more effective when the intended image is abstract, rather than when it aims to recreate traditional photography, meaning that authentic-appearing images aren't always better (Qin et al.). Furthermore, generative AI can provide fashion companies with substantial improvements in the speed and cost of visual content production (Reid). For example, Swedish fashion company Zalando reported its ability to reduce image-production time from six to eight weeks to three to four days while reducing associated costs by 90%. The company also uses AI-generated digital model replicas to create consistent imagery across campaigns and product pages, allowing it to respond more quickly to rapidly changing fashion trends (Reid). However, these production benefits do not necessarily indicate that consumers will respond positively to AI-generated imagery.

Psychological Responses to AI

While generative AI offers significant benefits to advertisers, its increasingly human-like content can also elicit complex psychological responses from consumers. One example of this is the uncanny valley- a negative psychological reaction that occurs in humans when something feels almost human, but not fully. It evokes eeriness and uneasiness, even if the image appears realistic (Kishnani). Rapp et al. found that after participants were presented with 20 AI images and asked to report their reactions, many respondents said that the images felt 'artificial' or 'strange'. This factor was influenced by the image's ability to accurately represent the intended product, as well as its technical quality, suggesting that quality and realism alone do not inherently generate more positive feelings among consumers.

Another study found that perceived AI authorship reduced feelings of authenticity among respondents, which in turn elicited negative emotional responses and even moral disgust. However, this applies mainly to emotional content, compared to factual content, in which the negative emotional response and low authenticity rating were less prominent. For instance, when the chatbot was attempting to convince the participants to purchase from a brand using emotional appeal, this received more negative feedback than when the chatbot was solely used for editing or sharing factual information (Kirk and Givi).

It is unclear whether people are able to distinguish between AI and human-created images. According to Högemann et al., while on average, people struggle to distinguish between AI-generated and human-made images, younger people are significantly more accurate in doing

so than older populations. They did not find gender to be a significant predictor of detection accuracy. In another study (Ivan) performed with a sample size of 831 people looking at dating profiles (with a mix of real and AI-generated profiles), while overall detection accuracy was low, women were significantly better at determining which images were AI, although they did report more false positives (Ivan). One notable difference in the methodology of these studies was that Högemann's study featured mainly images of landscapes and objects, while Ivan's study was focused solely on human images. Ivan's study also had a larger sample size with equivalent numbers of men and women, while Högemann's had double the number of men as women. This suggests that gender is more relevant for AI detection abilities when human faces are involved.

In a study performed on the effectiveness of COVID-19 vaccine advertisement strategies, researchers discovered that people's prior emotional state before interacting with AI can be indicative of their responses. For instance, when the participants were angry, interacting with a human was more effective, whereas when participants felt embarrassed or neutral, they tended to prefer interacting with the AI bot. The study also outlined the concept of anthropomorphization, or attributing humanlike qualities to non-human entities. However, there is significant overlap between anthropomorphization and uncanny valley psychological responses, as more human-like AI can often trigger discomfort and uneasiness if something feels 'off' (Diwanji et al.).

These psychological responses to AI-generated content extend beyond immediate emotions, influencing consumers' evaluations of the advertised brand and their willingness to purchase. Existing research suggests that perceptions of authenticity, ethics, and overall attitudes toward AI-generated content can shape consumer responses and behavioral intentions (Avery et al.; Herdimansyah; Yıldırım and Dursun). In particular, research on AI-generated fashion and digital advertising indicates that consumers may evaluate AI content not only based on its visual quality, but also on what its use communicates about the brand's authenticity and values (Herdimansyah). These findings suggest that emotional reactions to AI-generated imagery may have broader implications for brand trust and purchase intention, making these outcomes important to consider alongside consumers' immediate psychological responses.

Zhang and Hur found that consumers evaluated AI-generated advertising images differently from human-generated images, demonstrating that the perceived use of AI can influence broader consumer attitudes toward advertising. Similarly, research on AI-generated luxury advertising found that factors including trust in AI, perceived humanness, informativeness, and novelty influenced consumers' responses and purchase intentions (Jung et al.). Research on AI-generated brand personalities has also found that AI-generated images can influence brand preference and purchase intention, suggesting that consumers' responses to AI imagery can extend beyond their immediate reaction to the image itself (Park and Ahn). While these studies indicate differences between responses to AI and human-made images, existing research has not fully established how consumers' emotional reactions to AI-generated product imagery relate to perceived authenticity, brand trust, and purchase intention. Examining these relationships may therefore provide a more complete understanding of how AI-generated imagery influences consumer decision-making.

AI Disclosure and Brand Evaluation

In a 2025 case study, public reactions to AI fashion models were measured using social media engagement. Due to the artistic and creative nature of fashion, consumers expect higher levels of authenticity compared to other industries. This led to lower AI acceptance for fashion-related AI-generated media, since viewers were expecting human-generated content. The study argues that in this sense, public reactions can serve as an 'informal form of governance', since if there is a strong consumer rejection of AI-generated content, companies will be incentivized to avoid using it in advertising and product imaging (Herdimansyah).

The emotional side effects and authenticity expectations have led to significant debates surrounding the ethicality of AI disclosure. Since there is no consistent or legally mandated standard of AI disclosure, one study proposed an AI attribution system that indicated varying levels of AI involvement. They cited three main benefits of clear AI disclosure: improved public perceptions, reduced legal risks, and more ethical reporting. Disclosing AI allows consumers to form a more informed evaluation of the media, rather than feeling deceived or uncomfortable, as the uncanny valley often suggests. Additionally, clear AI disclosure can mitigate legal risk for intellectual property and misleading content, while holding companies accountable for their AI use (Avery et al.).

A study performed by Yıldırım and Dursun measured three perception variables -perceived manipulation, perceived ethics, and purchase intention- for human-generated media, AI-generated media with disclosure, and AI-generated media without disclosure (all testing the same skincare product). They found that perceived manipulation was increased for both disclosed and undisclosed AI, which in turn reduced the participants' perceived ethics, as they viewed the AI as unfair or dishonest towards customers. Additionally, improved perceived ethics was correlated with higher purchase intention; however, perceived manipulation and purchase intention were not directly related. This demonstrates that a lot of the emotional reactions to AI stem from perceived dishonesty and inauthenticity, which are often impacted by AI disclosure.

The effects of AI disclosure may also vary according to consumers' familiarity with artificial intelligence. Shi et al. found that consumers with higher levels of AI literacy evaluated AI-generated advertisements more positively and were less sensitive to whether AI use was explicitly disclosed, while consumers with lower AI literacy responded more negatively to AI-generated advertising. These findings suggest that consumers' prior experience and knowledge of AI may influence how they interpret AI-generated content and its disclosure, highlighting the importance of considering individual differences when examining consumer responses to AI advertising.

Gaps In Existing Literature

Although existing research has established that generative AI can provide advertisers with greater efficiency, flexibility, and creative capabilities, several gaps remain in understanding how consumers respond to AI-generated product imagery. First, much of the existing research examines AI in broad advertising contexts, such as AI influencers, chatbots, or written marketing

communications, rather than focusing specifically on AI-generated product and fashion imagery. This is important because visual content may produce different psychological responses than other forms of AI-generated communication.

Second, existing studies have identified several individual consumer responses to AI, including perceptions of authenticity, trust, eeriness, ethical concerns, and purchase intention, but these factors are often examined separately. For example, research has demonstrated that AI disclosure can influence authenticity and advertising evaluations, while other studies have examined emotional responses such as uneasiness or uncanniness. However, less research has examined how these psychological responses interact within the same consumer decision-making process.

Third, existing findings suggest that consumer reactions to AI-generated advertising are not universally positive or negative and may depend on factors such as disclosure, the purpose of AI use, and the characteristics of the advertisement. Therefore, further research is needed to examine how consumers' ability to identify AI-generated imagery relates to their emotional responses and subsequent evaluations of the brand. The present study addresses these gaps by directly comparing AI-generated and human-generated fashion imagery and examining consumers' AI detection ability, emotional responses, perceived authenticity, brand trust, and purchase intention in relation to each other.

Methods and Data

This study used a between-subjects experimental design to examine differences in consumer responses to human-created and AI-generated fashion imagery. Participants were assigned to view either human-created or AI-generated images featuring the same clothing products. The study examined AI detection ability, emotional response, perceived authenticity, brand trust, purchase intention, and responses to AI disclosure.

This study included 50 participants, with each randomly assigned to either group A (real image) or group B (AI-generated image) (Table 1). There were 25 respondents in group A, and 25 respondents in group B. Thirty-nine of the total respondents were women, and 11 were men. The participants' ages ranged from 14 to 76, with a mean age of 26.5. Participants provided informed consent before completing the survey and were free to decline participation. No personally identifying information was collected, and all responses were anonymized for analysis.

Table 1: Summary Statistics of the Respondents

Variable	Classification	Group A	Group A	Group B	Group B
		Number	Percentage	Number	Percentage
Sex	Male Female	7 18	28% 72%	4 21	16% 84%

Age	<26.5 >=26.5	15 10	60% 40%	22 3	88% 12%

Survey Breakdown

Participants in group A were presented with three human images from the fashion company Aerie (see Figure 1). Three images from Aerie were selected because the company has publicly stated that it does not use AI-generated models in its product imagery. Their official policy is, “We commit to never use AI to generate bodies or change the people and bodies in our images.” (Aerie). The participants were then prompted with a series of questions. First, they were asked to rate on a scale of 1-5 how much they thought the image was altered with AI, with 1 being completely unaltered, 3 being modified, and 5 being completely AI-generated. Following that, they were asked to rate their emotions towards the image, in terms of uneasiness, discomfort, and a free-response question for their own descriptors. This emotion scale aligns with the Richins standardized scale (Richins).

After rating their emotions, an attention checker prompted them to choose the middle choice. This was included to find participants who were not paying attention to the questions, which was important due to the large number of rating questions in the survey. Next, they were asked to rate how authentic and genuine the images appeared. These factors measured the perceived authenticity of the images, according to the Napoli et al. scale (Napoli et al.). Finally, they were asked a variety of questions about brand trust and purchase intention, including: I trust the brand shown in the advertisement, I would seek more information about this product, I would consider visiting this brand's website, and I would consider purchasing this product.

The next part of the survey is the AI disclosure section. It gives participants the prompt, “Imagine that the fashion company confirms that the promotional images you just viewed were generated or significantly modified using artificial intelligence.” Then, it asks (both on a scale of 1-5, with 1 being least favorable and 5 being most): how would this information affect your overall evaluation of the brand, and how likely would you be to purchase a product from this brand? Next, it asks how important it is for companies to disclose when promotional images are generated or significantly altered with AI, and gives a free-response question to answer how much and why the participant's brand evaluation would change.

The next section is a series of demographic questions in order to gather information about how participants' backgrounds impact their AI perceptions. These questions include: participant age, gender, online shopping frequency (scale of 1-5 with 5 being once a week or more), prior awareness of AI use in fashion marketing, AI tool use frequency (scale of 1-5, 5 is very often), and how often the participant notices AI in digital advertisements (scale of 1-5, 5 is very often). Finally, the survey includes two free-response questions asking what features they use to spot AI use in advertisements, and if they have any closing thoughts or reactions.

The survey for group B was nearly identical, except that they were shown images of models generated with ChatGPT, wearing the same clothes as the Aerie models (see Figure 1).

ChatGPT was prompted to generate models wearing the same clothes and in the same background as the Aerie models, but was otherwise given little guidance.

Image Set A - human models (Aerie)

Image Set B- AI-generated models

Figure 1: Image Sets

Data Analysis

Data were analyzed using R. Descriptive statistics were first calculated for all major variables. Cronbach's alpha was used to assess the internal consistency of the emotional-response scale, authenticity scale, and purchase intention scale. Independent-samples t-tests were used to compare participants' responses to AI-generated and human-created images across measures of detection ability, emotional response, authenticity, and brand-related evaluations. Regression analyses were conducted to examine whether demographic characteristics and prior AI familiarity predicted AI detection ability. Statistical significance was evaluated at $\alpha = .05$. Correlation analyses were also run to determine correlations between brand and product evaluation variables.

Results

Scale Reliability

Internal consistency for the multi-item measures (combined uneasiness and discomfort factors) was acceptable (Cronbach's $\alpha = 0.74$), indicating adequate reliability for the survey scales used in subsequent analyses (Tavakol and Dennick). The internal consistency for feelings of authenticity (authentic, genuine, visually appealing) was acceptable as well, with a Cronbach's α of 0.74. The consistency of purchase intention variables (intention to seek more information about the brand, intention to visit the brand's website or social media page, and consideration to purchase products shown) was excellent, with a Cronbach's α of 0.91 (Tavakol and Dennick).

AI Image Detection Ability

Across the full sample, participants' perceived AI-likelihood ratings showed differences between the AI-generated and real-image conditions. However, because participants were assigned to view only one image condition, these results measure differences in perceived AI-likelihood between groups rather than individual detection accuracy.

Gender was significantly associated with differences in perceived AI-likelihood ratings. Women gave higher AI-likelihood ratings to the AI-generated images (Mean = 2.71) than the human-made images (Mean = 2.11), indicating a significant difference in their responses ($p = .042$). In contrast, men did not show a significant difference between their responses to AI-generated images (Mean = 2.25) and human-made images (Mean = 2.71; $p = .31$).

Age was also associated with differences in perceived AI-likelihood ratings. Younger participants (<26.56 years) showed a significant difference between their responses to AI-generated images (M = 2.59) and human-made images (M = 1.93; $p = .045$). Older participants (≥26.56 years), however, did not show a significant difference between their responses to AI-generated images (M = 3.00) and human-made images (M = 2.80; $p = .765$).

Self-reported familiarity with AI tools was not significantly associated with perceived AI-likelihood ratings ($t = 1.09$, $p = .292$). Participants who reported shopping online more frequently showed slightly greater differences in perceived AI-likelihood ratings, although this relationship was not statistically significant ($t = 1.80$, $p = .079$).

Table 2: Factors Associated with AI Image Detection Responses

Variable	Group	AI-Generated Images, M	Real Images, M	*t*	*p*	Significant?
Gender	Women	2.71	2.11	2.106	**.042**	**Yes**
	Men	2.25	2.71	-1.063	.310	No
Age	Younger (<26.56 years)	2.59	1.93	2.089	**.045**	**Yes**
	Older (≥26.56 years)	3.00	2.80	0.337	.765	No
AI tool familiarity	Overall	3.00	2.444	1.09	.292	No
Online shopping frequency	Overall	2.65	2.111	1.80	.079	No

Emotional Response to AI-Generated Images

Participants reported greater discomfort and uneasiness when viewing AI-generated images compared to real ones, and the difference was marginally significant ($t = 1.79$, $p = 0.08$) (Figure 2).

However, in the free-response question for the human-made images, many respondents expressed ideas of the images appearing carefree and welcoming (regardless of their demographic background), such as, “Even though the style of clothing doesn't align with my personal style, the posing of the models feels very welcoming and makes me feel as if the clothes would be comfortable to wear. I wouldn't purchase any of the three specific clothes shown, but I would be open to other items from the brand.” Others commented on how the models weren’t wearing too much makeup and looked like people you might see in real life. Respondents who viewed the AI-generated images have very different responses, with many commenting on how they look like typical models. One participant even noted, “Models have always had an uncanny,

unrealistic quality to them. However, with the rise in AI, the unrealistic appearance of models goes deeper into a suspicion that the people you are comparing yourself to aren't even real. That's a really gross feeling and is something humans were never meant to experience." While the emotion scales for both the real and AI-generated images yielded similar responses, the open-ended questions demonstrated that people tended to feel more connected to the human-made images in terms of relatability and self-acceptance.

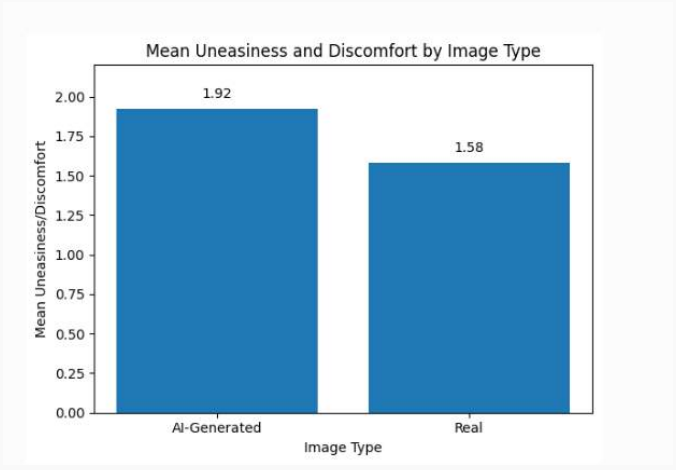


Figure 2: Results of t-test comparing emotions in AI vs. real images

Perceived Discomfort, Brand Trust, and Purchase Intention

To examine the relationships among participants' emotional responses, brand trust, and purchase intention, feelings of discomfort or uneasiness were associated with brand trust, controlling for perceived authenticity and age were examined. The model was significant ($F(4,45) = 13.28$, $p < .001$), explaining 54.2% of the variance in brand trust (adjusted $R^2 = .542$). The strongest predictor found in the regression was authenticity (**β**=0.63, p<0.001). This demonstrates that high perceived authenticity contributes to high brand trust, while low perceived authenticity leads to lower brand trust. Additionally, emotional discomfort was an important factor in brand trust, as lower uneasiness ratings were found to significantly increase brand trust (**β**=-0.43, p=0.005). This result demonstrated that the greater one's discomfort with the images was, the lower their overall trust of the brand would be. Finally, age was also significantly associated with brand trust, with older participants reporting lower overall trust (**β**=-0.02, p=0.001).

A second regression suggests that greater brand trust was associated with higher purchase intention when controlling for discomfort or uneasiness, perceived authenticity, and age, $b = 0.897$, $SE = 0.205$, $t(45) = 4.38$, $p < .001$.

Then, brand trust's mediation of the relationship between discomfort or uneasiness and purchase intention was explored. A mediation analysis using 10,000 bootstrap simulations revealed a significant indirect effect through brand trust, $ab = -0.378$, 95% CI [−0.811, −0.084], $p = .007$. The direct association between discomfort or uneasiness and purchase intention after accounting for brand trust was not significant, $b = 0.006$, $SE = 0.217$, $t(45) = 0.03$, $p = .978$.

These results suggest that greater feelings of discomfort or uneasiness were associated with lower brand trust, which in turn was associated with lower purchase intention.

Brand Evaluation After AI Disclosure

When participants were informed that an image was generated using AI, their responses were strongly correlated with demographics and personal beliefs. Reported AI-disclosure importance was strongly correlated with a negative brand evaluation change ($r = –.58$, $p < .001$). This means that participants who placed greater importance on AI disclosure showed larger decreases in brand evaluation after learning that an image was AI-generated. When asked 'Why would the use of AI make your evaluation of the brand more favorable, less favorable, or remain unchanged?', many respondents claimed their negative shift in post-AI disclosure brand evaluation was due to distrust of the product in real life, perceived laziness from the brand, and environmental concerns about AI use cited by younger people in particular. One participant in particular said that if AI was disclosed, "It would make my evaluation of the brand less favorable because the use of AI leads me to believe that the item's true form isn't realistically displayed and therefore makes me more hesitant to purchase".

Age differences were significant as well. Based on the Pearson correlation, age and brand evaluation change following AI disclosure are strongly correlated, with younger participants having a more negative brand evaluation change and older participants having an indifferent or positive change ($p=0.00011$). However, age and beliefs about AI disclosure were correlated as well; younger participants had significantly stronger beliefs about AI disclosure, with the majority of younger respondents in favor of it ($t=-2.16$, $p=0.035$). Older participants, on the other hand, were mainly indifferent to AI disclosure as a whole.

Next, regression models were used to examine whether prior AI tool use was associated with participants' responses to AI disclosure. More frequent AI tool use was associated with lower perceived importance of AI disclosure, $b = −0.257$, $SE = 0.113$, $t(47) = −2.28$, $p = .027$, controlling for age. Perceived importance of AI disclosure was, in turn, associated with change in brand evaluation when controlling for AI tool use frequency and age, $b = −0.352$, $SE = 0.097$, $t(46) = −3.63$, $p < .001$.

Next, it was explored if perceived importance of AI disclosure mediated the relationship between AI tool use frequency and change in brand evaluation. The indirect effect was significant based on 10,000 bootstrap simulations, $ab = 0.090$, 95% CI [0.010, 0.192], $p = .022$. The direct association between AI tool use frequency and change in brand evaluation after accounting for disclosure importance was not significant, $b = 0.080$, $SE = 0.079$, $t(46) = 1.02$, $p = .313$. The total effect of AI tool use frequency on change in brand evaluation was significant, $b = 0.171$, 95% bootstrap CI [0.016, 0.317], $p = .031$. These findings suggest an indirect relationship between AI tool use frequency and changes in brand evaluation through perceived AI disclosure importance.

Discussion

AI vs. Real - Consumer Evaluation

While on average, people were unable to consistently identify the AI-generated and human-made images, certain demographic groups showed greater ability to differentiate. Women and younger people were significantly more accurate in their detection abilities than men and older people. This aligns with prior studies that women tend to be more accurate in identifying human faces generated by AI than men are. Ivan suggests that women's higher AI-image detection accuracy may reflect greater skepticism toward visual deception, although their higher rate of false-positive identifications indicates that this may represent a more skeptical response tendency rather than objectively superior detection ability. The most likely reason younger audiences outperform older participants in AI image detection is their familiarity with AI-generated content as well as differing AI-detection strategies.

This raises a new question: if people are, on average, unable to differentiate between real and AI-generated content, can AI still influence their evaluation of the brand?

Psychological Responses to AI

While our study found that feelings of discomfort were more prevalent in those shown the AI images as opposed to the human-made images, this difference was marginally significant. However, when asked how the images made them feel in their own words, participants shown the AI image were much more likely to report something feeling ‘off’ or ‘artificial’. This highlights the psychological phenomenon of the Uncanny Valley, the feeling that occurs when something is almost human but has nonhuman characteristics (Kishnani). This demonstrates that even though they may not be able to fully identify whether an image is AI-generated, humans still tend to perceive AI and human images differently.

In addition to the emotional response, brand trust, perceived authenticity, and purchase intention were related as well. The mediation analysis showed that emotional responses were mediated by brand trust: rather than having a direct effect on purchase intention, negative emotional responses lower brand trust, which in turn lowers purchase intention. Therefore, marketers should consider not only whether consumers recognize AI-generated imagery, but also whether the imagery produces emotional responses that affect their relationship with the brand.

AI Disclosure and Implications

The desire for AI disclosure varied strongly by age - older participants were mainly indifferent, while younger participants were more supportive of disclosure. However, younger respondents also demonstrated a significantly lower brand evaluation post-AI disclosure. Older respondents, on the other hand, maintained their initial evaluation, with some even raising it. This brings up a dilemma for marketers- do the benefits of AI use outweigh the costs?

From a purely marketing perspective, nondisclosure may sometimes preserve brand evaluations. From an ethical perspective, however, prioritizing potentially higher brand evaluations by withholding information is unfair to consumers. Several respondents in our study cited uneasiness at the thought of being shown images that don’t accurately depict the products

they are looking at if AI use was disclosed, even though they were in favor of AI disclosure. If consumers already experience greater uneasiness toward AI-generated imagery, hiding AI use may preserve short-term evaluations but creates a transparency problem if consumers later discover that the imagery was synthetic. To many companies, the solution seems clear: don't disclose AI use, and consumers will maintain their initial brand evaluation.

Still, many ethical arguments favor AI disclosure in AI-generated advertisements. It provides consumers with greater transparency, allowing them to make informed judgements about the content they are being shown. Additionally, studies have demonstrated that while marginally significant, humans tend to have more negative emotional evaluations of AI images than human-made ones, which incentivizes marketers to use human content (Rapp et al.).

Consumers deserve transparent, nondeceptive information about the products they are shown, including AI disclosure and responsible representation. According to a study conducted by Alleva et al., women who viewed undisclosed AI-generated models that depicted a 'fit ideal' reported significantly lower appearance satisfaction and body appreciation than women who viewed body-diverse human images. However, in the third group of people who were shown AI images with disclosure, body appreciation was higher - although it did not reach the level of the women viewing body-diverse images. This suggests that exposure to idealized AI imagery may negatively affect self-perception even when consumers are aware that the images are AI-generated, although disclosure may partially mitigate these effects. Therefore, the question for marketers should not simply be whether AI use should be disclosed, but how AI-generated content can be implemented transparently without compromising accurate product representation or consumer well-being. AI disclosure alone will not erase the negative emotional and brand evaluations. However, a combination of AI disclosure and realistic models that represent a variety of customers can help mitigate these negative effects and provide consumers with the representation and information they need to feel confident with the products they are purchasing. Rather than using AI to create "ideal" models, if fashion companies are to use AI, they should focus on accurate product representation and showcasing diverse models so consumers can make informed purchasing decisions without being exposed to unrealistic standards of appearance.

Broader Implications

The findings of this study have implications for both consumers and marketers as generative AI becomes increasingly integrated into visual advertising. Although AI-generated imagery offers brands a cost-effective and flexible alternative to traditional photography, the results suggest that its use can influence perceptions of authenticity, emotional responses, trust, and brand evaluations. Because participants were generally unable to consistently distinguish AI-generated images from human-created ones, consumers will increasingly rely on AI disclosure rather than self-identification to determine the origins of an advertisement.

This study also helps address gaps in the existing literature by examining AI-generated fashion imagery across multiple consumer outcomes within a single study, including detection ability, emotional response, perceived authenticity, brand trust, purchase intention, and responses

to AI disclosure. Additionally, the study examines how age, gender, and familiarity with AI technology may shape responses, providing a more nuanced understanding of individual differences in consumer reactions to AI advertising.

For marketers, these findings suggest that the benefits of AI-generated advertising should be weighed against potential downsides for consumer trust and well-being. Transparency, accurate product representation, and responsible use of AI-generated models may become increasingly important as the technology becomes more prevalent. Ultimately, AI should not be viewed as inherently beneficial or harmful to advertising; instead, its effectiveness and ethical implications depend on how it is implemented and communicated to customers.

Limitations

A primary limitation of this study is its relatively small sample size (N = 50), with only 25 participants per experimental condition. While this sample was sufficient to identify several statistically significant relationships, the limited sample size reduces statistical power and may have limited the ability to detect smaller effects. This limitation is particularly relevant for the regression and mediation analyses, which typically require larger samples depending on the number of predictors and expected effect sizes (Green; Schoemann et al.). Additionally, since there were significantly more women in the study than men, the generalizability of the observed gender differences is lower. Age is also unevenly distributed, with only 26% of respondents being over 26.5 years old. Finally, this study features a convenience sample- a group of participants who were chosen due to convenience and proximity, meaning they are not necessarily reflective of the demographics and opinions of the broader population.

Conclusion

This study demonstrates that AI-generated fashion imagery can influence consumers in ways that extend beyond their ability to identify whether an image was created by AI. Although participants were generally unable to consistently distinguish AI-generated images from human-created images, their emotional responses, perceptions of brands, and purchasing intentions were also important considerations. The findings suggest that brand trust may play a role in connecting consumers' emotional responses to their purchase intentions, while familiarity with AI may influence how consumers respond to AI disclosure. These findings highlight a dilemma for marketers: AI can provide greater efficiency and creative flexibility, but its use may also create concerns surrounding authenticity, transparency, and consumer well-being. As generative AI continues to develop, fashion companies should consider not only the effectiveness of AI-generated advertising, but also whether its use provides consumers with accurate, transparent, and responsible representations of their products.

Works Cited

"Welcome to Aerie." Aerie, American Eagle Outfitters, 2026, Aerie Real webpage

Alleva, Jessica M., et al. "The Effects of Exposure to Imagery Created with Generative Artificial Intelligence (Gen-AI) on Young Women's Body Image: Do Image Type and Disclosure Matter?" Computers in Human Behavior: Artificial Humans, vol. 9, 2026, article 100339. https://doi.org/10.1016/j.chbah.2026.100339.

Avery, James J., et al. "Attributing AI Authorship: Towards a System of Icons for Legal and Ethical Disclosure." Northwestern Journal of Technology and Intellectual Property, vol. 22, no. 1, 2024, pp. 1–54.

Diwanji, Viraj S., et al. "Consumers' Emotional Responses to AI-Generated versus Human-Generated Content: The Role of Perceived Agency, Affect and Gaze in Health Marketing." International Journal of Human-Computer Interaction, vol. 41, no. 19, 2025, pp. 12270–12290. https://doi.org/10.1080/10447318.2025.2454954.

Ford, James, et al. "AI Advertising: An Overview and Guidelines." Journal of Business Research, vol. 166, 2023, article 114124. https://doi.org/10.1016/j.jbusres.2023.114124.

Green, Samuel B. "How Many Subjects Does It Take to Do a Regression Analysis?" Multivariate Behavioral Research, vol. 26, no. 3, 1991, pp. 499–510. https://doi.org/10.1207/S15327906MBR2603_7.

Herdimansyah, A. H. "Is AI Truly in Style?: A Case Study of the Public Reactions to AI-Generated Fashion Models." Proceedings of the Bakrie International Conference on Communication, Management, Politics & Accounting (BICOMPACT), 2025.

Högemann, Maximilian, et al. "What You See Is Not What You Get Anymore: A Mixed-Methods Approach on Human Perception of AI-Generated Images." Frontiers in Artificial Intelligence, vol. 8, 2025, article 1707336. https://doi.org/10.3389/frai.2025.1707336.

Ivan, Lorena. "Visual Deception in Online Dating: How Gender Shapes AI-Generated Image Detection." Computers in Human Behavior: Artificial Humans, vol. 6, 2025, article 100208. https://doi.org/10.1016/j.chbah.2025.100208.

Jung, T., et al. "Artificial Creativity in Luxury Advertising: How Trust and Perceived Humanness Drive Consumer Response to AI-Generated Content." Journal of Retailing and Consumer Services, vol. 87, 2025, article 104403. https://doi.org/10.1016/j.jretconser.2025.104403.

Kirk, Colleen P., and J. Givi. "The AI-Authorship Effect: Understanding Authenticity, Moral Disgust, and Consumer Responses to AI-Generated Marketing Communications." Journal of Business Research, vol. 186, 2025, article 114984. https://doi.org/10.1016/j.jbusres.2024.114984.

Kishnani, D. K. "The Uncanny Valley: An Empirical Study on Human Perceptions of AI-Generated Text and Images." Massachusetts Institute of Technology, 2025. Master's thesis. https://hdl.handle.net/1721.1/159096.

Lee, C. H. "Can Consumers Spot AI and Real Photos in 2024?" Conjointly, 17 Oct. 2024.

Napoli, Julie, et al. “Measuring Consumer-Based Brand Authenticity.” Journal of Business Research, vol. 67, no. 6, 2014, pp. 1090–1098. https://doi.org/10.1016/j.jbusres.2013.06.001.

Park, J., et al. “Traditional vs. AI-Generated Brand Personalities: Impact on Brand Preference and Purchase Intention.” Journal of Retailing and Consumer Services, vol. 81, 2024, article 104009. https://doi.org/10.1016/j.jretconser.2024.104009.

Qin, M., et al. “Falsity Images Also Have Real Benefits? The Impact of AI-Generated Advertisements’ Visual Styles and Reach Methods on Consumer Consultation Intentions.” Journal of Retailing and Consumer Services, vol. 91, 2026, article 104772. https://doi.org/10.1016/j.jretconser.2026.104772.

Rapp, A., et al. “How Do People Experience the Images Created by Generative Artificial Intelligence? An Exploration of People’s Perceptions, Appraisals, and Emotions Related to a Gen-AI Text-to-Image Model and Its Creations.” International Journal of Human-Computer Studies, vol. 193, 2025, article 103375. https://doi.org/10.1016/j.ijhcs.2024.103375.

Reid, Hannah. “Zalando Uses AI to Speed Up Marketing Campaigns, Cut Costs.” Reuters, 7 May 2025.

Richins, Marsha L. “Measuring Emotions in the Consumption Experience.” Journal of Consumer Research, vol. 24, no. 2, 1997, pp. 127–146. https://doi.org/10.1086/209499.

Schoemann, Alexander M., et al. “Determining Power and Sample Size for Simple and Complex Mediation Models.” Social Psychological and Personality Science, vol. 8, no. 4, 2017, pp. 379–386. https://doi.org/10.1177/1948550617715068.

Shi, M., et al. “Understanding Purchase Intention for GenAI-Enabled Museum Cultural and Creative Products Using a SOR Model.” Scientific Reports, vol. 16, 2026, article 5858. https://doi.org/10.1038/s41598-026-36224-8.

Tavakol, Mohsen, and Reg Dennick. “Making Sense of Cronbach’s Alpha.” International Journal of Medical Education, vol. 2, 2011, pp. 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd.

Yıldırım, E., and F. Dursun. “Disclosure Matters: Perceived Manipulation, Perceived Ethics, and Purchase Intention toward AI Influencers in Social Media Marketing.” Journal of Theoretical and Applied Electronic Commerce Research, vol. 21, no. 6, 2026, article 194. https://doi.org/10.3390/jtaer21060194.

Zhang, L., and C. Hur. “The Impact of Generative AI Images on Consumer Attitudes in Advertising.” Administrative Sciences, vol. 15, no. 10, 2025, article 395. https://doi.org/10.3390/admsci15100395.

What Are the Best Economic Predictors for Armed Conflict and Peace?
By Wang Qisheng

Introduction

Armed violence poses a major impediment to human progress, eroding human development and keeping societies locked into cycles of poverty and instability. In 2024 alone, the Uppsala Conflict Data Program saw 61 active state-based armed conflicts worldwide, the highest number recorded since data collection began in 1946, and roughly 160,000 people died from organized violence. By 2025, the death toll had climbed to approximately 244,600, making it the second-bloodiest year globally since the Rwandan genocide of 1994. This violence has driven displacement to unprecedented levels: by the end of 2024, 123.2 million people worldwide had been forcibly displaced by conflict, persecution, and human rights violations, marking roughly a decade of consecutive year-on-year increases. This runs counter to a common post-Cold War assumption that armed conflict has been in long-term decline: UCDP data shows the number of active state-based conflicts actually rose steadily from 2010 to 2023, reversing decades of relative stability. The effect of armed violence can also be seen on infrastructure. According to the Fifth Rapid Damage and Needs Assessment (RDNA5), as of 2025, direct damage in Ukraine had reached over $195 billion, with the housing, transport, and energy sectors bearing the brunt of the destruction. Looking beyond the damage already sustained, the report estimates that full recovery and reconstruction will cost Ukraine almost $588 billion over the next decade, a sum nearly three times the country's projected 2025 GDP. Sector-by-sector, the greatest long-term needs lie in transport (over $96 billion), energy (nearly $91 billion), and housing (almost $90 billion), underscoring how thoroughly the war has hollowed out the infrastructure that ordinary life depends on. Together, these figures underscore that armed conflict is not a distant or abstract threat, but a defining and worsening feature of the present global landscape, one whose costs are measured not only in lives lost but in the resources diverted from building a more peaceful world. These figures raise a pressing policy question: research from the World Bank and United Nations' *Pathways for Peace* study finds that investment in conflict prevention - including institutional reforms to strengthen governance and the rule of law, economic measures like targeted development aid and employment programmes, is dramatically more cost-effective than crisis response, with scaled-up preventive action estimated to save the global community between $5 billion and $70 billion annually. Hence, if the economic predictors of conflict can be identified and tracked early, targeted prevention could avert both the human toll and the trillions of dollars conflict currently consumes - a case for shifting resources from reaction to prevention.

The challenge, for scholars and policy makers as well, is to identify which economic factors breed conflict and which sustain peace. Despite decades of economic globalisation and institutional reform, neither prosperity nor integration has proven sufficient to secure peace. On the one hand, economic growth has expanded across much of the world and markets are more interconnected than ever. On the other hand, civil wars and interstate tensions persist, emerging

across markedly different economic contexts. This variation is evident in how similar conditions produce divergent outcomes: resource-rich states, for instance, may achieve stability while others descend into predation. Likewise, economic downturns can prompt institutional reform, yet in different settings they instead trigger unrest and rebellion. Even deep trade interdependence, often assumed to deter conflict, does not consistently prevent escalation. Taken together, these patterns suggest that neither prosperity nor integration alone can reliably secure peace.

This article argues that economic predictors of conflict operate through a set of interdependent processes rather than isolated variables. Economic contraction alters individual incentives; entrenched inequalities structure causes collective grievance; concentrated rents affect both state capacity and the feasibility of violent competition; and systemic interdependence reshapes the external calculus of escalation. None of these forces is decisive on its own. Their impact depends on the institutional arrangements through which societies absorb, redistribute, or politicize economic pressure. Peace, in this sense, is not simply the by-product of growth but the outcome of economic structures embedded within resilient and inclusive governance.

The sections that follow develop this framework in stages, examining macroeconomic instability, horizontal inequalities, resource dependence, economic diversification, and trade interdependence before returning to the central role of institutional mediation.

Section I- Macroeconomic Instability and the Opportunity Cost of Violence

Macroeconomic instability alters not only material welfare but the strategic environment within which political actors calculate risk, reward, and survival. Declining incomes, stalled growth, inflationary volatility, and fiscal crisis reshape the relative returns to participation in legal markets versus armed mobilization. When formal economic prospects deteriorate, the opportunity cost of rebellion falls; violence becomes economically conceivable even where ideological motivations remain constant. Instability, in this sense, operates as a structural modifier of incentives rather than a direct trigger of conflict.

This logic underpins the opportunity-cost framework most prominently advanced by Paul Collier and Anke Hoeffler. Drawing on cross-national data covering over 160 countries between 1960 and 1999, they find that a doubling of GDP per capita reduces the risk of civil war onset by roughly half, while each additional percentage point of economic growth lowers conflict probability by about one percentage point. These results suggest that poverty and stagnation do not merely correlate with conflict but materially reduce the opportunity costs of insurgent participation. Subsequent work by Edward Miguel and collaborators strengthens this claim by exploiting exogenous income variation: in Sub-Saharan Africa, a 5 percent decline in economic growth – instructed through rainfall shocks – is associated with nearly a 50 percent increase in the likelihood of civil war. Taken together, these findings shift the analytical focus away from grievance alone and toward feasibility: rebellion emerges not only where dissatisfaction exists, but where economic conditions lower the costs of organization and participation.

Yet macroeconomic decline does more than recruit insurgents, it destabilizes the bargaining equilibrium between state and challenger. In standard bargaining models of conflict, war reflects a failure to reach a negotiated settlement under conditions of uncertainty or an inability to make credible future commitments. Economic contraction intensifies both. As revenues shrink, governments face reduced capacity to buy off opposition through patronage, redistribution, or credible reform commitments. Simultaneously, weakened fiscal positions may signal declining state strength, encouraging challengers to revise upward their expectations of victory. Information asymmetries therefore widen, causing both sides to face growing uncertainty about the other's true strength and resolve. A government facing fiscal contraction may struggle to credibly signal whether its weakness is temporary or terminal, while challengers cannot easily verify the state's remaining capacity to repress or co-opt opposition. This mutual uncertainty makes it harder for either side to accurately gauge the other's bargaining position, increasing the risk that one party miscalculates and opts for violence over a settlement both might otherwise prefer. Credible commitments hence become harder to sustain, the result is not automatic war but a narrowing of the negotiated space.

The societal effects of macroeconomic instability are equally consequential to the direct costs of war, and often function as a precondition for conflict rather than merely a consequence of it. In particular, high youth unemployment expands the pool of individuals for whom participation in armed groups offers a relative economic advantage. Where formal and informal labour markets are weak and opportunities for migration are limited, the range of viable livelihoods narrows considerably. Under such conditions, armed organisations can function as alternative employers, offering not only wages but also protection and a sense of social mobility otherwise unavailable. This dynamic was evident in Sierra Leone's civil war (1991–2002). The Revolutionary United Front drew heavily on unemployed, unskilled rural youth who had few alternatives beyond subsistence agriculture or informal and often illicit diamond mining. This pattern persisted after the war ended, with youth unemployment reaching an estimated 65% by 2009, continuing to drive instability even in peacetime. This suggests that economic exclusion functions as more than a background condition; it materially lowers the opportunity cost of joining an armed group, a relationship supported by broader research showing that unemployment reduces the relative cost of rebellion by supplying an abundant pool of recruits with little to lose. It is worth noting, however, that economic explanations alone are insufficient: interventions that improve employment and income in fragile settings have shown only limited effect on reducing violence or mobilisation, suggesting that unemployment is best understood as an enabling condition for conflict rather than its sole cause.

While Collier's opportunity-cost model explains the economic mechanism by which unemployment lowers the barrier to joining an armed group, it risks flattening the motivations involved into simple economic rationality. Paul Richards' fieldwork on Sierra Leone complicates this picture: rather than framing RUF recruitment as pure predation or greed, he argues the conflict emerged from a genuine rural crisis, land-poor, marginalised youth excluded from a collapsing patrimonial state, for whom joining the RUF offered not just income but status and

belonging otherwise unavailable to them. The two views are not mutually exclusive. Collier's model helps explain why unemployment expands the pool of potential recruits, while Richards' account explains why, for many young people in Sierra Leone, that choice was less a calculated economic decision than a response to systemic exclusion.

This dynamic was evident in Sierra Leone prior to the outbreak of civil war in 1991. During the 1980s, GDP per capita contracted sharply – falling by more than 50 percent over the decade – reflecting prolonged economic decline and state retrenchment. At the same time, youth unemployment was estimated to affect up to 60 percent of young people, leaving a large cohort economically marginalised and excluded from formal opportunity structures.

These conditions did not predetermine violence. Rather, they lowered the threshold for mobilisation, making recruitment easier once political entrepreneurs supplied the necessary organisation and narrative.

At the state level, macroeconomic fragility strikes at the very foundations of governance. As inflation erodes real public-sector wages, bureaucratic loyalty frays; as fiscal crisis deepens, military funding shrinks and intelligence networks atrophy. Hence, governments find themselves increasingly unable to credibly signal long-term concessions precisely when doing so matters most. The late-1980s collapse in Yugoslavia illustrates how quickly these pressures can become self-reinforcing. By 1989, inflation had reached approximately 2,700 percent, hollowing out the purchasing power of federal institutions and the officials who staffed them. Richer republics such as Slovenia and Croatia, watching their own revenues disappear into a depreciating federal budget, grew increasingly unwilling to subsidise poorer regions like Bosnia and Macedonia - transforming what had once been routine budgetary negotiations into contests over political survival. Simultaneously, a mounting external debt crisis forced the federal government into IMF austerity conditions that further stripped its capacity to offer meaningful concessions to any faction. The result was not war by design, but institutional disintegration by default: a federal authority too fiscally exhausted to hold together a state that economic pressure had already begun to pull apart.

Crucially, macroeconomic instability is permissive rather than deterministic - it reshapes the landscape within which political actors operate, but does not predetermine their choices. Ghana's experience in the late 1970s and early 1980s makes this clear. Inflation exceeded 100 percent and real incomes collapsed, yet rather than descending into violence, the Rawlings government responded with sweeping structural reforms under the IMF's Economic Recovery Programme from 1983, gradually restoring fiscal discipline and rebuilding state credibility. Economic stress was severe, but institutions - however imperfect - remained adaptive enough to absorb the pressure. Yugoslavia, as discussed, told a very different story: comparable fiscal strain did not encounter adaptive institutions but exposed fault lines that federal arrangements had long papered over, converting economic crises into political disintegration. Sri Lanka presents a third and perhaps most revealing variant. Despite comparatively higher income levels, the systematic exclusion of Tamil populations from political representation and formal employment - entrenched through discriminatory language policies and university admissions quotas - meant

that economic stress landed on communities already primed for mobilisation. Prosperity offered no buffer when the institutions distributing its benefits remained structurally exclusionary. What these three cases collectively reveal is not that economics is irrelevant, but that its consequences are always filtered through the political arrangements that surround it, a point to which this article will return.

To summarise my first point, macroeconomic fundamentals therefore constitute a foundational layer in the political economy of conflict. They shape individual incentives, alter bargaining dynamics, and condition state capacity. However, their effects depend on how economic stress is distributed and politicised-an issue that requires attention to horizontal inequalities and group-based exclusion, to which the next section turns.

Section II - Horizontal Inequalities and the Political Construction of Grievance

National income averages conceal the social architecture through which economic advantage and exclusion are experienced. Growth may occur while entire communities remain systematically marginalised; poverty may decline even as disparities between identity groups harden. It is within these patterned inequalities-distributed across ethnic, religious, or regional lines-that economic grievance becomes politically potent.

The analytical contribution of horizontal inequality theory, most closely associated with Frances Stewart, lies in its shift from individual deprivation to group-based comparison. Individuals rarely rebel simply because they are poor. Mobilisation becomes likely when economic disadvantage is shared, visible, and mapped onto socially meaningful identities. Inequality acquires political force when it is experienced collectively and interpreted as structural injustice rather than personal misfortune.

This mechanism operates through both perception and organisation. Disparities across groups shape everyday encounters with the state-access to land, education, employment, and political representation-gradually producing shared experiences of exclusion. Over time, these experiences can be framed and mobilised by political leaders, activists, or armed groups who translate economic inequalities into collective grievances and demands. Empirical research by Lars-Erik Cederman and colleagues provides strong support for this dynamic. Drawing on cross-national data on ethnic groups, they find that groups excluded from political power and with average incomes at least 20 percent below the national mean are significantly more likely to engage in rebellion. More broadly, their analysis shows that politically excluded groups face conflict risks up to two to three times higher than those included in state power. These findings suggest that conflict does not arise from inequality in the abstract, but from inequality that is structured along identity lines and reinforced through political exclusion.

The experience of Rwanda illustrates how distributional inequalities can become embedded within political and institutional structures. Prior to the 1994 genocide, Rwanda was one of the most densely populated countries in Africa, with over 300 people per square kilometre of arable land, placing intense pressure on scarce resources. Access to land, cattle ownership, and formal employment was unevenly distributed across ethnic groups, with Tutsi households

owning, on average, more cattle and holding a disproportionate share of salaried positions. At the same time, economic conditions were deteriorating: coffee prices—Rwanda's primary export—fell by nearly 50 percent in the late 1980s, while structural adjustment policies reduced public-sector employment and increased the cost of basic goods. These pressures intensified competition over land and livelihoods, particularly among younger men who found themselves unable to inherit land or secure formal employment. Crucially, however, violence did not emerge from scarcity alone. The Habyarimana regime and the northern elite clique around it—known as the akazu—responded to the 1990 invasion by the Rwandan Patriotic Front, and to the economic strain that followed, by recasting Tutsi civilians as a fifth column responsible for the country's hardship. This narrative was propagated through state and party channels: the newspaper Kangura published the "Hutu Ten Commandments" branding Tutsi as enemies of Hutu solidarity, while Radio Télévision Libre des Mille Collines broadcast messaging that linked economic grievance directly to ethnic exclusion. The hardline CDR party, formed in 1992 as a rival to Habyarimana's own MRND, pushed this framing further, and local administrators used their control over land allocation and labour mobilisation to recruit the un- and under-employed youth most affected by the economic crisis into the Interahamwe militia. In this context, participation in violence was not simply a response to poverty, but to the perception—actively constructed by these elites—that access to economic opportunity and state power was systematically denied along identity lines. The Rwandan case thus demonstrates that it is the politicisation of inequality, executed through identifiable propaganda and patronage networks, rather than deprivation in itself, that shapes the scale and intensity of conflict.

Horizontal inequalities also illuminate conflict in relatively affluent contexts. In Northern Ireland, sustained economic development coexisted with systematic disparities in housing, employment, and political representation between Protestant and Catholic communities. Prior to the late 1960s, Catholics were significantly underrepresented in public-sector jobs and local government, and faced discrimination in housing allocation, which in turn affected voting rights under property-based electoral systems. By the mid-1960s, unemployment rates among Catholics were estimated to be more than twice those of Protestants in some areas. These inequalities became politically salient during the civil rights movement, which began with peaceful protests demanding equal access to housing, jobs, and voting rights. However, the state’s heavy-handed response and the persistence of institutional bias escalated tensions, leading to riots and, eventually, the emergence of armed groups such as the Irish Republican Army (IRA). Violence thus did not arise from poverty in absolute terms, but from the perception of systematic exclusion within a relatively prosperous society.

Yet, the trajectory of such inequalities is not predetermined. Institutional design plays a central role in shaping whether disparities harden into conflict or are channelled into accommodation. Power-sharing arrangements, targeted redistribution, and inclusive political representation can transform structural grievance into negotiated compromise. Where institutions entrench exclusion, however, inequality becomes increasingly salient and mobilisable.

Horizontal inequalities therefore constitute a distinct causal pathway linking economic structure to political violence. Whereas macroeconomic instability alters individual opportunity structures, distributional asymmetries organise collective grievance and identity-based mobilisation. Economic growth may raise national income while leaving the underlying architecture of exclusion intact. When it does, the foundations of conflict remain firmly in place.

The analysis now turns to a third mechanism-concentrated economic rents-through which wealth itself can reshape incentives, institutions, and the feasibility of sustained violence.

Section III - Resource Rents, Predation, and the Political Economy of Conflict Financing

While macroeconomic instability reshapes incentives and horizontal inequalities structure grievance, neither alone explains the persistence and scale of many modern conflicts. Armed violence requires material capacity. Weapons must be purchased, fighters paid, supply networks maintained, and territorial control enforced. It is here that concentrated economic rents-particularly from extractive resources-play a decisive role.

Resource dependence changes how both governments and armed groups behave. In countries where wealth comes largely from easily extractable resources such as oil, diamonds, timber, or minerals, controlling territory often means controlling income. As a result, conflict is no longer just about political demands or policy differences; it becomes a struggle over access to these valuable resources. Armed groups fight not only to challenge the state, but to gain direct control over revenue sources. In this context, economic activity itself becomes tied to violence, as resource extraction is used to fund and sustain conflict.

The argument of easily extractable resource wealth can reshape both political incentives and institutional development in ways that increase conflict risk-forms the foundation of the resource curse literature. Early contributions by Jeffrey Sachs and Andrew Warner (1995) demonstrate that countries heavily dependent on primary commodity exports tend to experience slower long-term growth and weaker institutional development. Building on this, conflict-focused research by Michael Ross (2004) shows that resource-rich states-particularly those reliant on geographically concentrated and easily extractable commodities such as oil or diamonds-are significantly more prone to civil war, and that such conflicts tend to last longer. The key mechanism lies not in scarcity, but in how abundance is politically and economically organised. Where governments can rely on resource revenues rather than taxation, they face weaker pressure to remain accountable to citizens. At the same time, armed groups can finance their activities by capturing resource-rich territory, reducing their dependence on popular support. Resource wealth thus alters both the incentives to compete for power and the capacity to sustain violence over time.

Resource rents distort the fiscal foundations of the state. Governments that finance themselves through extraction rather than taxation face weaker incentives to build inclusive political coalitions or provide public goods. Accountability erodes as rulers become less dependent on citizen compliance. At the same time, concentrated wealth raises the stakes of state

capture. Control of government translates directly into control over lucrative revenue streams, intensifying elite competition and lowering the threshold for violent contestation.

For non-state actors, resource rents lower the barriers to sustained insurgency. Armed groups no longer rely solely on popular support or external sponsorship; they can self-finance through territorial control. Violence becomes economically viable over long horizons.

The civil war in Angola offers a stark illustration. Rebel forces financed operations through diamond extraction and smuggling, while the government relied on oil revenues to procure advanced weaponry and international backing. Both sides possessed independent revenue sources insulated from civilian taxation. As a result, neither faced strong economic pressure to compromise. What began as ideological struggle evolved into a protracted contest over rent control, extending conflict for nearly three decades.

A similar pattern can be seen in Sierra Leone, where diamond-rich regions quickly became central to the conflict. Armed groups took control of mining areas and used the proceeds to fund their operations, trading diamonds for weapons and supplies. In this context, violence was sustained not only by underlying grievances, but by the steady flow of income generated from these resources. Control over territory therefore meant more than strategic advantage-it provided the financial means to keep the war going.

These dynamics are compounded by global commodity volatility: price shocks translate directly into fiscal instability, elite fragmentation, and social unrest. Resource wealth may inflate national income statistics, but when gains remain concentrated and employment effects limited, growth masks structural fragility. Under such conditions, control of the state becomes synonymous with control of rents, intensifying political competition and raising the stakes of violent contestation.

Yet resource wealth is not inherently destabilising. Its effects are conditional on institutional management. In countries such as Norway, transparent revenue governance and long-term investment frameworks have converted extractive wealth into broad-based development rather than elite capture. Similarly, Botswana leveraged diamond revenues through accountable institutions and public investment, avoiding both conflict and economic distortion. These cases underscore that it is not abundance itself that fuels violence, but rent concentration operating through weak political constraints.

Resource rents therefore constitute a third mechanism in the political economy of conflict: they finance mobilisation while corroding institutional restraint. Where opportunity costs fall and grievances intensify, rents supply the means through which violence becomes durable. Where institutions channel resource wealth into inclusive development, the same revenues can reinforce stability.

The analysis now turns to the inverse dynamic-how economic diversification and structural transformation diffuse rents, broaden opportunity, and raise the long-term costs of conflict.

Section IV - Economic Diversification and the Structural Foundations of Peace

If concentrated rents, volatility, and narrow production bases intensify conflict risk, economic diversification operates in the opposite direction. By broadening sources of income, employment, and state revenue, diversified economies reduce exposure to commodity shocks, diffuse economic power, and raise the opportunity costs of violence. Stability emerges not from growth alone, but from the structural composition of production.

From a political economy perspective, diversification transforms both incentives and institutions. Where economic activity is spread across manufacturing, services, and technology-intensive sectors, livelihoods become less vulnerable to sudden price swings. Fiscal revenues stabilize as governments draw from multiple tax bases rather than a single extractive stream. Employment expands in labor-absorbing industries, integrating broader segments of the population into formal markets. Each of these dynamics strengthens the social and fiscal foundations of the state while narrowing the appeal of armed mobilisation.

This process also alters the distribution of power. In rent-dependent economies, wealth concentrates around resource control, reinforcing elite competition and patronage politics. Diversified economies, by contrast, generate dispersed sources of income and investment, weakening the dominance of narrow rent-holding coalitions. Economic complexity fosters plural economic interests that favor predictability, rule enforcement, and institutional stability. Violence becomes increasingly costly to both firms and workers embedded in long-term productive relationships.

Historical development trajectories illustrate this shift most clearly when examined through specific country experiences. In South Korea, for example, the transition from an agrarian, war-torn economy in the 1950s to an export-oriented industrial power by the 1980s was driven by deliberate state investment in manufacturing, education, and infrastructure. Employment expanded rapidly in labour-intensive industries such as textiles and electronics, absorbing surplus rural labour and steadily raising household incomes. Over time, this industrial base upgraded into capital- and skill-intensive sectors, including automobiles and semiconductors, embedding large segments of the population into stable, formal employment. A similar trajectory can be observed in Taiwan, where land reform and small- and medium-enterprise development facilitated broad-based participation in export growth.

These cases demonstrate that diversification did more than increase output; it reshaped the structure of opportunity. By expanding access to employment and integrating workers into global production networks, industrialisation raised the opportunity cost of violence and reduced the appeal of insurgent activity. This economic explanation, however, cannot be fully separated from the political context in which it occurred. In South Korea, Taiwan, and Suharto's Indonesia, diversification proceeded under authoritarian regimes that combined developmental policy with the suppression of independent labour organisation, opposition parties, and ethnic or regional mobilisation. Political stability was therefore not purely a by-product of rising incomes; it was actively enforced, which itself constrained the channels through which grievance could organise into violence. Diversification and authoritarian control are best understood as complementary rather than competing explanations: industrial policy expanded the economic base broadly

enough to prevent large, marginalised populations from forming, while state repression foreclosed the political space in which any residual grievance might otherwise have mobilised. In this sense, the relative peace of these cases reflects both the distribution of economic participation across society and the coercive capacity of the state to manage the discontent that distribution did not fully resolve.

The stabilising effects of diversification extend beyond income levels. Broader production bases enhance state capacity by expanding taxable activity, allowing governments to finance public goods without reliance on volatile rents. Education systems, infrastructure investment, and regulatory institutions become economically necessary rather than politically optional. Over time, these investments reinforce institutional credibility and social trust - central buffers against conflict escalation during economic downturns.

Diversification also mitigates the volatility channel central to the resource curse. When export revenue is drawn from multiple sectors, shocks in one market do not immediately destabilise the entire economy. Fiscal planning becomes more predictable; social spending less cyclical; elite coalitions less prone to fragmentation during price collapses. Economic stress still occurs, but it is absorbed rather than amplified.

Yet diversification alone is not sufficient. Growth that remains spatially or socially concentrated can reproduce the same distributional grievances identified earlier. In Nigeria, for instance, periods of industrial and oil-driven expansion have coexisted with stark regional disparities, particularly between the more developed south and the economically marginalised north. Despite rising national income, limited job creation and uneven access to infrastructure have contributed to persistent unrest and insurgency in poorer regions. A similar pattern can be observed in China, where rapid industrialisation and export-led growth have been heavily concentrated in coastal provinces. While overall GDP has surged, inland regions have lagged behind, generating significant rural–urban and regional inequalities that have required continuous state intervention to manage social tensions. These cases suggest that structural transformation contributes to peace most effectively when it is broadly inclusive. Where access to employment, education, and infrastructure remains uneven, growth may alleviate poverty in aggregate terms while leaving underlying sources of grievance intact.

The argument now extends beyond domestic structure to the international system. Just as diversified production reshapes internal incentives, integration into cross-border trade and financial networks alters the external costs of conflict - the focus of the next section.

Section V - Interdependence, Global Markets, and the Conditional Logic of Commercial Peace

Economic structures shape conflict not only within states but also across the international system. As production networks, financial markets, and supply chains have become increasingly integrated, the costs associated with interstate conflict have expanded beyond territorial loss or military expenditure. War now disrupts investment flows, trade networks, technological

exchange, and global production systems. Under conditions of deep interdependence, economic relations themselves become constraints on escalation.

This insight forms the foundation of the commercial peace tradition. Liberal theories of interdependence, empirically developed by scholars such as John Oneal and Bruce Russett, argue that dense trade relationships increase the opportunity costs of conflict by linking national prosperity to external cooperation. Firms dependent on export markets, investors exposed to financial volatility, and workers embedded in global value chains all face substantial losses in the event of war. Economic integration therefore generates domestic constituencies with strong incentives to resist escalation.

More recent political economy research extends this logic beyond trade volumes alone. Financial openness and capital mobility amplify restraint by making economies immediately sensitive to geopolitical risk. Even the anticipation of conflict can trigger capital flight, currency depreciation, and investment withdrawal, imposing rapid economic penalties before hostilities begin. Interdependence thus operates through expectations as much as material disruption; the threat of instability can be economically costly enough to shape strategic decision-making.

Yet the pacifying effects of interdependence are neither automatic nor uniform. Economic ties may deter conflict when dependence is mutual, but asymmetric relationships can instead generate vulnerability and coercion. States able to weaponize trade access, energy supplies, or technological infrastructure may transform interdependence into geopolitical leverage. Under such conditions, economic integration ceases to function as a shared constraint and instead becomes an arena of strategic competition.

The relationship between Russia and Ukraine prior to 2022 illustrates this limitation. Despite extensive trade linkages and energy interdependence, geopolitical and security considerations ultimately outweighed economic costs. Rather than preventing conflict, interdependence exposed channels through which sanctions, financial exclusion, and energy disruptions became instruments of warfare. Economic integration altered the form of conflict without eliminating its possibility.

A similar dynamic characterizes cases of asymmetric economic dependence, where interdependence generates vulnerability rather than stability. In the relationship between Palestinian territories and Israel, heavy reliance on access to markets, employment, and essential infrastructure has not produced political stability. Instead, unequal dependence has reinforced perceptions of constraint and economic coercion, deepening rather than alleviating underlying tensions. A comparable, though distinct, pattern can be observed in cross-strait relations between Taiwan and China. Despite extensive trade and investment linkages-particularly Taiwan's integration into mainland manufacturing supply chains-economic interdependence has not resolved core political disputes. Rather, it has created a complex mix of mutual benefit and strategic vulnerability, in which economic ties coexist with persistent security concerns. Taken together, these cases illustrate that where interdependence is asymmetric or politically contested, economic integration may deepen insecurity rather than foster cooperation, especially in the absence of credible institutional mediation.

Conversely, regions characterized by dense and relatively balanced production networks demonstrate more durable stabilizing effects. In East Asia, cross-border supply chains linking manufacturing, finance, and technology sectors have increased the economic costs associated with military escalation among major trading partners. Although geopolitical tensions persist, deeply embedded commercial ties raise the risks of disruption for multiple actors simultaneously, encouraging restraint even amid strategic rivalry.

Interdependence therefore functions as a conditional mechanism of peace. It raises the external costs of violence when economic dependence is reciprocal, diversified, and institutionally embedded. Where dependence is asymmetric or securitized, however, economic ties may intensify competition rather than mitigate it. The contemporary emergence of "geo-economic conflict"-in which trade, finance, and technology flows become instruments of strategic pressure--demonstrates that globalization reshapes conflict dynamics without resolving underlying political disputes.

Conclusion

Across the preceding analysis, a consistent pattern emerges: no economic condition produces conflict in isolation. Macroeconomic crises do not inevitably generate rebellion; inequality does not always culminate in violence; resource wealth may sustain predation in some contexts while supporting prosperity in others; and economic interdependence can either restrain or intensify geopolitical rivalry. These divergent outcomes point toward a mediating variable that conditions the political consequences of economic structure-institutions.

Institutions shape how economic pressures are distributed, interpreted, and managed. States possessing credible administrative capacity and inclusive political arrangements are better able to absorb economic shocks through redistribution, negotiation, and policy adjustment. Fiscal downturns may prompt reform rather than collapse; social grievances can be channelled into electoral or bureaucratic processes rather than armed mobilisation. Economic stress remains present, but institutionalized mechanisms reduce the incentives for violent escalation.

Institutional mediation operates most visibly through the management of revenue and distribution. In resource-dependent economies, transparent fiscal frameworks and long-term investment strategies can prevent rent concentration from undermining accountability. The experience of Norway demonstrates how sovereign wealth management and rule—bound governance convert volatile resource income into stabilizing public investment. By contrast, where institutions permit discretionary control over rents, competition for state authority becomes existential, increasing the likelihood of violent contestation. Economic structure alone does not determine outcomes; governance determines whether wealth reinforces cooperation or fuels predation.

Institutions also influence how societies respond to inequality. Inclusive political systems capable of representation and redistribution reduce the salience of horizontal disparities by providing non-violent avenues for contestation. Power-sharing arrangements, decentralized governance, and credible legal protections transform group-based grievances into bargaining

processes rather than security dilemmas. Where such mechanisms are absent, economic exclusion more readily translates into identity-based mobilisation.

At the international level, institutional frameworks similarly condition the effects of interdependence. Trade networks embedded within predictable legal regimes and multilateral coordination mechanisms generate mutual expectations of restraint. In contrast, economic relations lacking institutional safeguards remain vulnerable to securitization and coercion. Interdependence constrains conflict only when supported by governance structures capable of managing asymmetry and uncertainty.

Taken together, these dynamics suggest that economic predictors function probabilistically rather than deterministically. Economic shocks alter incentives; inequalities structure grievances; rents enable mobilisation; diversification and trade reshape opportunity costs. Yet institutions determine whether these pressures culminate in accommodation or rupture. Peace, therefore, cannot be reduced to prosperity alone. It emerges from the interaction between economic structure and institutional capacity.

As global economic conditions become increasingly volatile-shaped by climate change, technological disruption, and fragmenting trade relations-the importance of this interactional perspective becomes more pronounced. Preventing conflict is therefore less about eliminating economic stress than about building institutions capable of managing it. Identifying economic predictors of war is inseparable from understanding the political structures through which they operate. The central challenge for policymakers is not simply to promote growth, but to shape economic systems resilient enough to sustain peace under conditions of uncertainty and change.

Works Cited

Acemoglu, Daron, Simon Johnson, and James A. Robinson. "An African Success Story: Botswana." In Search of Prosperity: Analytic Narratives on Economic Growth, edited by Dani Rodrik, Princeton UP, 2003, pp. 80–119.

Amsden, Alice H. Asia's Next Giant: South Korea and Late Industrialization. Oxford UP, 1989.

Cederman, Lars-Erik, Kristian Skrede Gleditsch, and Halvard Buhaug. Inequality, Grievances, and Civil War. Cambridge UP, 2013.

Collier, Paul, and Anke Hoeffler. "Greed and Grievance in Civil War." Oxford Economic Papers, vol. 56, no. 4, 2004, pp. 563–95.

David, James Ojochenemi. "Boko Haram and Niger Delta Avengers: Unraveling the Greed-Need-Creed Spectrum in Nigeria's Security-Development Nexus." Frontiers in Political Science, vol. 7, 2025, doi:10.3389/fpos.2025.1562472.

Davies, Shawn, Therése Pettersson, Margareta Sollenberg, and Magnus Öberg. "Organized Violence 1989–2024, and the Challenges of Identifying Civilian Victims." Journal of Peace Research, 2025, doi:10.1177/00223433251345636.

Davies, Shawn, Therése Pettersson, and Magnus Öberg. "Organized Violence 1989–2025, and Violent Political Protests." Journal of Peace Research, vol. 63, no. 4, 2026, doi:10.1093/jopres/xjag046.

Des Forges, Alison. Leave None to Tell the Story: Genocide in Rwanda. Human Rights Watch, 1999.

Fan, C. Cindy, and Mingjie Sun. "Regional Inequality in China, 1978–2006." Eurasian Geography and Economics, vol. 49, no. 1, 2008, pp. 1–20.

Global Witness. A Rough Trade: The Role of Companies and Governments in the Angolan Conflict. Global Witness, 1998.

International Crisis Group. Chokehold: Countering Israel's Grip on the West Bank Economy. Middle East and North Africa Report no. 255, International Crisis Group, 2026.

Larsen, Erling Røed. "Escaping the Resource Curse and the Dutch Disease? When and Why Norway Caught Up with and Forged Ahead of Its Neighbors." American Journal of Economics and Sociology, vol. 65, no. 3, 2006, pp. 605–40.

Mamdani, Mahmood. When Victims Become Killers: Colonialism, Nativism, and the Genocide in Rwanda. Princeton UP, 2001.

Mark, Jeremy, and Niels Graham. Relying on Old Enemies: The Challenge of Taiwan's Economic Ties to China. Atlantic Council, 17 Nov. 2023.

Miguel, Edward, Shanker Satyanath, and Ernest Sergenti. "Economic Shocks and Civil Conflict: An Instrumental Variables Approach." Journal of Political Economy, vol. 112, no. 4, 2004, pp. 725–53.

Oneal, John R., and Bruce Russett. Triangulating Peace: Democracy, Interdependence, and International Organizations. Norton, 2001.

Richards, Paul. Fighting for the Rain Forest: War, Youth and Resources in Sierra Leone. James Currey, 1996.

Ross, Michael L. "What Do We Know about Natural Resources and the Civil War?" Journal of Peace Research, vol. 41, no. 3, 2004, pp. 337–56.
Sachs, Jeffrey D., and Andrew M. Warner. "Natural Resource Abundance and Economic Growth." National Bureau of Economic Research, Working Paper 5398, Dec. 1995.
Skalamera, Morena. "The Geopolitics of Energy after the Invasion of Ukraine." The Washington Quarterly, vol. 46, no. 1, 2023, pp. 7–24.
Smillie, Ian, Lansana Gberie, and Ralph Hazleton. The Heart of the Matter: Sierra Leone, Diamonds and Human Security. Partnership Africa Canada, 2000.
Stewart, Frances, editor. Horizontal Inequalities and Conflict: Understanding Group Violence in Multiethnic Societies. Palgrave Macmillan, 2008.
Straus, Scott. The Order of Genocide: Race, Power, and War in Rwanda. Cornell UP, 2006.
United Nations High Commissioner for Refugees. Global Trends: Forced Displacement in 2024. UNHCR, 2025, www.unhcr.org/global-trends-report-2024.
Wade, Robert. Governing the Market: Economic Theory and the Role of Government in East Asian Industrialization. Princeton UP, 1990.
World Bank and United Nations. Pathways for Peace: Inclusive Approaches to Preventing Violent Conflict. World Bank, 2018.
World Bank, Government of Ukraine, European Commission, and United Nations. Ukraine: Fifth Rapid Damage and Needs Assessment (RDNA5), February 2022–December 2025. World Bank, Feb. 2026.

The Shahed Drone and the Economics of Modern Warfare By Siddhanth Dayanand

Abstract

Iran's Shahed-136 loitering munition, developed as a low-cost answer to decades of sanctions and military isolation, has become one of the most consequential weapons of recent conflicts. This paper traces the drone's trajectory from its origins in Iran's asymmetric-warfare strategy, through its proving grounds in the Middle East, to its industrial-scale adoption by Russia in Ukraine and its return to the Middle East in Iran's direct confrontations with Israel and the 2026 war with the United States and Israel. It argues that the Shahed's significance lies less in its destructive power—which prepared, technologically advanced defenders have repeatedly blunted—than in the economics it imposes: a weapon cheap enough to launch in mass forces defenders into unfavorable cost exchanges. The paper also examines the doctrine's real limitations, arguing that cost-imposition has proven disruptive but not decisive.

Introduction

Today, drones are crucial weapons in modern military technology and a familiar presence on the battlefield. Originally serving as expensive platforms for surveillance and reconnaissance, drones now function as key intelligence-gathering tools, precision strike weapons, and long-range attack systems. Out of all the different types of drone systems developed and fielded over the last twenty years, the one that stands out most is Iran's Shahed series of loitering munitions. Originally designed as low-cost precision weaponry to make up for Iran's military limitations, the Shahed drone is now utilized in virtually every conflict in the Middle East and Eastern Europe.

The Shahed drone's emergence signals larger transitions in both global politics and military technology, indicating a move away from historically defined militaries as nations begin exploring economically viable alternatives to traditional missiles and fighter aircraft, which have historically been extremely expensive. Due to decades of restrictions and sanctions that prevented Iran from procuring advanced military equipment, the country invested considerable resources into developing domestically produced drones. These efforts eventually produced the Shahed-136, which can fly long distances and conduct precision attacks on multiple types of targets at a fraction of the cost of traditional missile systems.

Following the Russian military offensive against Ukraine, global attention toward these drones exploded. Images of Shahed drones destroying targets in Ukraine transformed how these weapons are perceived, as they were previously associated almost exclusively with conflicts in the Middle East. Today, the Shahed drone is an important military asset for Iran while also symbolizing the growing utilization of unmanned systems in modern warfare. Its use in Ukraine demonstrates the strategic implications that low-cost technologies can have on battlefields around the globe.

Origins of Iran's Drone Program

Iran's drone program began during the Iran-Iraq War (1980-1988). During this time, Iran's military equipment was becoming outdated due to technological challenges as well as shortcomings resulting from the 1979 Iranian Revolution, which limited access to advanced weapon systems and interfered with military sales to Iran. Iranian military leaders therefore began developing the ability to produce their own military technologies.

One of the ways they accomplished this was through unmanned aerial vehicles (UAVs), which could be manufactured domestically using fewer resources than modern manned fighter aircraft while still conducting surveillance missions without risking human life. Between the 1980s and the mid-1990s, Iran made significant investments in research institutions, defense manufacturers, and military organizations focused on drone development, producing systems such as the Mohajer-1 in 1985 and the Ababil-1 the following year (M. Rubin). Although these early drones were mostly basic models used for surveillance purposes, the experience gained from working with these systems greatly aided the development of future drone programs.

While reconnaissance was originally the primary focus, Iranian military leadership eventually saw the versatility of drones as a means of attack. As early as the 2000s, Iranian engineers began producing more advanced UAVs that offered greater range, endurance, and accuracy than previous generations. This culminated in the Karrar prototype—developed in 2009—which was immediately followed by several new UAV platforms (M. Rubin). This effort to develop an increasingly wide and sophisticated range of UAVs intentionally aligned with Iran's broader strategy of asymmetric warfare, designed to counter technologically superior military opponents through cost-effective and creative means (M. Rubin).

The Shahed series was developed in this context by Iranian defense agencies. Rather than attempting to match the technological sophistication of their Western counterparts, the drones took years of technological experimentation to create a system that was reliable, inexpensive, and capable of being mass-produced.

Creation of the Shahed-136

The Shahed-136 drone is a prominent delivery system among the many developed in recent years by Iranian manufacturers. Unlike conventional reusable drones that return to their launch site after completing a mission, loitering munitions such as the Shahed-136 (often called "kamikaze" or "suicide" drones) serve the role of both missile and drone.

A loitering munition has the capability to hang around in a slow, controlled manner, allowing for much greater mission flexibility than, for example, a missile whose target is predetermined and, once launched, cannot be changed. In this context, greater mission flexibility could include bailing out before attacking the planned target, changing the target, or readjusting flight to improve effectiveness against the intended target. It is worth noting that this classification is not universally accepted: analysts at the Royal United Services Institute have questioned whether the Shahed-136 truly loiters, arguing that it has mainly been used for point-to-point strikes more akin to a cheap cruise missile than a true loitering munition.

Reflecting Iranian values of low cost and simplicity, the Shahed-136 features a familiar triangular wing design and utilizes as many commercially available parts as possible. Iran noticeably sources a great number of components from non-Iranian, Western manufacturers, with analysts at Conflict Armament Research finding that up to 82% of the components in an examined drone were manufactured by companies based in the United States (Bertrand). The anti-air systems required to intercept drones like the Shahed-136 are often far more expensive than the drones themselves. MANPADS are estimated to cost between $20,000 and $100,000 per unit, and surface-to-air missiles often exceed $100,000, while a Shahed-136 costs only $10,000 to $30,000 to produce (Zampronha and Albuquerque). The more affordable manufacturing costs of loitering munitions make it possible for militaries to deploy them in large numbers, increasing the probability that at least some will reach their intended target despite an enemy's air defense system.

Another notable advantage of the Shahed-136 is its range. The Iranian MADO MD-550 four-cylinder piston engine, thought to be reverse engineered from the Limbach L550E, provides the drone with an estimated range of 1,500 to 2,500 kilometers (Zampronha and Albuquerque). This allows drone operators to strike critical utility networks, military installations, and other strategic targets located far behind the front lines.

Initial Drone Combat in the Middle East

Iranian proxies in Yemen, Lebanon, Syria, and Iraq had been employing Iranian-made drones across the Middle East for years with support from the Quds Force before the Shahed-136 drew international attention during the conflict in Ukraine (Bashir). From a military perspective, these regional conflicts served as testing grounds for Iran's engineers and military planners to analyze the performance of their drone systems under actual battlefield conditions before facing more advanced Western military capabilities.

One of the more notable users of Iranian drones has been in Yemen, where Houthi rebels have regularly received Iranian assistance. Over the duration of the conflict, the Iranian-allied Houthi forces increasingly employed drones to strike military and critical infrastructure targets. Drone attacks in October 2022 shut down Yemeni oil exports and most of the country's customs revenues (Knights). The actions of the Houthis demonstrated the effectiveness of low-cost drones in conducting precision airstrikes and posing long-range threats to targets that would typically require much more expensive alternatives to achieve similar results (Knights). Although the physical damage caused by each drone strike may be relatively limited on a macro scale, the psychological and economic impact of such strikes has often proven far greater. Countries across the Middle East were therefore forced to make substantial investments in air defense systems and to enhance security around strategic infrastructure—an effort that continues to this day. In June 2026, the United States approved a sale to Kuwait of almost $2 billion worth of Anduril-made counter-unmanned aerial systems (Helou), in the wake of an Iranian drone strike on a Kuwaiti airport that killed one person, damaged airport facilities, and delayed flights.

Russia Acquires the Shahed

The Shahed's journey out of the Middle East began with a problem that had nothing to do with Iran. By the fall of 2022, roughly six months into its full-scale invasion of Ukraine, Russia was burning through its stockpile of precision cruise missiles far faster than it could replace them, especially as Western sanctions cut off access to many of the imported components those missiles required. Moscow needed a cheap, long-range strike weapon that could be acquired quickly and in volume, and Iran had spent the previous two decades building exactly that. On October 6, 2022, Iranian representatives traveled to Moscow to finalize supply terms and a broader partnership surrounding the drones (VGI). Within a few weeks, Shahed-131s and Shahed-136s were appearing in the skies over Ukrainian cities under new Russian designations—the Geran-1 and Geran-2—which, combined with missile attacks, left large portions of Ukraine in blackout (VGI).

What made the arrangement significant, however, was not the initial deliveries but what followed them. Iran did not simply sell Russia a finished product; it exported the entire production model, with Iranian engineers reportedly providing direct technical support as Moscow developed its own manufacturing capacity (IRIA News). By the summer of 2023, Russia had begun producing its own Gerans at the Alabuga Special Economic Zone in Tatarstan, an expansive industrial complex that has since grown to include composite fuselage assembly lines, engine production areas, foundries, and its own worker housing (defence-blog). The output figures tell the story of how quickly the operation scaled: Alabuga produced 2,738 drones in 2023, more than doubled that pace across the first nine months of 2024, and by mid-2025 was churning out roughly 2,700 Shahed-type drones every month alongside thousands of decoy variants meant to bait the enemy into expending resources for defense (VGI). By 2024, Western intelligence agencies assessed that Russia could manufacture the drones domestically at scale without relying on Iranian imports at all (IRIA News). A weapon as scrappy and born out of desperation as the Shahed series proved so useful that a relatively large power such as Russia was willing to incorporate it into its military doctrine.

The Shahed in Ukraine

Once domestic production in Russia came online, the scale of employment became difficult to overstate. Ukrainian defense intelligence estimated that Russia planned to produce roughly 45,700 Shahed-type strike drones in 2025 alone, alongside approximately 34,000 unarmed decoys (Skibitskyi). Russian forces launched more than 38,000 long-range attack drones and decoys at Ukraine between January and November of 2025, accounting for nearly two-thirds of all such launches since the invasion began in February 2022 (Borsari). On a single night in September 2025, Russia sent 810 Shahed-type drones and decoys into Ukrainian airspace, the largest drone attack of the entire war up to that point (IRIA News).

This is precisely the strategy that the Shahed's economics were designed to enable. Ukrainian interception rates against the Geran have remained high, typically falling between 80 and 97 percent, yet the sheer numbers being launched meant that hundreds could still reach their

targets every month, striking energy grids, military installations, and residential areas (Borsari). The Shahed does not need to hit anything to be effective, which reflects the cold economic logic described earlier. The simple presence of one in the sky imposes costs on the defender—a dynamic the Houthis had already demonstrated in Yemen, now replicated on a much more industrial scale.

The drone itself has also evolved considerably in Russian hands. Early Geran-2s relied on pre-programmed coordinates and basic satellite navigation, making them vulnerable to Ukrainian jamming, but later iterations have included anti-jamming antennas, onboard infrared cameras, mesh communication networks that allow operators to steer drones deep inside Ukrainian airspace, and even an Nvidia Jetson Orin computer board, suggesting early attempts at automated target recognition (Borsari). The weapon Iran designed to be cheap and simple is, ironically, becoming steadily more sophisticated in Ukraine, all while holding on to the principle of low cost that keeps it advantageous to use.

Return to the Middle East

While the Shahed was being mass-produced in Russian factories, the weapon made a dramatic homecoming. On the night of April 13-14, 2024, Iran launched a direct attack on Israel from its own soil for the first time, conducting a five-hour barrage that involved an estimated 170 drones, 30 cruise missiles, and 120 ballistic missiles (UANI). The operation, codenamed True Promise, was retaliation for the assassination of General Mohammad Reza Zahedi, commander of the Quds Force in Syria and Lebanon, who was killed in a strike on a building adjacent to the Iranian embassy in Damascus (U. Rubin). This marked a fundamental shift, as Iran was now waging war directly with Iranian-made weapons rather than distributing them to proxies.

The attack also demonstrated the Shahed's limitations against a prepared, technologically advanced opponent. Israel, operating alongside the United States, Britain, France, and other partners, intercepted roughly 99 percent of the incoming projectiles, and not one of the 170 drones penetrated Israeli airspace (GlobalSecurity). The slow speed that gives the Shahed its loitering-munition advantage also gave defenders hours of warning, and a multinational coalition stretching across Jordan, Iraq, and Syria shot the drones down long before they reached their targets. Yet the drones still served a purpose consistent with their design philosophy: to saturate and overwhelm Israeli air defenses (UANI), forcing defenders to expend expensive interceptors and attention while ballistic missiles followed behind.

The pattern repeated on a larger scale in June 2025, when Israeli strikes on Iranian nuclear and military facilities triggered a twelve-day war between the two countries. Iran launched approximately 550 ballistic missiles and around 1,000 drones at Israel over the course of the conflict, with the vast majority of the drones failing even to reach Israel's borders; of those that did pose a threat, nearly 99 percent were intercepted, and only a single drone struck a populated area (Times of Israel). Against the most heavily defended airspace in the region, the Shahed proved far less destructive than in Ukraine. But destruction, once again, was only part of the motive behind a conflict like this one. Every drone launched forced Israel and its allies to

activate fighter jets, warships, and air defense batteries whose operating costs dwarfed the price of the drones they were destroying, effectively drawing them into the economic trap that the drone was designed to exploit.

Limitations of the Shahed Doctrine

The Shahed's advantages come with real constraints that qualify its reputation. Its slow, propeller-driven flight—the same quality that grants it loitering flexibility—also makes it audible, trackable, and vulnerable to defenders who often have hours of warning. Its accuracy is modest, and its relatively small warhead limits the damage any single drone can inflict; in one period of the Ukraine war, only a small fraction of launched Shaheds reportedly reached their targets. Taken individually, a Shahed is a militarily marginal weapon. Its power emerges only in mass, which means the entire doctrine depends on sustaining production and launch volumes that a determined defender can, over time, contest.

More fundamentally, cost-imposition is not the same as victory. For all the economic strain the Shahed places on defenders, no drone campaign examined here has proven strategically decisive. The Houthis harassed shipping and infrastructure without altering the outcome of the war in Yemen; Russia's massed strikes have degraded Ukraine's power grid without breaking its lines; Iran's direct barrages against Israel were absorbed by layered defenses. The Shahed is, in this sense, a spoiler weapon—it exhausts, distracts, and bleeds an opponent, but it does not conquer. A strategy built on imposing costs assumes the adversary will eventually decide those costs are not worth bearing, and so far none has.

Finally, the very economic logic at the heart of the Shahed's success is beginning to erode. The drone's advantage was always the exchange ratio: cheap attacker, expensive defender. But the emergence of low-cost interceptor drones—the same Ukrainian-developed systems the United States has since sought to acquire—threatens to invert that math, pitting a few-thousand-dollar defender against a tens-of-thousands-dollar attacker. Iran's reliance on foreign-manufactured components, up to 82 percent of them Western-made in at least one examined drone, adds a second vulnerability: the supply chain that makes the Shahed cheap also makes it interdictable. The weapon that rewrote the economics of war may yet be undone by the same economics it exploited.

The 2026 Iran War

The uneasy calm following the twelve-day war did not last. On February 28, 2026, the United States and Israel began a series of coordinated strikes against Iran, aimed at the country's nuclear and ballistic missile programs and at inducing regime change; the opening attacks killed several senior Iranian officials (Commons Library). Over the course of the conflict, Supreme Leader Ali Khamenei was killed and succeeded by his son (Commons Library). Iran's retaliation was the largest demonstration of the Shahed doctrine to date, launching counterstrikes against Israel, U.S. military bases in the region, and military and civilian locations in Arab states (Commons Library). In the opening days of the conflict, Shahed-136s launched in coordinated

waves struck targets in Bahrain, Kuwait, and Dubai, and two drones hit the U.S. Embassy compound in Riyadh, triggering a fire (Gulf News). Whatever limitations the drone had shown against Israel's layered defenses in 2024 and 2025, the 2026 war proved that when spread across an entire region's worth of targets, the Shahed could still impose enormous costs. Gulf states found themselves firing interceptors worth millions of dollars apiece at drones costing as little as $20,000—an exchange ratio that overwhelmingly favors the attacker regardless of how many drones are shot down (Simkins).

The most remarkable development of the war, however, came from the other side. When Operation Epic Fury began, U.S. Central Command's newly formed Task Force Scorpion Strike employed low-cost one-way attack drones in combat for the first time (Robinson). The weapon in question, the Low-cost Uncrewed Combat Attack System, or LUCAS, was not merely inspired by the Shahed-136—it was reverse-engineered from it. The United States had covertly acquired downed Shahed-136s for technical analysis, initially producing replicas for counter-drone training before adapting them into an operational strike weapon (Simkins). Built by the Arizona-based firm SpektreWorks and priced at roughly $35,000 per unit, the LUCAS shares the Shahed's triangular delta-wing design and even uses remarkably similar engine technology, while adding a nose-mounted camera system and satellite connectivity that allows operators to re-task drones mid-flight (Simkins). The logic was dumb: contrasting the roughly $35,000 cost of the drone against the multimillion-dollar interceptors used to shoot such weapons down yields a cost exchange overwhelmingly in the attacker's favor (Simkins).

The symbolism here is difficult to overstate. Iran developed the Shahed because sanctions and isolation left it unable to match Western precision weapons, forcing it to pursue cheap, mass-producible alternatives. Two decades later, the wealthiest and most technologically advanced military in the world—one with no shortage of million-dollar cruise missiles—studied Iran's weapon, copied it, and turned it against its creator. The Pentagon's embrace of what it now calls "attritable warfare," the doctrine that weapons should be cheap enough to lose in large numbers without budgetary crisis, represents a formal admission that the economic logic Iran pioneered out of desperation has become universal (Gulf News). Even the defensive side of the equation is being rewritten along the same lines: the most effective counter to the Shahed-136 is no longer a multimillion-dollar interceptor but cheap Ukrainian-developed interceptor drones costing a few thousand dollars, and the United States has drawn directly on Ukraine's anti-Shahed expertise (Tucker; Gulf News). As of this writing, the war remains unresolved—a memorandum of understanding signed in June 2026 collapsed within weeks amid renewed strikes and confrontations in the Strait of Hormuz (Al Jazeera)—but the Shahed's legacy is already settled. A weapon born from Iranian weakness has redefined how the strongest militaries on Earth think about the economics of war.

Works Cited

Al Amir, Khitam. "Two Drones Target US Embassy in Riyadh." Gulf News, 3 Mar. 2026, gulfnews.com/world/gulf/saudi/two-drones-target-us-embassy-in-riyadh-1.500461486.

Bashir, [First Name]. The Iranian Drone Program: Role and Scope of Its Influence in Iran's Foreign Policy. Rasanah—International Institute for Iranian Studies, June 2023, rasanah-iiis.org/english/wp-content/uploads/sites/2/2023/06/THE-IRANIAN-DRONE-PROGRAM-ROLE-AND-SCOPE-OF-ITS-INFLUENCE-IN-IRANS-FOREIGN-POLICY.pdf.

Bertrand, Natasha. "A Single Iranian Attack Drone Found to Contain Parts from More than a Dozen US Companies." CNN, 4 Jan. 2023, cnn.com/2023/01/04/politics/iranian-drone-parts-13-us-companies-ukraine-russia.

Borsari, Federico. "How Russian Drone Developers Outpace the West." Center for European Policy Analysis, cepa.org/article/how-russian-drone-developers-outpace-the-west/.

"How Russia Ramps Up Shahed Drone Production." Defence Blog, defence-blog.com/russia-ramps-up-shahed-drone-production/.

Helou, Agnes. "US Approves $2B Sale of Anduril Counter-Drone Systems to Kuwait." Breaking Defense, June 2026, breakingdefense.com/2026/06/us-approves-2b-sale-of-anduril-counter-drone-systems-to-kuwait/.

"Iranian Shahed-136 Drone Unveiled in UK Parliament." United Against Nuclear Iran, unitedagainstnucleariran.com/iranian-shahed-136-drone-unveiled-uk-parliament.

Knights, Michael, et al. "[Article Title]." CTC Sentinel, vol. 17, Apr. 2024, ctc.westpoint.edu/wp-content/uploads/2024/04/CTC-SENTINEL-042024.pdf.

Loft, Philip. "Israel/US-Iran Conflict 2026: Background and UK Response." House of Commons Library, 24 Apr. 2026, commonslibrary.parliament.uk/research-briefings/cbp-10521/.

"Operation True Promise." GlobalSecurity.org, 15 Apr. 2024, globalsecurity.org/military/world/war/operation-true-promise-20240415.htm.

Oliverio, Natalie. "Iran War Depleted US Patriot Missile Stockpiles, Creating Readiness Challenges, Experts Say." Military Times, 30 July 2026, militarytimes.com/news/your-military/2026/07/30/iran-war-depleted-us-patriot-missile-stockpiles-creating-readiness-challenges-experts-say/.

"Russia Increases Shahed Drone Production to 2,700 Units per Month, Ukrainian Intelligence Reports." IRIA News, ir-ia.com/news/russia-increases-shahed-drone-production-to-2700-units-per-month-ukrainian-intelligence-reports/.

Robinson, Dan. "US Struck Iran with Copies of Its Own Drones." The Register, 2 Mar. 2026, theregister.com/offbeat/2026/03/02/us-struck-iran-with-copies-of-its-own-drones/4833761.

Rubin, Michael. A Short History of the Iranian Drone Program. American Enterprise Institute, Aug. 2020, aei.org/wp-content/uploads/2020/08/A-short-history-of-the-Iranian-drone-program.pdf.

Rubin, Uzi. "Operation True Promise: Iran's Missile Attack on Israel." BESA Center, besacenter.org/operation-true-promise-irans-missile-attack-on-israel/.

Shamim, Sarah. "US-Iran Memorandum of Understanding Expires: How and Why It Fell Apart." Al Jazeera, 17 Aug. 2026, aljazeera.com/news/2026/8/17/us-iran-memorandum-of-understanding-expires-how-and-why-it-fell-apart.

Simkins, J. D. "US Confirms First Combat Use of LUCAS One-Way Attack Drone in Iran Strikes." Military Times, 28 Feb. 2026, militarytimes.com/news/your-military/2026/02/28/us-confirms-first-combat-use-of-lucas-one-way-attack-drone-in-iran-strikes/.

Skibitskyi, Vadym. "Ukraine's Defense Intelligence Clarifies Actual Shahed Drone Production in Russia." Defence UA, en.defence-ua.com/news/ukraines_defense_intelligence_clarifies_actual_shahed_drone_production_in_russia_amid_talk_of_1000_per_day-15472.html.

"The Evolution of Shaheds: How the Enemy's Weapon Developed." VGI, 13 Oct. 2025, vgi.com.ua/en/the-evolution-of-shaheds-how-the-enemys-weapon-developed/.

"The Israel-Iran War by the Numbers after 12 Days of Fighting." The Times of Israel, 27 June 2025, timesofisrael.com/the-israel-iran-war-by-the-numbers-after-12-days-of-fighting/.

Tucker, Patrick. "The Pentagon Replicated a Ukrainian-Style Drone Attack in Florida. Now It's Changing Its Counter-Drone Strategy." Defense One, 23 Apr. 2026, defenseone.com/threats/2026/04/pentagon-ukraine-counter-drone/413087/.

Zampronha, D., and A. Albuquerque. "An Exploratory Study about the HESA Shahed 136." Advances in Aerospace Science and Technology, 2024, scirp.org/journal/paperinformation?paperid=131823.

Did Republican Counties Face Greater Retaliatory Tariff Exposure in 2025?
By Jubili Bamzai-Wokhlu

Abstract

Background*:* U.S. trading partners maintain retaliatory tariffs arising from trade actions spanning the 2018 Section 232 and 301 disputes and the April 2025 reciprocal tariff order. Evidence of political targeting in the 2018 U.S.–China trade conflict raises questions about their partisan impact. This study assesses whether counties with higher Republican vote share are disproportionately exposed to them.

Methods*:* This study introduces a county-level Tariff Exposure Index (TEI), built from industry-specific tariff rates, export volumes, and employment shares, to measure local exposure. The dataset covered 8,172 tariff lines from five countries across 3,144 U.S. counties. Analyses were restricted to 3,095 counties with 2024 election data.

Results*:* Retaliating countries differed markedly in breadth and severity. Manufacturing accounted for 89% of tariff lines matched to industry sectors, while agriculture faced the highest average rates. Republican-leaning counties showed higher tariff exposure than Democratic-leaning counties (median 54.8 vs. 33.7 percentile; $r = 0.12$). Exposure among Republican counties fell as income rose, the association reversing sign near a county median household income of $61,000. After demographic controls, Republican counties overall showed no significant association with exposure, a null result that may mask this divergence. Manufacturing employment share was positively associated with Republican vote share ($r = 0.21$) and closely tracked exposure ($r = 0.94$), though the second association is partly built into the index.

Conclusions*:* The partisan exposure gap is largely accounted for by county economic geography, especially manufacturing concentration, rather than requiring a targeting explanation. The role of manufacturing may warrant further study.

Keywords: retaliatory tariffs; political geography; tariff exposure; Republican vote share; manufacturing; trade policy

Introduction

On April 2, 2025, President Trump signed an executive order declaring a national emergency over the United States' trade deficit (United States, Executive Office of the President). He designated the day "Liberation Day" and introduced a sweeping 10% baseline tariff on nearly all imports, as well as "reciprocal" country-specific tariff rates. These measures sent stock markets sharply lower (Treisman) and prompted swift retaliatory responses from multiple trading partners (International Trade Administration, Foreign Retaliations Database).

Retaliatory tariffs are import taxes imposed by one country in response to trade restrictions enacted by another. Unlike traditional trade barriers aimed primarily at correcting trade balances or protecting domestic industries, retaliatory tariffs are often explicitly strategic: they are devised to impose concentrated economic pain, signal resolve, and influence policy

behavior in the target country (Martin and Vergote). More broadly, political-economy research treats trade protection as an equilibrium outcome of organized political interests rather than a purely economic choice (Grossman and Helpman).

Despite their growing prominence, the intentions and effects of retaliatory trade policies remain poorly understood, in part because such large-scale, politically motivated trade conflicts are historically rare. Aside from the U.S.–China trade war of the late 2010s, the most recent parallel is the Smoot–Hawley tariff regime of the 1930s (Fajgelbaum et al.; Irwin). As seen in China's behavior in the first U.S.–China trade conflict, there is evidence that countries can enact politically targeted trade retaliation by focusing the economic pain of sanctions in a way designed to maximize the resulting political pain, targeting partisan electoral channels (Brutger et al.). Fetzer and Schwarz find that retaliatory tariff exposure from this conflict concentrated in Trump-supporting and swing areas. Kim and Margalit find strong evidence that Chinese tariffs systematically targeted U.S. goods with production concentrated in Republican counties, particularly when those counties were located in closely contested congressional districts. Blanchard et al. further find that Republican House candidates lost support in counties more exposed to tariff retaliation in the 2018 midterm elections. Beyond retaliation specifically, rising trade exposure has been shown to shift electoral outcomes and deepen political polarization across U.S. local labor markets (Autor et al., "Importing Political Polarization?").

This raises the question of whether such behavior extends to the retaliatory tariffs currently in effect against the United States. Empirical research addressing the political geography of these measures is scant. The current trade conflict differs in important ways from the first Trump presidency, whose disputes were largely bilateral and centered on China. In contrast, the 2025 tariff escalation is global in scope, involving retaliatory actions by a diverse set of trading partners with distinct strategic incentives and trade relationships with the United States. Moreover, the scale and uniformity of the new U.S. tariff regime may alter both the feasibility and the motivations of political targeting abroad. Accordingly, existing research has not fully examined sociopolitical trends in retaliatory tariff impact.

This paper revisits the political geography of retaliatory tariffs under these conditions, asking whether counties with higher Republican vote share exhibit greater estimated retaliatory tariff exposure. Employing a county-level Tariff Exposure Index to compare local exposures, the analysis tests the hypothesis that counties with higher Republican vote share are more exposed to retaliatory tariffs. The index measures exposure from pre-tariff trade patterns and announced tariff schedules; it does not measure realized economic incidence, which depends on pass-through and substitution (Amiti et al.), nor does it identify foreign-government intent. Distinguishing where retaliatory exposure falls from why particular products were selected matters because the two questions require different evidence and support different conclusions.

Prior research suggests that retaliatory tariffs can produce politically uneven geographic exposure, and two explanations compete. Political targeting is one: a retaliating government may select products made in politically salient constituencies in order to concentrate pressure on the opposing government. Economic considerations offer an alternative, because product selection

may instead reflect trade volumes, domestic substitutability, sector composition, or the retaliating country's own costs.

The present analysis therefore tests whether Republican-leaning counties show systematically higher estimated exposure, and whether any unadjusted partisan gap persists after accounting for observable county characteristics. Demographic controls proxy for blue-collar share, urban–rural differences, and goods-producing economies. A direct test of political product selection would require modeling the universe of eligible products and the selection process itself, which lies outside the scope of a county-level exposure analysis.

Methods

Research Design and Unit of Analysis

This study constructs a county-level Tariff Exposure Index (TEI) and relates it to county partisanship. The study uses a non-experimental design with descriptive and correlational components. Because the variables of interest cannot be experimentally manipulated, the analysis uses observational county-level data and statistical controls to estimate relationships between tariff exposure, partisanship, and local economic structure. Due to the observational design, the results can indicate whether associations are consistent with political targeting models but cannot demonstrate intent.

The county is the unit of analysis because county-level data capture variation in local industry structure, demographics, and voting patterns. Partisanship data, industry composition, and demographic variables are also consistently available at the county level. Of 3,144 U.S. counties and county equivalents in the initial dataset, 3,095 had available 2024 presidential election data and were included in the final analysis. Regression samples are marginally smaller because observations with missing covariate values were dropped listwise. All methods and raw data can be accessed via the following repository:
https://github.com/jnbamzaiwokhlu/retaliatory-tariff.

Data Sources and Variables

This study uses five public datasets to analyze county tariff exposure in relation to demographic and political factors. Retaliatory tariff schedules were obtained from the International Trade Administration's Foreign Retaliations Database, which tracks retaliatory tariffs imposed by trading partners in response to U.S. trade policy using official tariff schedules (International Trade Administration, Foreign Retaliations Database). The dataset reports tariff rates at the Harmonized System (HS) 8-digit product level and includes retaliatory actions by China, Canada, Russia, Turkey, and India, totaling 8,172 tariff lines.

The database records all retaliatory measures in effect at the time of collection, which arise from multiple U.S. trade actions, including the 2018 Section 232 and Section 301 tariffs and the April 2025 reciprocal tariff order. The measures were therefore imposed at different times: those from Canada and China date largely from 2025, whereas those imposed by Russia,

Turkey, and India date from the 2018–2019 Section 232 dispute (International Trade Administration, Foreign Retaliations Timeline). These older schedules, referred to here as legacy measures, remain in force and are recorded alongside the 2025 measures. Because the database records only measures in force, it did not include measures from the European Union (EU) or Mexico: the European Union's measures were suspended during the study period, and Mexico did not impose retaliatory tariffs on U.S. exports. Their absence is a property of the source rather than a sampling choice.

Export values were obtained from *USA Trade Online* and aggregated using 2023–2024 data (United States, Census Bureau, *USA Trade Online*). Export values are reported by HS code, converted to North American Industry Classification System (NAICS) industry, and used to weight tariff exposure by the economic importance of each industry. County employment shares were constructed using the 2023 County Business Patterns dataset, which reports employment and establishments by NAICS industry at the county level (United States, Census Bureau, *County Business Patterns*).

County-level demographic characteristics were drawn from the 2022 American Community Survey 5-year estimates, including population, median household income, educational attainment, racial composition, unemployment rate, and median age. These controls test whether observed partisan disparities stem from Republican vote share itself or the underlying geography of tradable sectors, including manufacturing-heavy, rural, lower-education, and goods-producing counties (United States, Census Bureau, *American Community Survey*).

County-level 2024 presidential election results from the MIT Election Data and Science Lab were used to measure county Republican vote share. The 2024 election was used because it provides the most recent measure of county partisanship and captures the voter base most relevant to current trade policy.

The primary dependent variable is county-level TEI, which measures local economic exposure to retaliatory tariffs. The primary independent variable is county Republican vote share in the 2024 presidential election. Control variables include log population, median household income, unemployment rate, college-degree share, racial composition, and median age. Manufacturing employment share is included as an exploratory variable because manufacturing concentration may link partisan geography to tariff exposure.

Construction of the Tariff Exposure Index

The TEI integrates county-level industry composition, national export volumes, and industry-level tariff rates to approximate local exposure to retaliatory tariffs, following the shift-share logic of local labor-market exposure measures used in the trade literature (Autor et al., "The China Syndrome"). Retaliatory tariff schedules were observed at the HS-8 product-code level and converted to NAICS industries using the concordance package in R (Liao et al.), which maps product codes to industry codes using an official United Nations crosswalk.

A quality-control procedure was applied before constructing the index. Tariff lines that could not be matched to NAICS industries were excluded. Lines mapping erroneously to public-administration placeholder codes, including NAICS 91–92, were manually recorded to their correct sectors after inspection when appropriate; 43 waste and scrap lines were reassigned in this way. Non-traded HS chapters, including chapters 97 and 000, were excluded. Of the 8,172 tariff lines in the original dataset, 7,635 were successfully mapped to NAICS sectors and retained for sector-level analysis. After this mapping process, a mean effective tariff rate was computed for each two-digit NAICS sector by averaging across tariff lines within that sector.

County-level exposure was computed as:

$$TEI_c = \sum_i (TariffRate_i \times ExportVolume_i \times EmploymentShare_{ci})$$

where *i* indexes each two-digit NAICS industry and *c* indexes each county. This produces a dollar-weighted measure capturing both the severity of tariffs on an industry and the degree to which a county's workforce is concentrated in that industry. Because the resulting values are not directly interpretable in dollar terms, county TEI values were converted to percentile ranks across all counties. The dependent variable therefore ranges from 0 to 100 and is reported throughout as a percentile.

Analytical Methods

Descriptive statistics summarize the distribution of retaliatory tariff rates across countries and industries. Continuous variables are reported using means, medians, interquartile ranges, and, where estimated, 95% confidence intervals. The structure of retaliatory tariff schedules was examined across the five retaliating countries by comparing breadth, precision, and severity. Breadth was measured by the number of HS chapters targeted, precision by the number of tariff lines targeted, and severity by mean effective tariff rates and their distribution.

County-level heat maps were used to visualize the geographic distribution of tariff exposure and 2024 Republican vote share. Manufacturing employment was examined as a potential mechanism linking partisan lean to tariff exposure using scatter plots of manufacturing share against both Republican vote share and tariff exposure.

Pearson correlation coefficients and ordinary least squares regressions evaluated the relationship between Republican vote share and tariff exposure. Regression specifications were introduced sequentially. The baseline model included only Republican vote share, aside from the state fixed effects and population weights applied to all specifications. Subsequent models added socioeconomic controls, including median household income and college-degree share. Additional models introduced an interaction between Republican vote share and median household income to test whether partisan exposure varied by county income level. Demographic specifications added log population, racial composition, unemployment rate, and median age. All models included state fixed effects and were weighted by log county population, which gives greater weight to more populous counties while preventing the largest metropolitan counties from dominating the estimates. Standard errors are conventional and are not clustered.

Ethical Considerations

This study uses publicly available population-level datasets. Because no individual-level information was collected, informed consent is not applicable.

Tools and AI Disclosure

All data cleaning, statistical analysis, and visualization were conducted in R version 4.4.1 (R Core Team). AI tools were used for assistance in debugging code, manuscript and citation formatting, and language editing, but all research design, variable construction, analysis, and interpretation were the author's own.

Results

This section presents results in four parts: (1) the structure of retaliatory tariff schedules across the five representative countries, (2) which industries bear the greatest tariff burden at the sector and product levels, (3) assessment of county-level exposure in politically salient areas, and (4) analysis of manufacturing employment as a possible mechanism.

Varying Tariff Strategy Across Retaliating Countries

As noted above, the schedules compared here were imposed at different times, the Canadian and Chinese dates from 2025 and the three legacy schedules from the 2018–2019 Section 232 dispute. Differences across countries therefore arise from the distinct U.S. actions being answered as well as from any difference in approach. These comparisons characterize differences in tariff coverage and severity and are not interpreted as evidence of political targeting or intent. Table 1 summarizes these schedules by country, and Figure 1 displays their variability across three dimensions: number of tariff lines targeted, mean effective tariff rate (ETR), and the number of HS chapters affected; these metrics express the precision, severity, and sectoral breadth of retaliation, respectively.

Tariff Lines Targeted counts how many specific products were named. China implemented the largest and most granular retaliation package, imposing tariffs on 6,265 tariff lines, or 77% of the 8,172 lines in the dataset. Canada also adopted a broad retaliatory strategy, applying tariffs across 1,795 lines, a further 22% of the dataset, although its schedule was less granular than China's. The remaining three schedules are far smaller: Russia covered 79 lines, Turkey 22, and India 11, together accounting for 1.4% of the dataset.

Mean Effective Tariff Rate (ETR) measures the overall severity of retaliation. Canada imposed the highest mean ETR (34.5%), with more than a quarter of tariff lines exceeding 50%. Canada's highest tariffs applied to motorcycles (100%) and citrus fruits (approximately 80%), while poultry and steel products faced tariffs between 65% and 70%. China's retaliatory tariffs were less severe on average, with a mean ETR of approximately 20%, though some agricultural products such as pork (65%) and tree nuts (55%) faced higher rates. Among the smaller schedules, Russia imposed the highest mean rate at 28.7%, followed by Turkey at 18.5% and India at 14.1%, across far fewer products.

Number of HS Chapters Targeted expresses sectoral breadth. China implemented the broadest retaliation strategy, covering 96 distinct HS chapters, compared to 56 chapters targeted by Canada. China spread its retaliatory tariffs across a wider set of industries, even though its average tariff rates were lower. The remaining schedules are far narrower still, covering four HS chapters for Russia, five for India, and one for Turkey.

Table 1. Retaliatory Tariff Schedules by Country

Country	Tariff lines	HS chapters	Mean ETR	Median ETR	Max ETR	Lines ≥ 25%	Lines ≥ 50%
2018 Section 232 comparators							
Russia	79	4	28.7%	30%	40%	100%	0%
Turkey	22	1	18.5%	12.5%	40%	40.9%	0%
India	11	5	14.1%	15%	15%	0%	0%
2025 measures							
China	6,265	96	20%	25%	65%	51.8%	1.2%
Canada	1,795	56	34.5%	25%	100%	100%	28.5%

Note. ETR = effective tariff rate. The final two columns report the share of a country's tariff lines at or above the stated rate. China's figures cover searchable product-specific schedules only and exclude the April 2025 blanket tariff escalation. Measures imposed by Russia, Turkey, and India are pre-existing Section 232 measures first imposed in 2018 and still in effect during the study period; those imposed by China and Canada date from 2025.

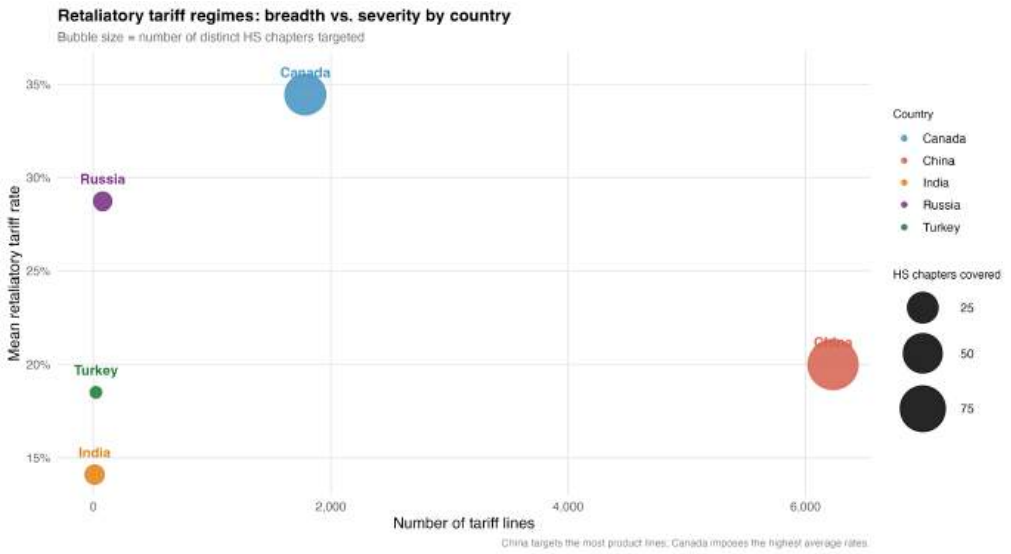


Figure 1. Retaliatory tariff breadth and severity by country.

Bubble plot comparing retaliatory tariff schedules imposed by Canada, China, Russia, Turkey, and India. The x-axis reports the number of tariff lines, the y-axis reports the mean effective tariff rate (ETR), and bubble size represents the number of Harmonized System (HS) chapters targeted.

Industries Impacted by Retaliatory Tariffs

Figure 2 shows the sectoral impact of tariffs by two-digit tradable industry NAICS codes. Of the 7,635 tariff lines successfully mapped to NAICS sectors, the majority impacted manufacturing industries, accounting for 6,797 lines (approximately 89%). Agricultural products represented the second largest category of retaliatory tariffs, involving 735 tariff lines, while mining, oil, and gas accounted for 100 lines and information for 3.

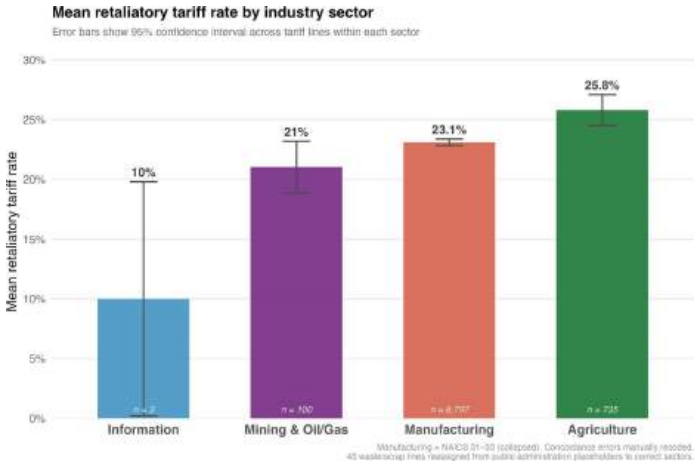


Figure 2. Mean retaliatory tariff rate by two-digit NAICS sector

Bars report the mean retaliatory tariff rate across tariff lines within each two-digit North American Industry Classification System (NAICS) sector, with manufacturing (NAICS 31–33) collapsed into a single category. Error bars represent 95% confidence intervals.

Figure 3 provides a more granular representation of targeted industries. Agricultural products, particularly fruits and nuts, coffee, vegetables, and tobacco, faced the highest tariffs. Within manufacturing, the targeted goods were relatively broad and included steel and other metal products, as well as apparel and leather goods.

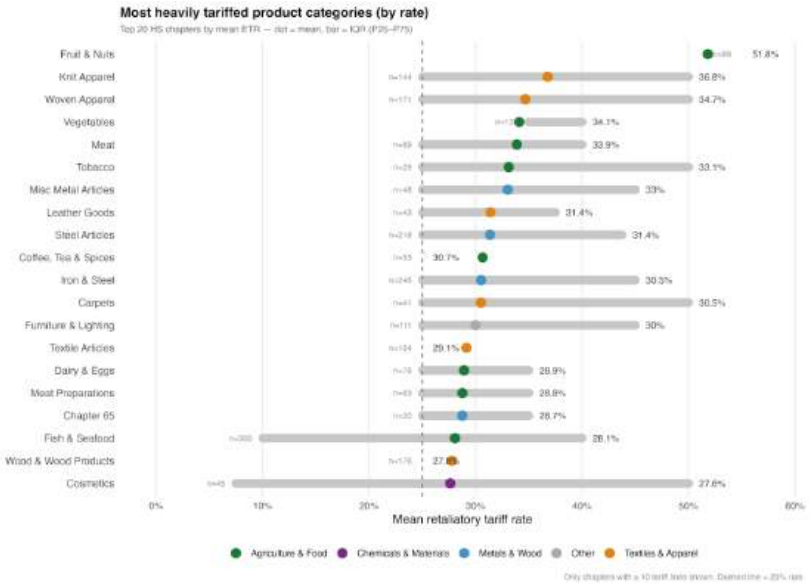


Figure 3. Retaliatory tariff rates for the most heavily tariffed product categories.

Mean retaliatory tariff rates are reported for selected Harmonized System (HS) product categories; gray bars indicate interquartile ranges. Broad sector categories are grouped.

Correlational Analysis of Tariff Exposure and Partisanship

Figure 4 provides a choropleth of county-level tariff exposure across the United States alongside 2024 Republican (GOP) vote share to illustrate geographic overlap. The industrial Midwest and Southeast regions (excluding Florida) appeared most vulnerable to exposure from retaliatory tariffs; as the map illustrates, these areas also tended to lean heavily Republican. Lower exposure levels appear more frequently in coastal metropolitan regions and service-oriented economies.

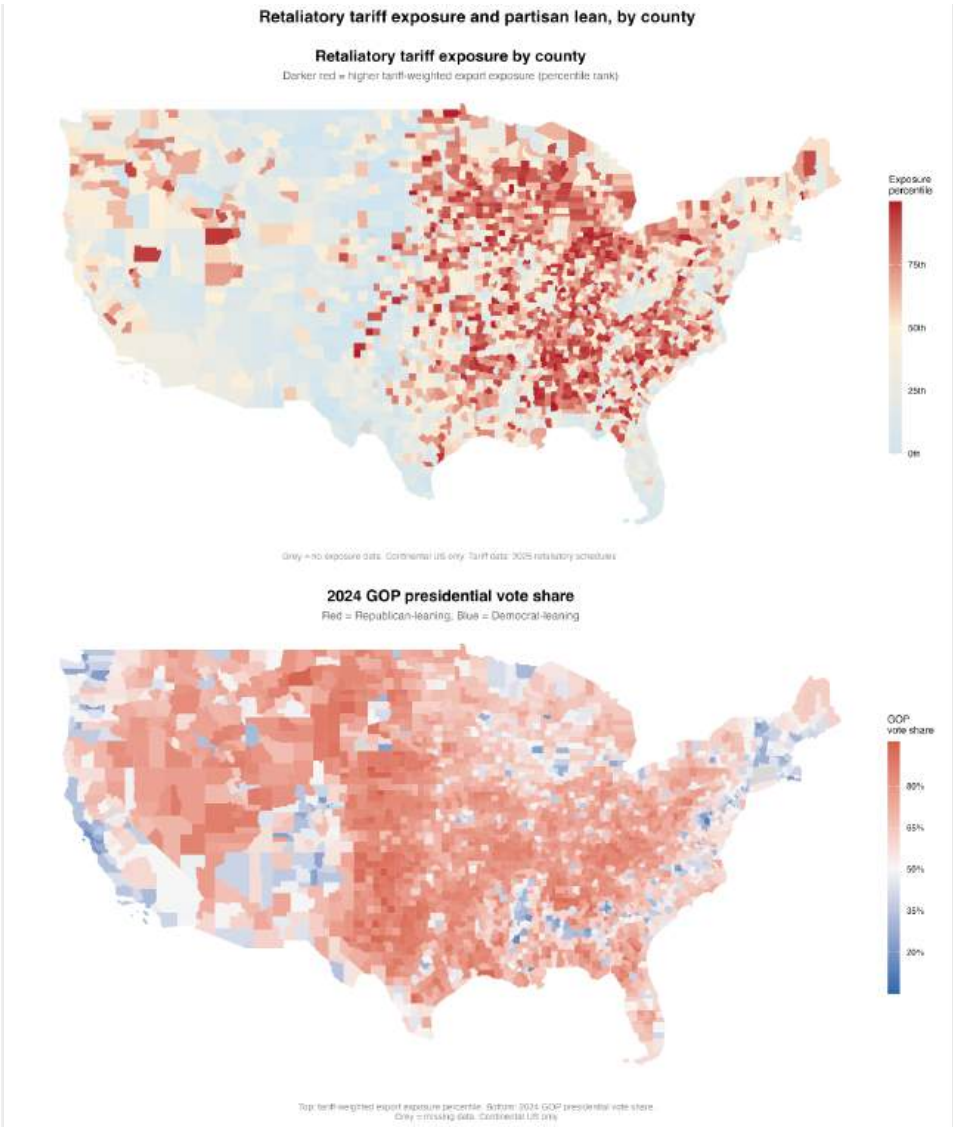


Figure 4. County-level choropleth maps of retaliatory tariff exposure and partisanship across the United States

The top panel displays the percentile rank of the county-level Tariff Exposure Index (TEI). The bottom panel shows county-level Republican vote share in the 2024 presidential election.

To further assess the apparent geographic overlap between tariff exposure and partisanship, Table 2 quantifies the difference in TEI exposure across counties based on partisan lean, showing that the average Republican county has much higher tariff exposure than the average Democratic county. The finer partisan breakdowns in Figure 5 show that moderately Republican counties (55–65% GOP) had the highest median tariff exposure, at the 56.6th percentile, with highly Republican counties a close second.

Table 2. Summary of Tariff Exposure for Democratic vs. Republican Counties.

Partisan category	Number of counties (N)	Mean TEI percentile	Median TEI percentile	25th–75th percentile range
All counties	3,095	50.1	50.3	25.3–75.1
Republican (>50% GOP)	2,622	52.4	54.8	27.4–77.8
Democratic (<50% GOP)	473	36.9	33.7	18.4–49.4

Note. Counties are classified as either Republican or Democratic depending on their vote share in the 2024 presidential election. Counties without such partisanship metrics are excluded from this analysis.

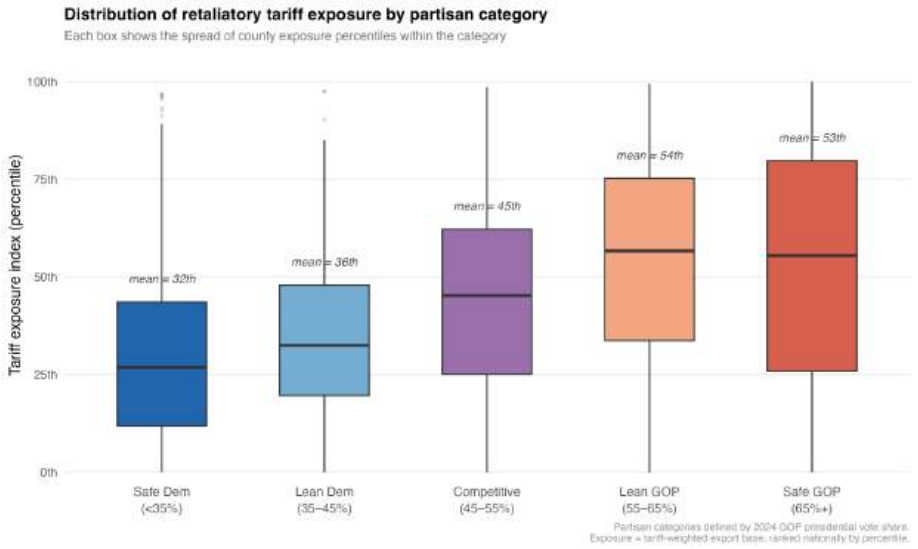


Figure 5. Distribution of tariff exposure by partisan category

Boxplots report county Tariff Exposure Index percentile ranks across 2024 presidential vote-share categories. Boxes show interquartile ranges, horizontal lines show medians, and labels report group means.

Figure 6 further characterizes the relationship between GOP vote share and tariff exposure on a continuous scale. The relationship is weakly positive (r = 0.12).

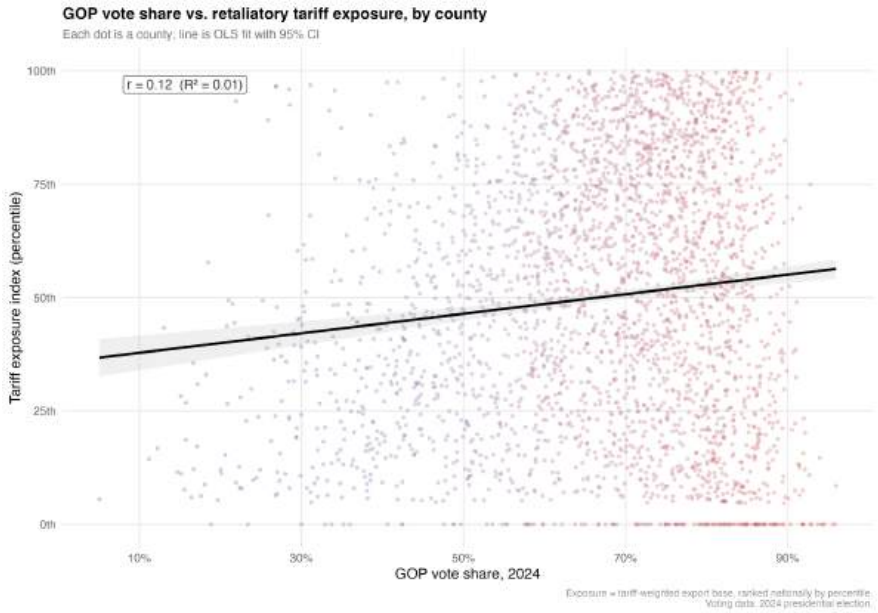


Figure 6. Scatterplot of GOP vote share versus retaliatory tariff exposure, by county

Each point is a county; the line is an ordinary least squares (OLS) fit. The y-axis shows the percentile rank of the Tariff Exposure Index (TEI), ranging from 0 to 100 and normalized by rank across counties. The faint gray band represents the 95% confidence interval.

Manufacturing Employment Underlying the Partisan Lean of Trade Exposure

Table 3 presents the effect of GOP vote share on tariff exposure across progressive model specifications. In the baseline model, GOP vote share is positively associated with tariff exposure, and this persists after adding economic controls (specification 2a). The income interaction term (specification 2b) reveals that the effect of GOP vote share decreases as county income rises, meaning that low-income Republican counties face disproportionate exposure. After full demographic controls (specification 3), the overall association became statistically insignificant; however, the income-dependent relationship persisted in the most comprehensive model (specification 4), indicating that low-income Republican counties experience elevated tariff exposure not fully explained by underlying county characteristics. Because log median household income is not mean-centered, the coefficients on GOP vote share in specifications 2b and 4 describe the marginal effect at an income of one dollar and should not be interpreted directly. Evaluated across the observed range of county incomes, the marginal effect of GOP vote share in specification 4 is positive in counties with median household income below roughly $61,000 and negative above that point, consistent with the near-zero average effect estimated in specification 3.

Table 3. OLS Estimates of GOP Vote Share and Retaliatory Tariff Exposure across Models

Variables	Baseline Model 1	Spec. 2a	Spec. 2b	Spec. 3	Spec. 4
GOP vote share (pp)	0.666*** (0.062)	0.349*** (0.076)	4.881*** (1.756)	−0.038 (0.116)	3.549** (1.457)
Log (median household income)	—	12.786* (7.022)	34.437*** (11.506)	13.901*** (4.754)	30.724*** (8.531)
GOP share × log(income)	—	—	−0.408** (0.159)	—	−0.322** (0.129)
Share college education	—	−75.210*** (13.693)	−81.961*** (13.155)	−101.035*** (10.573)	−106.979*** (10.808)
Log (population)	—	—	—	−1.057** (0.521)	−0.692 (0.529)
Share white	—	—	—	24.533* (14.613)	25.894* (14.299)
Unemployment rate	—	—	—	−118.717*** (14.370)	−114.560*** (14.662)
Observations	3,094	3,093	3,093	3,093	3,093
R^2	0.487	0.528	0.534	0.560	0.564
Adj. R^2	0.479	0.520	0.526	0.553	0.556

Note. All models include state fixed effects and are weighted by log county population. Coefficients are shown with standard errors in parentheses. *p < 0.10; **p < 0.05; ***p < 0.01. pp = percentage points. Sample sizes differ slightly across specifications because of missing covariate data. Median age was included in specifications 3 and 4 but is not shown. In specifications 2b and 4, log median household income is not mean-centered, so the GOP vote share coefficient gives the marginal effect at log(income) = 0 rather than at a typical county income.

To explore other potential mechanisms behind the disproportionate tariff exposure observed across counties and partisan groups, such as industrial composition, the relationship of manufacturing to tariff exposure was assessed further. Figure 7 shows that manufacturing employment is positively associated with both GOP vote share (Figure 7a) and tariff exposure (Figure 7b). This indicates that manufacturing counties are more likely to be Republican leaning, and that increases in manufacturing employment share strongly predict increases in tariff exposure ($r = 0.94$), although this second association is partly definitional because the index is built from employment shares. These patterns are consistent with a potential contribution of industrial composition to tariff exposure, but it was beyond the scope of this exploratory analysis to investigate the strength of this correlation accounting for other factors in additional multivariate analysis.

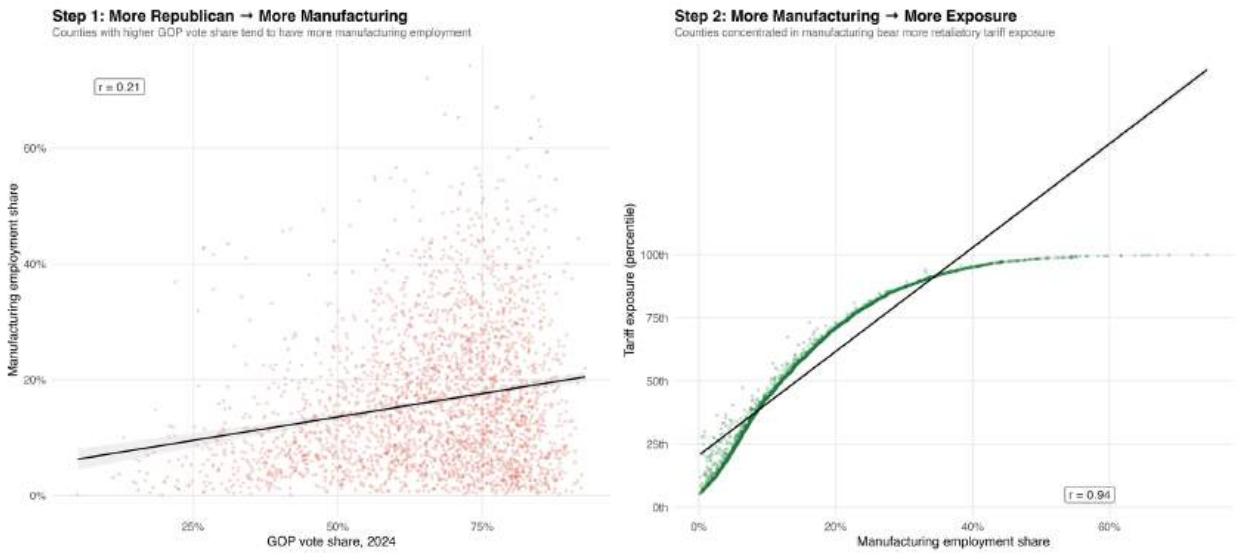


Figure 7. Manufacturing employment as a potential structural channel linking partisanship and tariff exposure

The left panel (7a) shows county manufacturing employment share against 2024 Republican vote share. The right panel (7b) shows manufacturing employment share against TEI percentile rank. Fitted lines summarize bivariate relationships.

Discussion

This study examines the extent to which the retaliatory tariffs in effect against the United States disproportionately impact Republican counties, and whether such patterns reflect deliberate political targeting or the underlying economic geography of export-intensive industries. The author hypothesized that results would be consistent with a political targeting model, with many tariff lines and a strong partisan exposure association. One methodological contribution of this study is the construction of the county-level Tariff Exposure Index (TEI), a composite metric that integrates tariff rates, employment shares, and export volumes to capture differences in tariff exposure across 3,144 counties. Linking this index to data on county partisanship yields three main findings. In brief, Republican-leaning counties are substantially more exposed to retaliatory tariffs than Democratic-leaning counties, but that gap is largely accounted for by where tradable industry is located rather than by partisanship itself.

Variation Across Retaliating Schedules

Retaliating countries differed substantially in their tariff strategies. China and Canada imposed by far the broadest and most severe measures, whereas the three legacy schedules were narrow and limited in scale. China pursued breadth, targeting many HS chapters at moderate average rates, while Canada concentrated its retaliation at higher rates on a smaller set of high-profile goods. This heterogeneity suggests that characterizing retaliation as a uniform phenomenon is misleading. It should be read with the legacy measures in mind, however: the Canadian and Chinese schedules answer 2025 tariff actions, whereas those of Russia, Turkey,

and India answer the 2018 Section 232 measures, so the contrast reflects differing occasions for retaliation as much as differing theories of leverage. This structure shapes the exposure analysis that follows: because the two largest schedules are also the broadest, retaliatory coverage extends across most manufacturing categories rather than concentrating in a few, which is what spreads exposure widely across manufacturing counties.

The Sectoral Burden of Retaliation

At the industry level, manufacturing bore the vast majority of retaliatory tariff lines, followed by agriculture. This follows from the structure of tariff schedules themselves: manufacturing spans far more product categories than any other tradable sector, so a broad retaliation list mechanically lands there. Importantly, agriculture faced the highest average tariff rates despite fewer lines, concentrated in high-value categories such as fruits and nuts and vegetables. Agriculture contains fewer tariff categories than manufacturing, which helps explain why it accounts for fewer tariff lines overall.

Partisanship, Income, and Exposure

Republican-leaning counties exhibited meaningfully higher tariff exposure than Democratic-leaning counties. However, the multivariate results do not require a targeting explanation. The positive relationship between GOP vote share and tariff exposure remained when economic controls were included, but became statistically insignificant once demographic controls were added. This indicates that broader structural characteristics of counties could account for the observed partisan pattern.

Correlational analysis further shows that manufacturing employment share is positively associated with Republican vote share ($r = 0.21$) and closely tracks tariff exposure, indicating that manufacturing share is a plausible mechanism linking partisanship and exposure. At the same time, interaction results indicate that the marginal effect of GOP vote share declines as income rises: low-income Republican counties face disproportionately high tariff exposure to a degree that persists even when demographic controls are included. In the fullest specification, the marginal effect of Republican vote share turns from positive to negative at a county median household income of roughly $61,000, near the middle of the observed range. Thus, the apparent null relationship after demographic controls may mask substantial differences by income, with high-income Republican counties appearing relatively insulated from tariff exposure while low-income Republican counties bear a much larger retaliatory tariff burden.

Intent is notoriously difficult to establish from outcome data alone. These results therefore do not necessarily dispel a political targeting model. They may instead indicate that high-income Republican counties, whose economies are less concentrated in tradable goods, are structurally insulated from trade shocks. These findings cannot speak to the intent of foreign strategists, but they do indicate a substantial disparity in unadjusted partisan trade exposure. That distinction is the one drawn at the outset: this analysis establishes where retaliatory exposure falls, not why particular products were selected, and the second question would require evidence

of a different kind. This pattern is consistent with low-income Republican counties being disproportionately concentrated in goods-producing, manufacturing-based economies. The hypothesis was therefore only partially supported. Although Republican-leaning counties showed higher exposure in unadjusted comparisons, the relationship weakened after demographic and socioeconomic controls were added, suggesting that local economic demographics, especially manufacturing employment share, play an important role in observed tariff-exposure patterns.

Relation to Prior Literature

In this unprecedented trade war of global scale, these findings help address a gap in the existing literature regarding the impacts of multilateral trade conflicts. Previous findings primarily assess the strategies used in bilateral conflicts. Kim and Margalit, as well as Fetzer and Schwarz, found evidence of deliberate political targeting in the 2018–2019 trade war, particularly by China. The present study finds a similar raw correlation in 2025 data, but the relationship is explained by controls in a way that does not require a deliberate targeting interpretation.

Several factors may account for this difference. The 2025 retaliatory regime was multilateral and involved countries whose trade relationships with the United States differ considerably from China's. The breadth of the underlying U.S. tariff increases may also have left retaliating countries with less room to selectively target politically sensitive industries, as nearly all traded goods were subject to the 10% baseline tariff. It is also possible that the political targeting found in prior work operated through electoral mechanisms, such as politically important counties, that are more difficult to detect in the cross-sectional exposure data used here. Fetzer and Schwarz also found that during an earlier EU–U.S. dispute the European Union targeted products made in politically important districts while minimizing domestic cost by selecting goods for which alternative suppliers were readily available; the absence of EU measures from the present dataset may therefore bias estimates of partisan exposure downward.

Limitations

Several limitations bear on how these results should be read. The analysis is cross-sectional and cannot speak to dynamic effects, including whether tariff exposure influenced subsequent electoral outcomes. One constraint is definitional to the index itself. Because the Tariff Exposure Index is constructed from county employment shares, the association between manufacturing employment and tariff exposure in Figure 7b is partly definitional. Manufacturing dominates both tariff lines and tariff-weighted export value, so a county's index value largely tracks its manufacturing share by construction; the substantive link is the first one, that Republican-leaning counties have higher manufacturing employment ($r = 0.21$), which is measured independently of the index.

Exposure is also measured from schedules rather than from trade. The index records tariff rates as enacted rather than realized commerce, and weights industries by total national exports rather than by exports to retaliating countries, so countries trading primarily with non-retaliating partners may be assigned exposure they do not in practice face.

The dataset itself is heterogeneous, pooling measures imposed at different times, from the 2018 Section 232 and Section 301 disputes through the April 2025 reciprocal tariff order, and records only measures in force. The legacy measures are small in the aggregate, accounting for 112 of the 8,172 tariff lines, so the exposure index is driven almost entirely by measures imposed in 2025; the legacy measures matter chiefly for comparisons across retaliating countries. The European Union and Mexico are absent because the database records only measures in force, the former's measures having been suspended and the latter having imposed none. Future work should incorporate EU tariff data and examine whether exposure patterns differ across retaliating blocs.

Finally, the models rely on observational controls. Unmeasured county characteristics correlated with both partisanship and industrial structure may account for part of the association, and several controls, notably educational attainment and income, are themselves shaped by local industry composition, so adjusting for them may absorb part of the relationship the analysis seeks to measure. The adjusted estimates should therefore be read as descriptive rather than causal.

Conclusion

The retaliatory tariffs now in effect against the United States are broader and more multilateral than in previous U.S. trade conflicts, and this study analyzes their political geography at the county level. Republican counties were disproportionately exposed. The counties that bore the brunt were those most closely tied to manufacturing concentration and lower educational attainment. Countries most exposed to retaliation also voted most heavily for the administration whose protectionist policies provoked the retaliation in the first place (Hartig et al.). Whether this pattern reflects deliberate foreign political targeting, the natural geography of tradable industries, or some combination remains an open and consequential question.

Overall, the partisan exposure gap appears to follow the economic geography of tradable industries more than Republican vote share alone. The costs of trade conflict therefore fall on the same manufacturing communities that supply much of the political support for it, a consequence policymakers should weigh when considering trade policy. Future research should incorporate EU and Mexican tariff data and examine whether county-level tariff exposure translates into electoral responses in later elections.

Acknowledgements

The author acknowledges Dr. Eric VanLoo-Hsu for his supervision of this research, and Mr. Ryan McNickle for his supervision of the independent research process.

Works Cited

Amiti, Mary, et al. “The Impact of the 2018 Tariffs on Prices and Welfare.” Journal of Economic Perspectives, vol. 33, no. 4, 2019, pp. 187–210, https://doi.org/10.1257/jep.33.4.187.

Autor, David H., et al. “The China Syndrome: Local Labor Market Effects of Import Competition in the United States.” American Economic Review, vol. 103, no. 6, 2013, pp. 2121–68, https://doi.org/10.1257/aer.103.6.2121.

Autor, David H., et al. “Importing Political Polarization? The Electoral Consequences of Rising Trade Exposure.” American Economic Review, vol. 110, no. 10, 2020, pp. 3139–83, https://doi.org/10.1257/aer.20170011.

Blanchard, Emily J., et al. “Did Trump’s Trade War Impact the 2018 Election?” Journal of International Economics, vol. 148, 2024, article 103891, https://doi.org/10.1016/j.jinteco.2024.103891.

Brutger, Ryan, et al. “Trade Wars and Election Interference.” The Review of International Organizations, vol. 18, no. 1, 2023, pp. 1–25, https://doi.org/10.1007/s11558-022-09464-2.

Fajgelbaum, Pablo D., et al. “The Return to Protectionism.” The Quarterly Journal of Economics, vol. 135, no. 1, 2020, pp. 1–55, https://doi.org/10.1093/qje/qjz036.

Fetzer, Thiemo, and Carlo Schwarz. “Tariffs and Politics: Evidence from Trump’s Trade Wars.” The Economic Journal, vol. 131, no. 636, 2021, pp. 1717–41, https://doi.org/10.1093/ej/ueaa122.

Grossman, Gene M., and Elhanan Helpman. “Protection for Sale.” The American Economic Review, vol. 84, no. 4, 1994, pp. 833–50. JSTOR, https://www.jstor.org/stable/2118033.

Hartig, Hannah, et al. “Behind Trump’s 2024 Victory: A More Racially and Ethnically Diverse Voter Coalition.” Pew Research Center, 26 June 2025, https://www.pewresearch.org/politics/2025/06/26/behind-trumps-2024-victory-a-more-racially-and-ethnically-diverse-voter-coalition/.

International Trade Administration. Foreign Retaliations Database. U.S. Department of Commerce, https://www.trade.gov/foreign-retaliations-database. Accessed 1 Oct. 2025.

International Trade Administration. “Foreign Retaliations Timeline.” U.S. Department of Commerce, https://www.trade.gov/feature-article/foreign-retaliations-timeline. Accessed 1 Oct. 2025.

Irwin, Douglas A. Clashing over Commerce: A History of U.S. Trade Policy. U of Chicago P, 2017, https://doi.org/10.7208/chicago/9780226399010.001.0001.

Kim, Sung Eun, and Yotam Margalit. “Tariffs as Electoral Weapons: The Political Geography of the US–China Trade War.” International Organization, vol. 75, no. 1, 2021, pp. 1–38, https://doi.org/10.1017/S0020818320000612.

Liao, Steven, et al. Concordance: Product Concordance. Version 2.0.0, 2020, https://CRAN.R-project.org/package=concordance. R package.

Martin, Alberto, and Wouter Vergote. “On the Role of Retaliation in Trade Agreements.” Journal of International Economics, vol. 76, no. 1, 2008, pp. 61–77, https://doi.org/10.1016/j.jinteco.2008.03.009.

MIT Election Data and Science Lab. County Presidential Election Returns, 2000–2024. Harvard Dataverse, 2018, https://doi.org/10.7910/DVN/VOQCHQ.

R Core Team. R: A Language and Environment for Statistical Computing. Version 4.4.1, R Foundation for Statistical Computing, 2024, https://www.R-project.org/.

Treisman, Rachel. “Word of the Week: Trump Calls Tariffs ‘the Most Beautiful Word.’ Here’s Its History.” NPR, 9 Apr. 2025, https://www.npr.org/2025/04/09/nx-s1-5355661/tariffs-history-meaning.

United States, Census Bureau. American Community Survey 5-Year Estimates, 2018–2022. 2022, https://www.census.gov/programs-surveys/acs.

United States, Census Bureau. County Business Patterns. 2023, https://www.census.gov/programs-surveys/cbp.html.

United States, Census Bureau. USA Trade Online. 2024, https://usatrade.census.gov/.

United States, Executive Office of the President. Executive Order 14257: Regulating Imports with a Reciprocal Tariff to Rectify Trade Practices That Contribute to Large and Persistent Annual United States Goods Trade Deficits. 2 Apr. 2025. Federal Register, vol. 90, 7 Apr. 2025, pp. 15041–49.

Quiet on the Set: A Task-Based Analysis of the Impact of AI on Workers in the US Motion Picture and Video Industries

By Tyler Frank [1]*, Mentor By Kenneth Frank[2], Jane Park[3]

Abstract

This paper explores the impact of general-purpose artificial intelligence systems on creative occupations within the US motion picture and video industries, defined as the intersection of NAICS 512100 and SOC-major group 27-0000, including the impact of software built to incorporate and/or enhance artificial intelligence capabilities. Building upon previously published task-based exposure methods, this work created an updated rubric reflecting the rapid growth of AI systems in the past three years, as well as their specific applications in this sector. This rubric was then applied to O*NET database task statements, with Claude Fable 5 and GPT-5.6 Sol scoring independently as validation checks. The findings indicate an employment-weighted 39% exposure level across 26 of 27 occupations with reported employment. Additionally, wage-weighting scores indicate that this exposure is spread across income levels industry-wide and that little correlation exists between income and exposure. The work concludes that AI can have a transformative impact on these professions but makes no predictions as to the timeline or level at which these technologies may be adopted.

Introduction

Stakes

Since the release of ChatGPT in November of 2022 (OpenAI, "Introducing ChatGPT"), AI systems have rapidly progressed, inciting fears of labor replacement and job loss. Various studies were conducted regarding this, but sufficient innovation has occurred in the field such that these results are now dated, and the predictions they made regarding further adoption are now able to be evaluated.

Furthermore, these studies all focus on the entirety of the United States economy. While this economy-wide reporting is beneficial, it is distinct from reporting on a single sector within the broader economy. The roles in any of these sectors, and the individuals who perform them, are different from any other sector and thus AI exposure is heterogeneous, rather than being standard across the entire economy.

In recognition of this, the following paper seeks to build upon the prior studies conducted, adopting the classification system of one prior work (Eloundou et al.). This classification system was then applied to the film industry. This sector of the economy is highly unionized, and these unions have been some of the loudest voices concerning AI to date. Many

1. Jesuit High School, 4701 North Himes Ave, Tampa, FL, 33614, USA
2. Molloy University, 1000 Hempstead Ave, Rockville Centre, NY, 11570, USA
3. University of Southern California, 3551 Trousdale Pkwy, Los Angeles, CA, 90089, USA

*tfrank33764@gmail.com

of the largest film occupation unions have released comprehensive statements regarding AI and collective bargained around its potential impacts (International Alliance of Theatrical Stage Employees; SAG-AFTRA; Directors Guild of America, *2026 Memorandum of Agreement*; Writers Guild of America). Thus, an investigation into the extent to which occupations within this industry are exposed would be a beneficial contribution to those who are employed in it, and that the results of this study contribute to the broader field of research.

What This Paper Does

This paper conveys the application of a novel exposure rubric built in July of 2026 to 645 O*NET task statements that span across 27 BLS OEWS occupations at the intersection of NAICS 512100 (Motion Picture and Video Industries) and SOC-major group 27-0000.

This rubric classifies tasks into one of three categories. The first of these categories, E0 indicates that the task is not exposed to AI, that is it does not score in the other two categories. The second category, E1 describes direct exposure to a general purpose AI system, in that a competent worker can use just this model without further specialization to perform the task in half the current time required or less. Finally, the E2 category shows an indirect exposure whereby a worker can perform the task with AI, halving the time to completion, but only when additional software specializes the AI capabilities to that specific task. A simple example of this could be software which uses AI to generate a song. This distinction between E1 and E2 is important, as within it lies a difference in adoption likelihood. While E1's requirement is merely access to a frontier model, E2 necessitates further specialized software, which may or may not be built and/or widely accessed.

The application of this rubric was performed by three raters: the author, Claude Fable 5, and GPT-5.6 Sol. Once scoring was completed and discrepancies were resolved, exposure was aggregated to the occupation level at three levels: Conservative, Standard, and Expansive. These three levels differ in their weighting of the classification of E2 tasks. The Conservative level excludes E2, the Standard Level includes half of the E2 tasks, and the Expansive level assumes full E2 exposure. These calculations follow the system of the above-mentioned work and reveal the level of exposure based upon different extents of this software creation.

Contribution

This work contributes three things. The first thing, largely explained above, is reporting on AI exposure within a single sector of the economy, as opposed to the entirety of it. Second, comparison. By adopting and refining a prior work's classification, this paper can demonstrate how task classifications shift over time and attribute these shifts to capability growth or rubric changes. Finally, a rubric which puts exact limits on current capabilities. This refinement allows for further exploration of classification shifts in the future.

Literature Review

This work sits between three distinct categories of prior research: task-based automation exposure, AI specific exposure indices, and film-industry creative labor analysis.

Task-Based Measurement of Automation Exposure

This paper determines exposure at the task level, a well-established method of determining exposure first performed by Autor, Levy, and Murnane. This method was specifically chosen over others, such as scoring at the occupation level, as performed by Frey and Osborne. The estimate which those authors put forth, namely that 47% of all jobs in the US were at risk of computerization, was heavily contested when further analysis was performed at the task level. In fact, Arntz, Gregory, and Zierahn estimated that only 9% of US workers were in high-risk jobs (14). Furthermore, task level exposure was selected in recognition of the finding that technology, while automating tasks and replacing labor, can also create new tasks, maintaining or potentially expanding the labor pool (Acemoglu and Restrepo).

AI-Specific Exposure

Beyond expanding upon previous studies which utilized task statements to measure exposure, this effort sits within a long line of work which quantifies exposure to AI. Brynjolfsson and Mitchell developed a rubric to measure exposure to machine learning. This rubric was then applied to O*NET detailed work activities (DWAs) in the following year, showing that exposure was essentially uncorrelated with wages ($r = -0.14$) (Brynjolfsson et al.). Webb measured the overlap between patents and O*NET task statements, finding that across the entire US economy, AI exposure is greater in high-skill tasks.

Felten, Raj, and Seamans presented an alternative method by creating an AI exposure index which linked AI progress benchmarks to 52 O*NET abilities based upon survey responses. However, this method was rejected in favor of tasks statements because of the greater granularity of tasks, which allows for a better understanding of changes and advancements made in AI automation over time.

Where Brynjolfsson, Mitchell, and Rock find exposure essentially uncorrelated with wages, Eloundou et al. report exposure rising with both wage and education. This work analyzed O*NET task statements for the entire US economy, introducing a rubric and aggregation system that were adapted in this paper. Furthermore, this work provided scoring metrics which this work used to analyze changes in task classification over time.

Film Industry Creative Labor Analysis

Beyond economy-wide AI exposure, this work is situated in various efforts to analyze creative labor within the film industry. Caves organize creative production around a variety of structural conditions, such as demand uncertainty, artists' non-monetary preferences for the work, and complementarity among specialists' inputs. This literature, however, is descriptive and is not measured or quantified.

AI impacts were quantified in the film industry in Future Unscripted, which predicted that over 21% of film, television, and animation jobs would be disrupted by 2026 based upon a survey of 150 executives and managers industry wide (CVL Economics 34).

This work, unlike prior research, focuses on quantifying a single, homogenous sector of the economy by applying an exposure rubric to O*NET task statements.

Methods

Task Statements

Task Statements were collected from the O*NET 29.0 database (National Center for O*NET Development). A task statement describes a task performed in the broader occupation it corresponds to. It can be classified as "core" or as "supplementary". For this work, 645 of these task statements were collected.

Employment and Wage Data

Data on employment totals and mean wages were collected from the May 2025 release of the Bureau of Labor Statistics' Occupational Employment and Wage Statistics program (BLS OEWS) (U.S. Bureau of Labor Statistics, *Occupational Employment and Wage Statistics*). Data is classified by both industry (NAICS, or North American Industry Classification System) and by occupation (SOC, or Standard Occupation Code) (Office of Management and Budget; U.S. Bureau of Labor Statistics, *2018 Standard Occupational Classification System*).

Chosen Occupations

The subject of this paper is the intersection of NAICS 512100, the motion picture and video industry, and SOC major group 27-0000 (and its corresponding O*NET group), covering Arts, Design, Entertainment, Sports, and Media Occupations. This combination yields 32 distinct OEWS occupations. Of these 32, 27 were scored. The 5 occupations excluded from scoring were "All Other" categories which encapsulated roles not held in standard occupations. Consequently, their tasks were not included in the O*NET database, and they thus could not be scored by the methods of this paper. Their employment totals represented only 7.5% of the intersection's total employment, which was 132,470.

Furthermore, both employment and wage statistics were suppressed in BLS reporting on certain occupations. BLS occupation Interpreters and Translators (SOC Code 27-3091) was not given an employment statistic and instead received a mark of "**", all 26 other occupations were given an employment total. Similarly, 7 of the 27 occupations were not given a mean wage statistic and instead were denoted "*" in fields for wage statistics. These 7 occupations were Actors (SOC Code 27-2011); News Analysts, Reporters, and Journalists (27-2023); Musicians and Singers (27-2042); Court Reporters and Simultaneous Captioners (27-3092); Broadcast Announcers and Radio Disc Jockeys (27-3011); Music Directors and Composers (27-2041); and

Dancers (27-2031). Besides these 7 occupations named, all BLS occupations included both employment and wage statistics.

Exposure Rubric

The results of this paper were found by scoring these O*NET tasks with an exposure rubric. This rubric was created in July of 2026, but its structure borrows heavily from Eloundou et al. The rubric grades each task individually upon the ability for a general-purpose AI system to reduce the time required to complete the task, at the same quality level, by half. A summary of the rubric is provided below.

No Exposure (E0)

Time required to complete tasks at current standards cannot be halved by access to AI systems.

Pure Exposure (E1)

Access to a general-purpose AI system, as defined, can reduce the time required to complete the task (including reviewing/fixing as necessary) by at least half, while maintaining the current level of output.

Exposure by Extended Software (E2)

Access to a general-purpose AI system cannot reduce the time required to complete the task by half. But access to a separate software product built to incorporate the capabilities of a frontier model can reduce the required time by half and maintain the output level.

This rubric is capability-based, meaning that tasks were scored based upon the capability of AI to perform it. The likelihood of any given task to be automated was not considered a factor. The current frontier model as of 2026 was defined to have the following capabilities: chat features, agentic coding, agentic web search, spreadsheet and document assistance, specialized skills, and MCP connectors (Anthropic, “Claude Fable 5 and Claude Mythos 5”; OpenAI, “GPT-5.6”).

Task Weighting

Each task was weighted based upon its classification. “Core” tasks were assigned a value of 2, while “supplementary” tasks were given a value of 1. Furthermore, tasks which were not classified were assigned a value of 1. These weightings were applied to each task for aggregations to the occupation level. This choice was borrowed from Eloundou et al. In total, there were 1,065 weight units (420 tasks assigned a value of 2, 225 tasks assigned a value of 1).

Scoring Protocol

All tasks were originally scored by the author manually. Once scores were assigned for all tasks, the tasks were then scored independently by both Claude Fable 5 and GPT-5.6 Sol. For this test, LLMs received the rubric text used by the human scorer verbatim. The LLMs were called once per occupation and were given the occupation title and task statement. The LLM then scored these tasks, without access to human scores, their own previous scores, and the scores of the other

LLM. This amounted to a total of 31 calls per model, as some BLS occupations are split into multiple O*NET occupations.

This scoring was performed in July of 2026 through the API services of both Anthropic and OpenAI, respectively (Anthropic, "Claude Fable 5 and Claude Mythos 5"; OpenAI, "GPT-5.6"). The default temperature of each model was maintained, as neither model's service accepted the parameter in the versions used. An audit trail was produced alongside the results, which were provided in a structured JSON format.

These LLM results served as a robustness check, validation that the manual scoring applied the rubric consistently and correctly. Table 1 contains a comparison of results between each of these three scorers. It is important to note that the agreement between the two models is not viewed as independent validation.

Table 1. Inter-scorer agreement on all 645 task statements. Human score "blind" refers to the original human results, before disagreements were resolved. κ = Cohen's kappa (Cohen, "A Coefficient of Agreement"; Cohen, "Weighted Kappa"). Generated in Python 3.12.10 with Matplotlib 3.11.1 and NumPy 2.4.6 (Python Software Foundation; Hunter; Harris et al.).

Pair	% agreement	Unweighted κ	Quadratic κ (reordered)
Human (blind)-Claude	82.5%	0.706	0.771
Human (blind)-GPT	78.4%	0.706	0.736
Claude-GPT	84.0%	0.744	0.827

κ = Cohen's kappa. "Blind" denotes human scores recorded before any model output was seen. The quadratic κ is computed on the reordered scale E0 < E2 < E1.

The results of this table indicate substantial agreement in all three pairs. Additionally, it should be noted that the percentages of agreement are comparable to those of Eloundou et al.

The quadratic kappa (reordered) metric is calculated in the traditional way (Cohen, "A Coefficient of Agreement"; Cohen, "Weighted Kappa"), except that the ordering for the statistic was modified, so that the ordering is E0 < E2 < E1. This choice was made to more accurately account for vast disagreements, as E2 is closer to E0 than E1 is. This decision was made because E2 and E0 share the belief that a current general purpose AI system cannot perform the given task, while E0 and E1 do not agree upon that, nor anything else.

Intra-Rater Reliability

In addition to the external checks deployed in this project, namely scoring performed by LLMs, an intra-rater check was performed to ensure that the rubric was faithfully applied to all task statements.

This check involved the rescoring of 50 tasks, at a date after the original scoring. These 50 tasks were drawn by implementing a stratified random sample from the blind score distribution (E0 = 25, E1 = 17, E2 = 8) with a seed of 42. This seed was selected at random.

The results of this rescore were recorded and then compared against the original results to ensure consistency. This comparison is shown in Table 2.

Table 2. Intra-rater agreement. Generated in Python 3.12.10 with Matplotlib 3.11.1 and NumPy 2.4.6 (Python Software Foundation; Hunter; Harris et al.).

Pair	% agreement	Unweighted κ	Quadratic κ (reordered)
Original-rescore	90.0%	0.844	0.935

Fifty tasks drawn by stratified random sample (seed 42) and rescored blind after the original pass. Abbreviations as in Table 1.

Index Construction

After scoring was completed, the final human results, after disagreements were resolved, were aggregated to their corresponding occupations at three different levels of exposure. Each of these are meant to represent exposure to AI automation at different extents.

Conservative Exposure is calculated by counting all tasks with a classification of E1 as exposed, while excluding those scored E2 (E0 = 0, E1 = 1, E2= 0). This score is the lowest of the three and is meant to represent the lower bound of AI exposure within the industry.

Standard Exposure is calculated by counting all tasks with a classification of E1 and half of the tasks with a score of E2 as exposed (E0 = 0, E1 = 1, E2= 0.5). This decision is an arbitrary one, not an estimate of partial adoption. It was selected because it is the midpoint of the conservative and expansive levels of exposure, and for consistency's sake because it was used in Eloundou et al.

Expansive exposure is calculated by counting all tasks with a classification of E1 and of E2 as exposed (E0 = 0, E1 = 1, E2= 1). This result would require total adoption of AI systems, including a variety of external software systems beyond a tradition system and is thus the most extreme level of AI exposure imagined in this work. It represents an upper bound or maximum level of AI exposure.

Once these exposure metrics were calculated, an index was created, aggregating the occupational exposures into industry-wide exposure metrics. This was done in four different ways: unweighted occupational mean, employment-weighted, employment-weighted (wage sample), and wage-bill-weighted.

The unweighted occupational mean is a simple average of the exposure level of each of the 27 occupations. The employment-weighted statistic is calculated by taking the summation of the product of each occupation's exposure by its total employment and dividing it by the summation of each of the occupations employment, or the total employment of the industry. The employment-weighted (wage sample) metric is calculated the same way, but on a modified sample to mirror that of the wage-bill-weighted stat. Wage-Bill-Weighted numbers are calculated by taking the summation of the product of the exposure level of each occupation by its total

employment times its mean wage and dividing it by the summation of the product of its mean wage by its total employment.

Of the twenty-seven occupations scored, one (Interpreters and Translators) did not have a disclosed total employment, while seven occupations (Actors; News Analysts, Reporters, and Journalists; Musicians and Singers; Court Reporters and Simultaneous Captioners; Broadcast Announcers and Radio Disc Jockeys; Music Directors and Composers; and Dancers) did not have disclosed mean wage statistics.

Due to these gaps in BLS OEWS data, employment-weighted statistics included only twenty-six occupations. Likewise, wage-weighted statistics included nineteen occupations. To account for this discrepancy, the employment-weighted (wage sample) category was created to standardize the results between employment-weighted and wage-bill-weighted for comparison.

Disagreement Resolution

After scoring was performed by hand and with LLMs, disagreements were found. Of these disagreements, tasks which had the result of two LLMs in agreement against one human were reviewed. For this review, the human scorer was provided with just the tasks which required rescoring, and with comments explaining the reasoning of both LLMs. With this information, the human rescored each of these tasks, providing a comment explaining his final decision, whether he upheld his original decision or switched in agreement with the AI models. Of the 645 tasks originally scored, 59 were revised to the decision of the LLMs (9.1%). The agreement statistics above were calculated using the original human scores, denoted Human Score (Blind).

Results and Discussion

Task Level Score Distribution

The 645 tasks scored corresponded to a total of 1,065 weight units. The distributions of these tasks and these units are shown in Tables 3 and 4, respectively.

Table 3. Unweighted task counts (n = 645). Generated in Python 3.12.10 with Matplotlib 3.11.1 and NumPy 2.4.6 (Python Software Foundation; Hunter; Harris et al.).

Rater	E0	E1	E2
Human (blind)	325 (50.4%)	220 (34.1%)	100 (15.5%)
GPT	270 (41.9%)	244 (37.8%)	131 (20.3%)
Claude	341 (52.9%)	221 (34.3%)	83 (12.9%)
Human (final)	318 (49.3%)	232 (36.0%)	95 (14.7%)

Unweighted task counts (n = 645). E0 = not exposed; E1 = exposed to a general-purpose model alone; E2 = exposed only via model-based software.

Table 4. Weight units (n = 1,065). Generated in Python 3.12.10 with Matplotlib 3.11.1 and NumPy 2.4.6 (Python Software Foundation; Hunter; Harris et al.).

Rater	E0	E1	E2
Human (blind)	532 (50.0%)	365 (34.3%)	168 (15.8%)
GPT	442 (41.5%)	402 (37.7%)	221 (20.8%)
Claude	560 (52.6%)	368 (34.6%)	137 (12.9%)
Human (final)	521 (48.9%)	383 (36.0%)	161 (15.1%)

Weight units (n = 1,065); core tasks weighted 2, supplementary 1. Categories as in Table 3.

As of July 2026, the largest group of tasks with regards to AI exposure are still those unexposed (48.9%). Yet, over a third of tasks (36%) are purely exposed to general purpose AI systems, and another 15.1% are exposed by extended software. Thus, when combining these groups, just over half of the tasks in creative occupations in the motion picture and video industries are exposed.

Occupation Level Exposure

The exposure level of each of the 27 occupations scored is displayed in Table 5. The results are displayed alphabetically.

Table 5. Occupation-level AI exposure scores, BLS mean employment, and BLS mean wages (U.S. Bureau of Labor Statistics, *Occupational Employment and Wage Statistics*). Generated in Python 3.12.10 with Matplotlib 3.11.1 and NumPy 2.4.6 (Python Software Foundation; Hunter; Harris et al.).

Occupation title	SOC code	Conser-vative	Standard	Expansive	Employment	Mean wage
Actors	27-2011	0.097	0.129	0.161	19,270	*
Art Directors	27-1011	0.536	0.571	0.607	3,080	$161,480
Audio and Video Technicians	27-4011	0.350	0.438	0.525	8,420	$75,230
Broadcast Announcers and Radio DJs	27-3011	0.256	0.349	0.442	170	*
Broadcast Technicians	27-4012	0.125	0.175	0.225	1,080	$92,450
Camera Operators, TV, Video, and Film	27-4031	0.167	0.200	0.233	7,580	$96,070
Court Reporters and Captioners	27-3092	0.458	0.458	0.458	230	*
Craft Artists	27-1012	0.367	0.383	0.400	130	$110,250
Dancers	27-2031	0.083	0.083	0.083	110	*
Editors	27-3041	0.588	0.618	0.647	7,540	$124,610
Fashion Designers	27-1022	0.394	0.394	0.394	1,110	$112,730
Film and Video Editors	27-4032	0.143	0.500	0.857	13,280	$93,720
Fine Artists	27-1013	0.361	0.361	0.361	840	$176,370
Graphic Designers	27-1024	0.639	0.778	0.917	2,020	$95,690
Interpreters and Translators	27-3091	0.680	0.720	0.760	**	$84,420
Lighting Technicians	27-4015	0.000	0.031	0.063	2,040	$114,120
Music Directors and Composers	27-2041	0.246	0.500	0.754	150	*
Musicians and Singers	27-2042	0.021	0.115	0.208	230	*
News Analysts, Reporters, and Journalists	27-3023	0.600	0.600	0.600	730	*
Photographers	27-4021	0.191	0.255	0.319	1,300	$112,500
Producers and Directors	27-2012	0.348	0.418	0.489	37,700	$133,970
Public Relations Specialists	27-3031	0.576	0.576	0.576	1,650	$101,290
Set and Exhibit Designers	27-1027	0.326	0.446	0.565	3,180	$104,210
Sound Engineering Technicians	27-4014	0.111	0.222	0.333	2,820	$100,450
Special Effects Artists and Animators	27-1014	0.316	0.632	0.947	5,910	$116,410
Technical Writers	27-3042	0.800	0.833	0.867	100	$72,650
Writers and Authors	27-3043	0.750	0.750	0.750	1,890	$103,840

* wage suppressed by BLS; ** employment suppressed by BLS. Suppression is why the employment-weighted sample is n = 26 and the wage-bill-weighted sample n = 19.

Values of * and ** indicate suppression of the statistic being reported. These suppressions are why employment-weighted populations are limited to size n=26 and wage-bill-weighted samples are limited to size n=19.

At a conservative and standard level, the most exposed occupation is Technical Writer. At the expansive level, it is Special Effects Artists and Animators. The least exposed occupation at all three levels is Lighting Technician. Occupational exposure for the population at the standard level ranges from 0.031 to 0.833 with a median of 0.438 and a standard deviation of 0.220.

11 of the 27 occupations scored have at least half of their tasks exposed to AI at a standard level. This figure jumps to 13 of the 27 occupations at an expansive level of exposure.

Furthermore, 20 of the 27 occupations have a standard exposure of at least 0.25. In fact, at the standard level all 27 occupations have at least one task exposed and 25 of those occupations have a core task exposed. The only occupations without a core task exposed are Camera Operators and Lighting Technicians.

Bound Width

Bound width is, by construction, the share of E2 tasks in an occupation. Because conservative exposure only counts E1 tasks and Expansive Exposure counts both E1 and E2 tasks alike. Occupations therefore fall into two extremes, those who only have exposed tasks scored E1 and thus have no bound width and those who have all or most exposed tasks labeled E2 and thus have a wide bound width.

Seven occupations have this first extreme of no bound width: Court Reporters, Dancers, Fashion Designers, Fine Artists, News Analysts, PR Specialists, Writers and Authors. At the other extreme, three occupations have bound widths greater than 0.5: Film and Video Editors, Special Effect Artists and Animators, and Music Directors and Composers. These differences in Bound Width are made available in Figure 1.

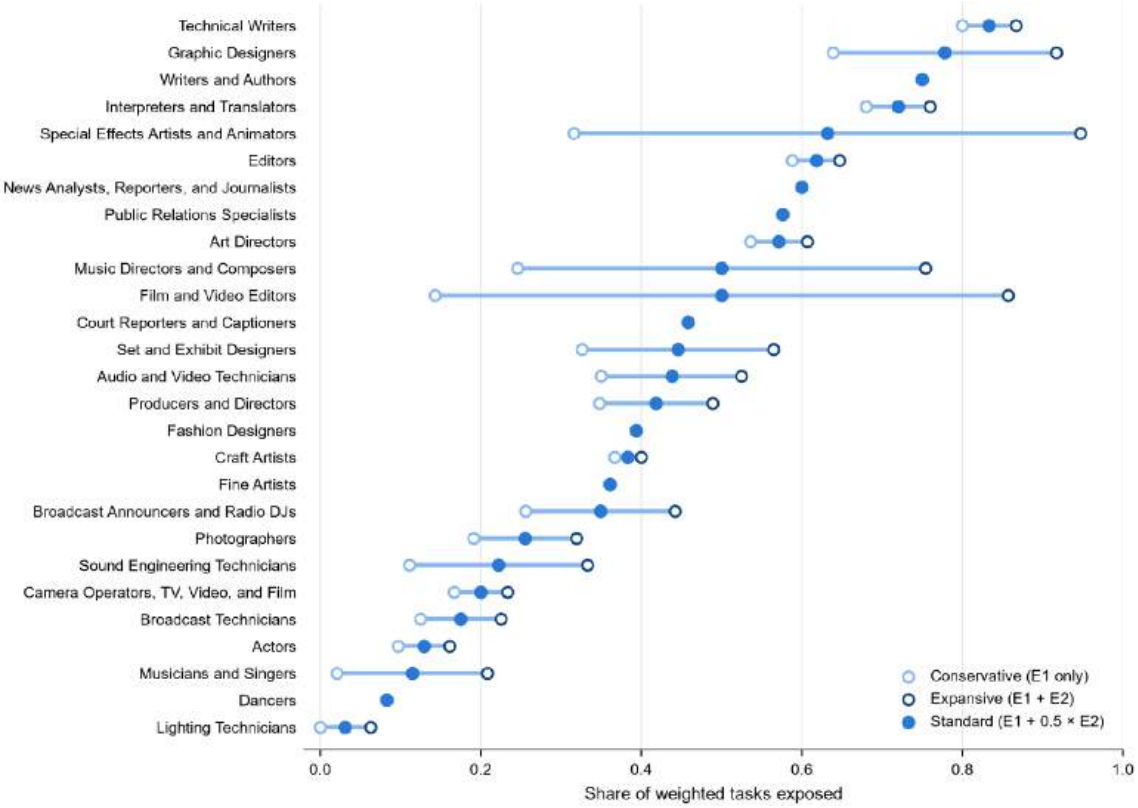


Figure 1. Bounds at conservative, standard and expansive exposure for each of the 27 occupations scored. The width of each bound is indicated by the space between the dots with open circles. Generated in Python 3.12.10 with Matplotlib 3.11.1 and NumPy 2.4.6 (Python Software Foundation; Hunter; Harris et al.).

Index

After aggregation to the occupation level, an index was created to further consolidate the results into sector wide exposure figures. This index features four fields, which are enumerated above. The entirety of the index is included in Table 6.

Table 6. Sector-wide AI exposure index. Generated in Python 3.12.10 with Matplotlib 3.11.1 and NumPy 2.4.6 (Python Software Foundation; Hunter; Harris et al.).

Field	Conservative	Standard	Expansive
Unweighted mean occupational exposure (n = 27)	0.353	0.427	0.502
Employment-weighted index (n = 26)	0.293	0.395	0.497
Employment-weighted index, wage sample (n = 19)	0.329	0.445	0.561
Wage-bill-weighted index (n = 19)	0.339	0.447	0.555

Sector-wide exposure. Sample sizes differ because of BLS suppression (see Table 5).

The Unweighted Mean Occupational Exposure field was calculated by computing the average exposure of each of the 27 occupations, treating each equally and disregarding BLS OEWS data. This field showed an average occupational exposure of 0.427, indicating that the average occupation has just under half of their tasks exposed to artificial intelligence.

The Employment-Weighted Index computed the industry's exposure with regards to employment totals. This metric accounts for differences in the number of employees found in these occupations and thus constructs a more accurate representation of the percentage of employees exposed to AI. This figure, at 0.395, is less than the simple unweighted average of 0.427, showing that the entire industry is slightly less exposed when accounting for employment differences between occupations. This decrease is driven primarily by Actors, who account for 15.7% of covered employment at a standard exposure of 0.129.

The Employment-Weighted Index, Wage Sample was constructed for the purpose of comparisons between employment-weighted statistics and wage-bill-weighted statistics. When comparing these two, similar results are seen. Of the three levels of exposure, the largest gap occurs at the conservative level, and this gap is only 0.010 index points. Because these two indices are computed over the same nineteen occupations and have such negligible differences, a comparison of them indicates that the exposure is not concentrated at either end of the sector's wage distribution but rather spread across the entirety of the wage spectrum. To explore this relationship further, a linear regression is displayed in Figure 2.

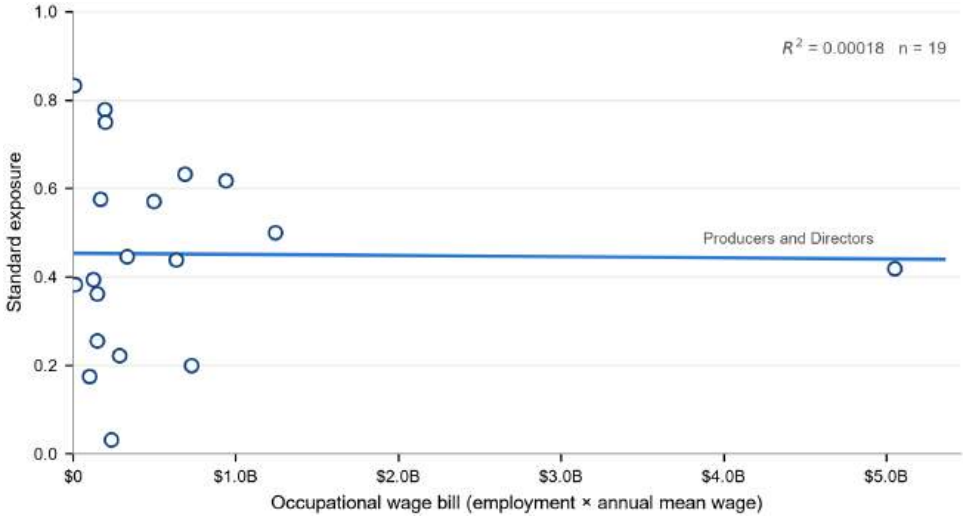


Figure 2. OLS regression of standard AI exposure on occupational wage bill (n = 19). Wage bill is the product of BLS OEWS employment and annual mean wage. Generated in Python 3.12.10 with Matplotlib 3.11.1 and NumPy 2.4.6 (Python Software Foundation; Hunter; Harris et al.).

By looking at this graph, two things become clear. First, Producers and Directors as an occupation is a large outlier from the other occupations. Its wage bill sits at $5.05 Billion, over $3.8B greater than the next largest wage bill, belonging to film and video editors at $1.2B. Second, wage bill is uncorrelated with exposure in this sample. The regression above has an r-squared value of 0.00017, showing that essentially none of the variance between standard exposure is explained by a relationship with wage bill. Additionally, this conclusion holds when excluding Producers and Directors. A regression of this graph, shown in Figure 3, results in an r-squared value of 0.0055, still essentially negligible. Overall, it should be concluded that exposure is uncorrelated with wage bills in this index.

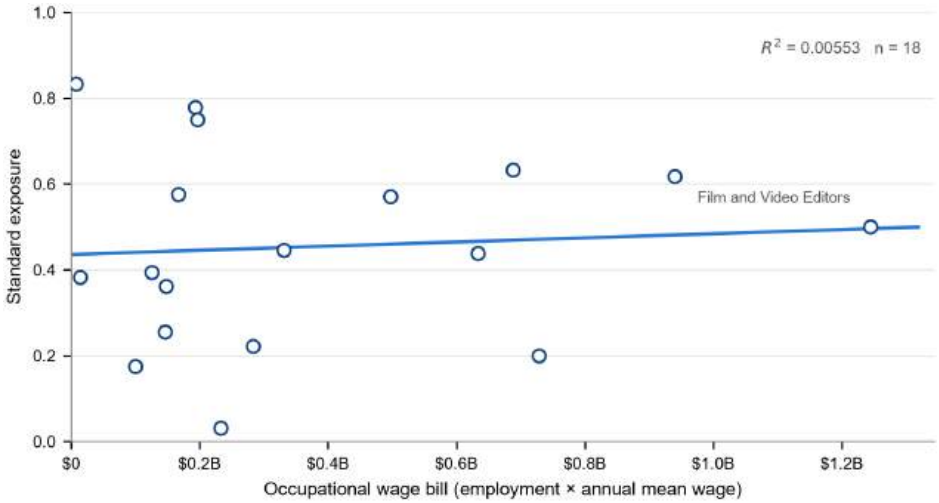


Figure 3. OLS regression of standard AI exposure on occupational wage bill, excluding occupation of Producers and Directors (n=18). Generated in Python 3.12.10 with Matplotlib 3.11.1 and NumPy 2.4.6 (Python Software Foundation; Hunter; Harris et al.).

E1 vs. E2 Realization

In this paper, there is a difference between E1 exposure and E2 exposure in that E1 exposure is more immediate. Because of this, occupations with similar exposure levels, but different compositions of E1/E2 tasks may have vastly different timelines and likelihoods of adoption. To illustrate this point, consider two occupations from this sample: Editors, and Film and Video Editors. These two tasks have similar Standard Exposure levels, 0.618 for Editors versus 0.500 for Film and Video Editors. Yet, at the conservative level Editors are exposed at a level 4x greater than Film and Video Editors and at the Expansive level Film and Video Editors are more exposed than Editors. Because of these differences in composition, Editors are currently more vulnerable to automation. Presently, any competent worker with access to a frontier model can perform the task of an Editor in half the time. This is not the case for Film and Video Editors, who have more of their tasks exposed, but solely to advanced software. This software is not as available and is not yet as developed, and thus this occupation is presently insulated from adoption to a greater extent.

Prior Estimates

The findings of this work will now be compared with those of Eloundou et al. For this comparison to take place, the task-level scoring results of that work were sourced (OpenAI, *GPTs-Are-GPTs*), then filtered such that their results align specifically with the occupations that were scored in this paper. However, it must be noted that the present work and Eloundou et al. are different works which contain different rubrics. Thus, this comparison is not one to one. Furthermore, it is important to mention that four tasks have been added to O*NET in the period between this paper's publication and the time of this work, and thus the comparison only features 641 tasks rather than the full 645 scored in this work.

Table 7 presents a matrix comparing their results against this paper's, by score. This first notable trend is a shift of tasks from E0 to either E1 or E2. 55 tasks of the 260 originally tagged unexposed were tagged with some extent of exposure in this work (~21%). This shift reflects the growth in AI capabilities between 2023 and 2026. Next, it must be acknowledged that some tasks shifted from directly exposed in their work (E1) to either indirectly exposed (E2) or unexposed (E0) in this work. Upon review, this shift appears to be due to rubric differences, rather than any trend. Specifically, 16 of the 29 tasks involved some form of guiding, training, or discussing something with other people. Furthermore, 8 of those remaining 13 involve some form of verbal communication. Both of these capabilities were considered out of scope of a general purpose AI system in this work, but that is not necessarily true for Eloundou et al. Finally, with regards to tasks deemed E2 by Eloundou, two important trends are revealed. First, 42% of those tasks were upgraded to E1 in this work. Once again, showing advancement in the capabilities of frontier AI models. Oppositely however, 83 of the 235 tasks originally scored as E2 (35%) were downgraded in this work to a score of E0. This trend reflects that tasks which may be able to be automated by AI-powered software are more insulated from adoption presently, as this software has lagged the growth of frontier models.

Table 7. Task-level exposure scores, Eloundou et al. (OpenAI, *GPTs-Are-GPTs*) versus this paper. Generated in Python 3.12.10 with Matplotlib 3.11.1 and NumPy 2.4.6 (Python Software Foundation; Hunter; Harris et al.).

Eloundou et al. (rows) / this work (columns)	E0	E1	E2	Row total
E0	205	24	31	260
E1	29	107	10	146
E2	83	99	53	235
Column total	317	230	94	641

Task-level crosswalk over the 641 tasks common to both studies. Categories as in Table 3.

Capability vs. Adoption

The tasks described as being exposed in this paper are tasks which are capable of being performed with AI, either directly with a frontier model or with extended software. However, it must be noted that this capability does not translate directly into adoption. Many other factors, beyond the scope of this work, play into the timeline and the extent to which these AI products are adopted. With these things in mind, it once again must be made clear that the purpose of this paper is not to predict adoption, but rather specifically to analyze what tasks are able to be automated by harnessing the capabilities of the current frontier models.

Furthermore, another major factor in the rate of adoption, which is unique to the film industry, exists, labor unions. Each of the four largest labor unions in the industry (IATSE, SAG-AFTRA, DGA, WGA) have released statements regarding AI and have bargained for protections against the technology (International Alliance of Theatrical Stage Employees; SAG-AFTRA; Directors Guild of America, *2026 Memorandum of Agreement*; Writers Guild of America). Of these four, SAG-AFTRA particularly stands out as they have strict policies in place regarding notification and consent for the creation of AI replicas of actors (SAG-AFTRA, sec. 64.A.(2)(a)). Similarly, the Directors Guild of America (DGA) bargained in June of 2026 such that directors now have control over AI-generated footage, and that they can attend employer funded skill enhancement programs (Directors Guild of America, "DGA National Board Unanimously Approves"; Directors Guild of America, "DGA Membership Ratifies"). All while previous safeguards from 2023 ensure that director's work must be continued to be performed in person (Directors Guild of America, "DGA National Board Unanimously Approves"). All the policies put out by these unions, and those they continue to adopt, will serve to limit the impact that artificial intelligence can have on labor in the industry. For this reason, specifically in this project's chosen occupation group, the capability of a model to perform a task does not inherently translate to AI being adopted to perform that task.

Limitations

Rater Design

The rubric of this project was created and subsequently applied to the 645 tasks by the same person. Furthermore, the validation of this single worker was two LLMs which cannot be viewed as two independent raters. Finally, disagreements between these results were resolved by the same human scorer. This rater design was accounted for by introducing the intra-rater reliability check, but this measured consistency rather than accuracy. Thus, a gap remained.

Task Statements

The weighting scheme of the O*NET Task Statements scored in this project are ordinal, rather than being based upon time. Thus, when aggregating to the occupation level two tasks which are both core are weighed the same, even if one takes twice as long to complete as the other. One method for accounting for this was O*NET's graded ratings for tasks, but this data was unevenly available across tasks and thus not a viable option.

Furthermore, these task statements themselves are abstractions, broad statements taken from surveys of workers. Thus, it is possible that the database obscures important parts of any given occupation. The potential occurrence of this is unknown and cannot be estimated.

Excluded Occupations

Five occupations which fit the criteria of the universe of this project were entirely excluded from scoring due to a lack of O*NET task statements. Beyond this, 1 of the 27 occupations scored had a suppression of employment totals, and 7 of the 27 occupations scored had a suppression of mean wage totals. This lack of public data limited the scope of this project, and the scope of the employment and wage-bill weighted metrics computed in its results.

Conclusion

Contribution Restated

This work contributes three things to the broader research field. First, a sector index. This work, unlike prior research, focuses on a single, homogenous sector. This allows a more focused analysis on a specific industry, rather than a broad, generalized report on the entire US economy. Next, a repeatable method of comparison. Because this work retains the three-level categorization of prior papers while defining each level more narrowly, individual tasks can be matched across studies and changes in categories can be attributed to either capability growth or rubric discrepancies. To that end, the third contribution of this work is found in the rubric, which puts precise limits on the capabilities of artificial intelligence as of the scoring date. This choice differs from prior works, which included only broader definitions on the capabilities of artificial intelligence.

Results Summarized

Upon weighing the 645 tasks scored, it was found that 51.1% of the weight units are exposed to AI automation across both the E1 and E2 levels. After aggregations to the occupation level were made, it was found that the average occupation was 42.7% exposed at the standard level, with levels of exposure ranging from 3% to 83%. Next, after constructing an index and weighing occupations based upon employment it was found that the industry is 39.5% exposed. Finally, this index standardized a population of nineteen occupations which had both employment and wage statistics, discovering approximately 44.5% exposure, with only a small margin between employment and wage-bill metrics. Furthermore, it was noted that composition of exposure (i.e., number of tasks exposed at E1 level versus E2 level) is an important factor in the realization of exposure presently.

Forward

Because artificial intelligence is rapidly improving, and its applications are rapidly expanding, these results can only provide a snapshot of current conditions. However, the rubric used in this project can be updated by replacing/augmenting the current capabilities. Thus, these same 645 tasks can be rescored at a future date, revealing the relevant improvements to frontier AI models for the creative economy while auditing the accuracy of predictions regarding software applications for AI.

Works Cited

Acemoglu, Daron, and Pascual Restrepo. "Automation and New Tasks: How Technology Displaces and Reinstates Labor." Journal of Economic Perspectives, vol. 33, no. 2, 2019, pp. 3–30, https://doi.org/10.1257/jep.33.2.3.

Anthropic. "Claude Fable 5 and Claude Mythos 5." 9 June 2026, https://www.anthropic.com/news/claude-fable-5-mythos-5. Accessed 25 July 2026.

Arntz, Melanie, et al. The Risk of Automation for Jobs in OECD Countries: A Comparative Analysis. OECD Publishing, 2016, https://doi.org/10.1787/5jlz9h56dvq7-en. OECD Social, Employment and Migration Working Papers No. 189.

Autor, David H., et al. "The Skill Content of Recent Technological Change: An Empirical Exploration." Quarterly Journal of Economics, vol. 118, no. 4, 2003, pp. 1279–1333, https://doi.org/10.1162/003355303322552801.

Brynjolfsson, Erik, and Tom Mitchell. "What Can Machine Learning Do? Workforce Implications." Science, vol. 358, no. 6370, 2017, pp. 1530–1534, https://doi.org/10.1126/science.aap8062.

Brynjolfsson, Erik, et al. "What Can Machines Learn, and What Does It Mean for Occupations and the Economy?" AEA Papers and Proceedings, vol. 108, 2018, pp. 43–47, https://doi.org/10.1257/pandp.20181019.

Caves, Richard E. "Contracts between Art and Commerce." Journal of Economic Perspectives, vol. 17, no. 2, 2003, pp. 73–84. JSTOR, https://www.jstor.org/stable/3216857.

Cohen, Jacob. "A Coefficient of Agreement for Nominal Scales." Educational and Psychological Measurement, vol. 20, no. 1, 1960, pp. 37–46, https://doi.org/10.1177/001316446002000104.

---. "Weighted Kappa: Nominal Scale Agreement with Provision for Scaled Disagreement or Partial Credit." Psychological Bulletin, vol. 70, no. 4, 1968, pp. 213–220, https://doi.org/10.1037/h0026256.

CVL Economics. Future Unscripted: The Impact of Generative Artificial Intelligence on Entertainment Industry Jobs. Commissioned by the Concept Art Association, the Animation Guild (IATSE Local 839), the Human Artistry Campaign, and the National Cartoonist Society Foundation, Jan. 2024, https://animationguild.org/wp-content/uploads/2024/01/Future-Unscripted-The-Impact-of-Generative-Artificial-Intelligence-on-Entertainment-Industry-Jobs-pages-1.pdf. Accessed 25 July 2026.

Directors Guild of America. "DGA Membership Ratifies New Contract by Overwhelming Margin." 26 June 2026, https://www.dga.org/news/pressreleases/2026/260625_dga_membership_ratifies_new_contract_by_overwhelming_margin. Accessed 25 July 2026.

---. "DGA National Board Unanimously Approves Tentative New Agreement." 12 June 2026, https://www.dga.org/news/pressreleases/2026/260612_dga_board_recommends_ratification_of_new_contract. Accessed 7 Sept. 2026.

Directors Guild of America, and Alliance of Motion Picture and Television Producers. 2026 Memorandum of Agreement for the DGA Basic Agreement and Freelance Live and Tape Television Agreement. Term beginning 1 July 2026, https://www.dga.org/Contracts/Agreements/2026-Memorandum-of-Agreement. Accessed 25 July 2026.

Eloundou, Tyna, et al. “GPTs Are GPTs: Labor Market Impact Potential of LLMs.” Science, vol. 384, no. 6702, 2024, pp. 1306–1308, https://doi.org/10.1126/science.adj0998.

Felten, Edward, et al. “Occupational, Industry, and Geographic Exposure to Artificial Intelligence: A Novel Dataset and Its Potential Uses.” Strategic Management Journal, vol. 42, no. 12, 2021, pp. 2195–2217, https://doi.org/10.1002/smj.3286.

Frey, Carl Benedikt, and Michael A. Osborne. “The Future of Employment: How Susceptible Are Jobs to Computerisation?” Technological Forecasting and Social Change, vol. 114, 2017, pp. 254–280, https://doi.org/10.1016/j.techfore.2016.08.019.

Harris, Charles R., et al. “Array Programming with NumPy.” Nature, vol. 585, no. 7825, 2020, pp. 357–362, https://doi.org/10.1038/s41586-020-2649-2. Version 2.4.6 used in this work.

Hunter, John D. “Matplotlib: A 2D Graphics Environment.” Computing in Science & Engineering, vol. 9, no. 3, 2007, pp. 90–95, https://doi.org/10.1109/MCSE.2007.55. Version 3.11.1 used in this work.

International Alliance of Theatrical Stage Employees, and Alliance of Motion Picture and Television Producers. 2024 IATSE Basic Agreement Memorandum of Agreement, Article XLIX (Artificial Intelligence). 3 July 2024, https://iatse.net/wp-content/uploads/2024/07/2024-IATSE-Basic-Agreement-MOA-FINAL.pdf. Accessed 25 July 2026.

National Center for O*NET Development. O*NET 29.0 Database. U.S. Department of Labor, Employment and Training Administration, Aug. 2024, https://www.onetcenter.org/db_releases.html. Accessed 25 July 2026.

Office of Management and Budget. North American Industry Classification System: United States, 2022. Executive Office of the President, 2022, https://www.census.gov/naics/reference_files_tools/2022_NAICS_Manual.pdf. Accessed 25 July 2026.

OpenAI. “GPT-5.6: Frontier Intelligence That Scales with Your Ambition.” 9 July 2026, https://openai.com/index/gpt-5-6/. Accessed 25 July 2026.

---. GPTs-Are-GPTs. Replication data and code, GitHub, 2023, https://github.com/openai/GPTs-are-GPTs. Accessed 25 July 2026.

---. “Introducing ChatGPT.” 30 Nov. 2022, https://openai.com/index/chatgpt/. Accessed 25 July 2026.

Python Software Foundation. Python. Version 3.12.10, 2025, https://www.python.org/. Accessed 25 July 2026.

SAG-AFTRA, and Alliance of Motion Picture and Television Producers. 2026 Theatrical-Television Memorandum of Agreement, Item 39 (Artificial Intelligence).

Signed 27 May 2026, ratified 5 June 2026, https://www.sagaftra.org/sites/default/files/2026-05/2026%20Theatrical-Television%20MOA.pdf. Accessed 25 July 2026.

U.S. Bureau of Labor Statistics. 2018 Standard Occupational Classification System. U.S. Department of Labor, 2018, https://www.bls.gov/soc/2018/. Accessed 25 July 2026.

---. Occupational Employment and Wage Statistics, May 2025. U.S. Department of Labor, 2026, https://www.bls.gov/oes/. Accessed 25 July 2026.

Webb, Michael. The Impact of Artificial Intelligence on the Labor Market. Working paper, Stanford University, Jan. 2020, https://www.michaelwebb.co/webb_ai.pdf. Accessed 25 July 2026.

Writers Guild of America, and Alliance of Motion Picture and Television Producers. Memorandum of Agreement for the 2026 WGA Theatrical and Television Basic Agreement, Article 72 (Generative Artificial Intelligence). 4 Apr. 2026, https://www.wga.org/uploadedfiles/contracts/2026_mba_moa.pdf. Accessed 25 July 2026.

Inflation as a Structural Barrier to Our Global Economy: Lessons from Argentina, The United States, and Countless Geopolitical Shocks By Keya Kumar

ABSTRACT

This paper delves into inflation as a structural barrier to global economic stability, through case studies from Argentina and the United States along with geopolitical shocks and their inflationary effects. It explores the root causes of inflation: demand-pull and cost-push mechanisms, monetary and fiscal policy failures, and supply chain vulnerabilities; this paper argues that inflation is not merely a cyclical issue in the global economy, but rather that it is a structurally embedded problem. Argentina's ongoing inflation crisis illuminates how foundational weaknesses and currency instability leads to severe inflation, while the United States reveals how even advanced economies are vulnerable to external shocks such as energy disruptions caused by geopolitical conflict. The paper further connects to how rising prices undermine decent work and prevent inclusive economic growth around the world. Comparative analysis using IMF data reinforces that inflation outcomes differ based on institutional strength and governance quality, showing that the gap between emerging and advanced economies is driven by how governments respond to shocks. Finally, coordinated solutions are proposed, including stronger central banks, more reliable supply chains, and full global cooperation. Ultimately, without sustained structural reform and solutions to combat inflation, inflation will continue to pose a barrier to the Global Economy.

Keywords: inflation, structural barriers, geopolitical shocks, monetary policy, global economy, cost-push inflation, demand-pull inflation

INTRODUCTION

In 2022, global inflation hit 8.7%, the world's highest inflation level in four decades. Around the world from the United States to Argentina, ordinary people watch firsthand as their money loses value faster than they can earn it. What previously appeared as a temporary post-COVID disruption has revealed itself as a structural vulnerability that is embedded in the global economy.

Inflation is the rise in prices of goods and services that results in the fall of purchasing power. By eroding the purchasing power of households, inflation makes essential goods less accessible and forces people to choose between necessities like food, energy, and housing. Even more, inflation increases the cost of energy and groceries, which brings up a prominent issue in households with people being unable to pay their growing bills.[1]

Argentina specifically has faced chronic inflation during its lifespan and although it has reduced recently, deep structural issues still exist, especially around currency instability and with regards to trust. Even the United States, a developed country, faces widespread inflation that continues to come back and affect the cost of energy, gas prices specifically. Inflation is not merely a cyclical economic issue in today's society, but moreover it is a structural barrier to

global economic stability that significantly impacts emerging markets like Argentina. Inflation also exposes vulnerabilities to advanced economies like the United States, specifically under geopolitical shocks such as energy disruptions during conflicts.

METHODS

This paper employs a comparative case-study methodology, drawing on publicly available macroeconomic data and published literature to analyze inflation as a structural global problem. The primary sources used include IMF World Economic Outlook Data Mapper for cross-national inflation rate comparisons,[2] academic literature analyzing Argentina's economy following Milei's reform program,[3] scholarly research on dollarization and economic stability,[4] Reuters reporting on IMF assessments of energy price impacts,[5] and The Guardian's coverage of geopolitical conflict and global oil prices.[6] Argentina and the United States were selected as contrasting case studies representing emerging and advanced economies respectively. Thematic analysis was applied across both cases to identify structural drivers of inflation, evaluate policy responses, and examine geopolitical transmission mechanisms. IMF cross-national data provided a quantitative foundation against which the qualitative case findings were contextualized.

RESULTS

What Drives Inflation and What Is Its Impact?

Inflation is caused by multiple different factors and there are many different types, some new ones being discovered even today. But, the most common type of inflation is demand-pull inflation that occurs when increased demand for goods and services outpaces the available supply, leading prices to increase. Some causes of demand-pull inflation include a strong confidence in consumers, increased government spending, and high foreign growth for companies. Cost-push inflation on the other hand, arises from increased production costs, such as wages or raw materials, which are then passed on to consumers in the form of extremely high prices. A few causes are increased prices for raw materials like oil, disruptions to the supply chain, and higher wages for many. Cost-push inflation is dangerous since it can lead to stagflation which is a combination of both high inflation and low economic growth.[1]

Additionally, monetary policy and fiscal policy are key drivers of inflation, although they do operate through different mechanisms, the combination of both has led to pronounced global inflation. Monetary Policy occurs when central banks implement policies that increase the money supply or lower interest rates, in turn increasing inflation. This increases liquidity in the economy and encourages borrowing and spending, which sharply drives prices upwards. By contrast, Monetary Policy can work to fight inflation as well when central banks raise interest rates in an attempt to slow down economic activity and reduce the amount of borrowing and spending, although this doesn't always work. Fiscal Policy on the other hand, drives inflation through government spending, deficit management, and taxation. When the government is the one to increase spending or cut down taxes, this puts money directly into the economy, in turn

increasing demand for housing and income. Many people argue that if the government runs large deficits without a proven plan to pay them back with future taxes, it will only cause even more inflation.

Inflation is relevant all over the globe and is a pressing global issue since it affects food availability, savings, and staggers economic growth. This in turn forces many into poverty. Inflation takes a considerable toll on the cost of living. The costs of any living essentials such as food, gas, and heating saw a great increase in price. Moreover, interest rates continue to increase, which lead less people to be able to buy housing and invest in property. Low income populations are often hit the hardest as they are forced to spend more money on necessities, leading to a greater gap between the rich and the poor. As is widely known, inflation and unemployment are related, but how exactly? By the Philips Curve Idea, there is an inverse relationship between inflation and unemployment: lower employment drives higher inflation. As unemployment falls and labor markets strengthen, wage competition will increase, in turn pushing up both production costs and prices, which illustrates the broader dynamic.[1]

With inflation comes many noteworthy challenges including economic slowdown, paycheck inequality, devaluation of currency, and struggling small businesses. Economic slowdown occurs when central banks raise interest rates in order to fight inflation, leading to recessions. Paycheck inequality is also common since the wealthy are not affected by inflation significantly, yet those in poverty are greatly affected. Devaluation of currency is when local currencies decrease in value due to inflation, this is exactly what happened in Argentina. With regards to the small business aspect, they begin to struggle due to higher production costs and lower consumer spending, together leading them towards major losses.[1] These mechanisms can be seen directly through the lens of Argentina, which demonstrates exactly what happens when demand-pull pressures, monetary failure, and institutional breakdowns all occur simultaneously.

Case Study 1: Argentina

In Argentina, inflation went over 140% in 2023. Necessities like bread and milk started to look unaffordable to many. The government attempted to control prices and implement monetary funds, but it was unaffected. The value of the local peso dropped so much that Argentina was forced to start using the United States dollar, known as peso devaluation. Argentina is a country with a long history of recurring inflation and it faced economic instability.[7]

Argentina's inflation crisis is not at all recent, it is rooted in decades of recurring instability, leading to Milei's reforms, in which he cut government spending and slowed down the creation of money. These reforms achieved short-term success but left a deep structural problem unresolved. Argentina has faced persistent high inflation, often reaching triple digits in recent years. Before Milei, inflation exceeded 200% annually, which severely damaged economic stability in the nation. During the late 20th century, Argentina saw a time of hyperinflation after its 2001 economic collapse, and to this day it faced chronic inflation. Argentina's economy was in such a deep depression that it began to lose value in its local peso relative to the U.S. dollar, which led to widespread lack of confidence in its national currency. Milei was able to

successfully reduce inflation in the short term, but structural problems remain and inflation is a systematic problem in Argentina. This historical instability is very closely related to structural weaknesses Argentina faces today such as the erosion of purchasing power and severe unemployment.[3]

Argentina's inflation stems from structural imbalances that remain even after efforts towards reformation. A few of the main causes of inflation are excessive money printing, chronic fiscal deficits, weak central bank credibility, and dollarization behavior. Historically, Argentina relied on and continues to rely on printing money in order to finance government spending. This expansion of money supply leads to a reduction in purchasing power, and even Milei's attempt to freeze money supply did not stop inflation. Next, Chronic Fiscal Deficits hurt the economy as government spending exceeds revenue and money borrowing or monetary financing become necessary for the country to survive. Milei introduced deficit reduction in an attempt to stabilize the economy, but it didn't reduce inflation completely.[3] With regards to Central Banks, the country has a great lack of trust in its monetary institutions. Even today, there are concerns about long-term credibility due to the banks being undermined by political interference. Finally, dollarization behavior led citizens and businesses to prefer holding the U.S. dollar over the local peso. The country later adopted the "dual currency" system, in which they use peso for transactions, but dollars for savings and in the long run. This alteration considerably weakened the demand for the peso and fuels instability in the struggling nation.[4] Together, all these factors influence each other as they cause further inflation, and even with reforms restoring the economy remains difficult; these structural issues directly affect the everyday life of Argentinians.

The effects of inflation in Argentina are widespread and reach beyond just political issues into social impact and increased poverty for citizens. First off, inflation has made basic goods like food and medicine increasingly unaffordable for many, leading to a loss of purchasing power. Wages continue to fall in order to keep up with these rising prices, hurting citizens even more. Inflation has also contributed to widespread poverty and a decline in living standards for Argentinians. The quality of life has significantly reduced and even with reform progress remains limited. Furthermore, due to this economic instability, investors are cautious and Argentina struggles to buy back this confidence from investors. Economic stabilization does not immediately translate to social improvement and there is a substantial bridge that still needs to be closed through new policies.[7]

Argentina is a key example of how inflation is self-driving and it grows when institutions and the government break down. Inflation persists due to the ideology many people have, the idea that inflation is expected to continue on. This leads to behaviors that only reinforce the already moving inflation, such as raising prices, converting pesos to dollars, and avoiding long-term commitments in fear of losing money. Argentinians lack confidence in their government policies, central bank stability, and even their national currency due to the fact that inflation follows them. Patel emphasizes how long-term success and the end of inflation for the nation depends on restoring credibility and increasing the people's confidence in their economy. All in all, the journey of inflation for Argentina tells the story of how controlling this growing

issue requires more than just changes to policy, but rather it demands rebuilding trust in economic institutions for the long run.[3] Argentina's crisis may appear to be a uniquely dysfunctional economy, but The United States proves otherwise as even the world's most powerful economy is not immune to the dangers of inflation.

Case Study 2: The United States

In the United States, inflation peaked at 9.1% in June 2022, making it the highest level in four decades. In recent years and following the sharp surge in 2021, The United States has faced inflation driven by a combination of pandemic-related and global supply disruptions. First off, COVID-19 was a major large-scale government stimulus that increased household income and led to substantial consumer spending after lockdowns. Yet, this ultimately created an excess demand in the country's economy and paved the path towards inflation. Similarly, supply chain disruptions spurred from shortages of key goods and the fact that supply could just not keep up with demand, leading prices to increase significantly. Due to this, there were shifts in the labor market as well with a tighter workforce for companies and a decrease in wage because of rising business costs. More importantly, energy costs were already rising post-pandemic and this set the foundation for even more inflation during further external shocks to come. Inflation showed itself due to a double shock—strong demand and a small supply in the industry, and this vulnerability became more apparent as geopolitical shocks emerged and introduced new pressures.[2]

Now, geopolitical conflicts and tensions that affect the energy market have become the main driver for inflation in the United States. Conflict in the Middle East right now has disrupted the global supply of oil. Oil prices proceed to surge and gas prices increase for everyone around the nation. This hike in energy prices affects not only transportation, but also food production and manufacturing. People are seeing higher gas prices than ever before, and this conflict in Iran is a key factor that is holding back The United States economy right now. The United States is part of the global energy market, and because of this, even domestic production is unable to fully protect the country from this global price increase. These continued increases in energy prices lead to higher inflation and slower economic growth in the foreseeable future. Energy shocks play the part of a multiplier as they raise costs across all parts of the nation, and since they are external, national policies find it difficult to control. Even the most advanced economies and developed nations, like The United States, are heavily affected by these geopolitical shocks.[5]

With regards to the role of Monetary Policy in all of this chaos, the Federal Reserve has responded to rising inflation through monetary policy, but there are still limitations in addressing these shocks. For example, interest rate hikes have had a rapid increase starting in 2022, which has reduced borrowing and slowed demand all together. This has somewhat helped bring inflation down from peak levels, but it brings up other issues as well, with Americans being unable to invest in property. Also, the Federal Reserve aims to stabilize inflation by around 2% by adjusting rates in order to stimulate economic activity. However, like many policies, Monetary Policy has its own limits. Monetary Policy in the United States is effective at

controlling demand-driven inflation, but it is not as effective when it comes to supply-side shocks, such as energy price hikes. Moreover, policy changes take time and considerable effort in order to deeply impact the economy for the better, which the United States does not have right now. In totality, the Federal Reserve is able to slow down inflation short-term, but it cannot directly control global supply disruptions or geopolitical events that are continuous driving factors of inflation today.[2]

In sum, The United States case clearly demonstrates that even developed countries with strong economies are vulnerable to inflation since demand shocks, supply and demand constraints, and geopolitical shocks can occur at any time and even occur simultaneously. Inflation begins to intensify when demand rises quickly, supply is constrained, and external shocks increase prices of important energy like gas. Monetary policy also takes time to properly reduce inflation and change does not occur immediately, especially during sudden external shocks. Broader implications exist as well, with inflation not only being internally driven, but rather it is shaped by the interconnectedness of our world today, and how this risks inflation hurting supply and demand as well as key energy markets. The United States experience illuminates how inflation is not limited to developed countries, it can affect anyone at any time, and is shaped by global issues that go beyond national control and affect the livelihoods of American citizens.[5] Understanding how inflation travels across national borders requires the examination of broader geopolitical forces that drive prices worldwide.

Global Dimension: Inflation and Geopolitics

Geopolitical tensions in energy-producing regions heavily disrupt supply and trigger widespread inflationary pressure around the world.[5] Today, countries' economies around the world are tied together and deeply linked through global trade networks. Disruptions in just one country affect the global supply, like oil for example, causing gas prices to increase for many around the globe. Major causes of these disruptions include conflict, trade restrictions, and political instability in any nation. The supply chain faces big disruptions during these times and it is difficult to quickly bounce back amidst this type of inflation. Inflation transmits throughout different nations quickly when delays and shortages cause increased production costs and then lead to higher consumer prices for everyone. Specifically, shipping delays can lead to limited access to raw materials, which then forces companies to charge more for products due to the inaccessibility of these precious items.[6]

Conflict-driven inflation is common in many countries and is especially prominent now during tensions in the Middle East. Foreign conflict leads to a reduced supply of key goods like energy, food, and materials in countries around the world. This increase in uncertainty directly influences market volatility and pushes investors away. Specifically in Europe, the region heavily depends on external energy sources to keep its economy running. When conflict disrupts this important supply, there is a rapid increase in both gas and electricity prices and this even begins to affect manufacturing and the cost of living. During tensions in the Middle East today, a major oil-producing region, other countries like the United States struggle with energy prices. Due to

unavailability of a high oil supply, Americans are faced with high gas prices that impact their day to day lives.[6]

Developing Countries are hit particularly hard during these times of conflict and amidst peak geopolitics. Energy and food prices are directly linked and as energy prices increase due to conflict in the Middle East, so do farming, transportation, and distribution costs. These price increases affect citizens' daily lives and emerging markets as well. Countries like Argentina, who are already facing inflation instability, are even more vulnerable to these external shocks. These geopolitical conflicts can worsen the already falling economies of many countries, making it even harder to bounce back. Together, inflation acts as a structural barrier that limits growth, increases inequality, and reduces purchasing power in both developed and developing nations. Taken together, geopolitical shocks unveil vulnerabilities and cracks in both advanced economies and emerging markets as inflation is globally transmitted and structurally embedded around the world.[5]

Comparative Analysis

IMF data shows inflation significantly varies across different countries, especially between emerging and advanced economies, The United States and Argentina are key examples of both. Emerging markets consistently experience higher and more volatile amounts of inflation than already emerging economies. For emerging economies, higher inflation rates and greater volatility over time have become the standard and these countries are often overlooked as they are seen to have gone beyond recovery. For advanced economies on the other hand, they face inflation as well, but it is often lower and more stable due to having stronger policy control. Inflation differences are systematic, not merely random.[2]

Going back to Argentina, its economy faces continuous inflation that is high and volatile to the market. Inflation often becomes persistent and self-reinforcing for developing countries, making it even harder to fully rid the economy of it. The United States on the other hand, aligning with IMF data, faces lower and more controlled and stabilized inflation. Even during shocks and geopolitics, inflation rises temporarily high, but it always stabilizes eventually. This is due to the strong credibility of U.S. institutions including market confidence and trust from the public.[2]

In order to fully understand how developing and developed economies compare when hit with inflation, the countries' economies need to be analyzed heavily. For example, by the IMF pattern, inflation spirals in some countries and not others not only due to exposure shocks, but from the ability for some countries to respond effectively. Trust also plays an important role in improving inflation. For stable economies, there are specific policies that go into place and are expectations by the government that have to be met to combat inflation. But for volatile economies, when governments aren't as reactive, they end up being unable to fully destabilize the storm. IMF data specifically shows how different inflation outcomes in separate nations are driven directly by governance and policy decisions. This gap in governance is exactly what makes inflation structural, rather than a merely cyclical, global problem. Ultimately, inflation is

shaped not only by global shocks, but rather by how institutions respond to them, and this remains the defining divide between why inflation hits harder in emerging economies versus advanced economies.[2] These discrepancies in governance do not merely produce different inflation outcomes, they instead affect whether entire populations have stable work, affordable living, and economic opportunities, which prevents progression towards inclusivity in the global workforce.

Employment and Inequality

The structural inflation patterns from Argentina and the United States carry the consequences for one of the most pressing issues in the world: unemployment. Inflation leads to a decrease in decent work by disrupting the creation of new jobs and it destabilizes economic growth. Inflation leads to less job opportunities, unfair wages, and economic instability for many. On a larger scale, inflation damages the progression towards a more equitable society. Inflation slows down economic development, not allowing for decent work. Also, investments, job hiring, and growth notably decline, causing investors to avoid investing in already struggling economies.[1]

Inflation undermines not only having a sustainable and stable economy, but also it weakens the progression towards productive employment. Inflation creates severe uncertainty in markets and in turn discourages long-term investment. It also reduces confidence in businesses and therefore delays expansion and innovation in our changing world. Emerging markets like Argentina that face recurring inflation limit the country's growth. Advanced economies also experience slowed growth during inflation spikes that are directly connected to global shocks. With regards to employment, business uncertainty leads towards reduced hiring and expansion. Rising costs also cause constant layoffs and freezes in hiring along with declining incomes. This ultimately leads to more insecurity in jobs and a reduced workforce in unstable economies, illuminating how inflation acts as a structural barrier to our global economy that prevents social progression.

Moreover, inflation erodes living standards and the economic well-being of nations, while also amplifying inequality that exists between nations. First, a decline in wages causes purchasing power to decrease for many and inflation also leads citizens to be unable to buy household essentials. Consumption is forced to be cut down and businesses begin to struggle during this time as well. On a social impact level, there is a disproportionate difference between how low-income vs wealthy populations live. This forces many people into poverty and increases the risk of poverty worldwide. Inflation shifts economic strain onto the most vulnerable of groups, but in reality it could affect any economy with equal severity, this undermines the "inclusive" aspect countries around the world strive to achieve. Furthermore, inflation widens the economic gap that exists between emerging and advanced economies, pushing even more towards the idea that inflation is a global structural issue that is not merely isolated to one region.[1]

This evidence suggests that inflation is not just cyclical, it is a roadblock that disrupts growth, employment, and living conditions simultaneously for those around the globe. Inflation prevents meaningful progression towards equity and it exposes the cracks that exist in both emerging markets and advanced economies. Addressing the impacts of inflation is essential for not only economic policy, but moreover in order to achieve sustainable and inclusive global progress around the world. Recognizing inflation as a structural barrier rather than a temporary disruption brings up the idea that it is no longer about reacting to inflation, but how to build economies that are resilient enough to prevent it.

DISCUSSION

The magnitude of inflation is so great that in order to solve this issue, multi-level solutions are needed, specifically independent central banks are critical to controlling inflation. In order to prevent political interference through monetary policy that could hinder solving inflation, independent central banks are required. They build credibility and stabilize the expectations that come with combatting inflation. For example, in countries like Argentina, who have an unstable inflation history, independent central banks can help reduce the risk of hyperinflation and lead to stability in the country's currency. Governments should also do their part in controlling spending and deficits in order to overcome inflation. Governments can limit excessive borrowing of money and prevent an unimaginable amount of money printing as well. Poor fiscal policy fuels inflation cycles and governments can help to stabilize currency and improve the confidence of investors instead by implementing policies.[3]

Moreover, central banks should have the ability to balance inflation control through growth. It is important to avoid overly aggressive hikes in interest rates to overcome inflation. Over-tightening leads to risks of recession, but under-tightening also prolongs inflation as well. Strengthening supply chains can also extensively reduce the vulnerability of countries to inflation. Diversifying production and reducing dependency on a single region leads countries to live more stable. A change to the way supply chains function will prevent cost-push inflation as a factor to the barriers in our global economy.[5]

On a global level, solutions include reducing the reliance on a single energy source and having coordinated policy responses in order to mitigate inflation shocks. Investing in renewables and multiple suppliers of oil, for example, can pave the path towards energy stability and reduce transmission of global inflation. This will also successfully prevent spikes in energy prices during conflicts since countries would not rely on one specific country for their key resources. Furthermore, cooperation between governments and other institutions like the IMF, World Bank, and coordinated central bank actions is the way to directly fight inflation. This will prevent spillover effects and further stabilize global markets during times of peak inflation. Ultimately, inflation requires structural, coordinated solutions and without sustained policy alignment and the work of countries around the globe, it will remain a long-term barrier that holds back our global economy.[6]

Inflation in our global economy is no longer a temporary phenomenon, it functions as a structural barrier that severely impacts economic stability in both emerging and advanced economies. Through Argentina's persistent inflation crisis, long-term weakness in structure like fiscal policy, currency instability, and declining trust in institutions, it is clear that inflation is deeply embedded in an economy. The United States' experience on the other hand, illustrates that even strong economies and institutions are not immune to inflation, external shocks like supply chain disruptions and geopolitical conflicts expose pre-existing structural vulnerabilities.

These cases highlight the reality: inflation is increasingly shaped by the global interconnectedness of today's world. Energy markets, international trade networks, and geopolitical tensions transfer inflation pressures across borders and make it difficult for a single country to fully control it. As a result, inflation continues to reinforce inequality, disrupts economic growth, and significantly limits progress in both developing and developed regions.

Now and in the future, inflation will remain a defining challenge in an increasingly volatile global economy. Addressing it requires more than short-term policy adjustments, it will demand coordinated international strategies, stronger institutional credibility, and many more resilient economic systems. In conclusion, without coordinated structural reform, inflation will not merely persist, it will widen the fault lines between nations, deepening inequality and foreclosing the possibility of sustainable global growth.

Works Cited

1. World Bank. Inflation: concepts, facts, and effects. https://thedocs.worldbank.org/en/doc/278221541081208802-0050022018/original/InflationChapter1.pdf, 2018.
2. International Monetary Fund. World economic outlook data mapper: Inflation rate. https://www.imf.org/external/datamapper/PCPIPCH@WEO/CIS, 2024.
3. R. Patel. An analysis of issues with the Argentine economy following Milei's inflation reform. https://sites.lsa.umich.edu/mje/2026/01/09/an-analysis-of-issues-with-the-argentine-economy-following-mileis-inflation-reform/, 2026.
4. J. A. T. Carvache, J. A. M. Correa, C. Sempertegui. Dollarized economies in Latin America. An inflationary analysis of pre, during and post pandemic. https://arxiv.org/abs/2501.12358, 2025.
5. A. Shalal, R. Campos. IMF says prolonged increase in energy prices could boost inflation, lower growth. https://www.reuters.com/world/imf-says-prolonged-increase-energy-prices-could-boost-inflation-lower-growth-2026-03-19/, 2026.
6. The Guardian. Iran war and the global economy: How rising oil prices are driving inflation. https://www.theguardian.com/news/ng-interactive/2026/mar/22/iran-war-global-economy-donald-trump-oil-prices-inflation, 2026.
7. Buenos Aires Times. Argentina posted higher-than-expected inflation ahead of Iran war. https://www.batimes.com.ar/news/economy/argentina-posted-higher-than-expected-inflation-ahead-of-iran-war.phtml, 2026.

Exchange Rate Fluctuations and Domestic Gold Prices in India (2015–2025): A Year-on-Year Analysis By Aditya Singh

Introduction

This paper investigates the extent of the relationship between the domestic price of gold and the US dollar-Indian rupee (USD/INR) exchange rate. This topic is of high importance, as in India it is viewed as more than an ordinary commodity, functioning instead as an economically significant asset that influences households, investors, financial institutions, and policymakers. Therefore, understanding the factors that affect domestic prices can provide insight into financial decisions and broader macroeconomic conditions.

Exchange rates are relevant to the domestic gold price because they transmit changes in international gold prices and global economic events to the Indian market. Fluctuations in the USD/INR exchange rate influence the price of imported commodities and internationally traded assets; therefore, examining the relationship between exchange rates and the domestic price of gold is important to understand how gold prices are determined.

Although previous research has suggested a relationship between exchange rates and domestic gold prices, relatively few studies attempt to isolate the effect of exchange-rate movements from other factors influencing domestic gold prices. This creates a gap for research that compares their annual change year-on-year while also considering movements in the international price. Accordingly, this paper asks the question: How do exchange rate fluctuations affect the domestic price of gold year on year?

Background

For centuries, gold has held immense significance in Indian society largely because of its economic, religious, and cultural importance. Along with this, gold has served as an important store of wealth for Indian households, which has made it more than just a luxury item and instead an asset class used to preserve purchasing power and provide financial security across income levels (World Gold Council, 2021). These cultural and investment functions have contributed to sustained household demand for gold in India. India ranks as the second-largest consumer of gold and the largest consumer of gold jewellery in the world (Jadhav, 2025). With annual consumption consistently ranging between 700 and 800 tonnes (MoneyControl News, 2026), gold is India’s second-largest imported commodity after crude oil (Business Today, 2026), and the demand for it holds immense economic importance. The scale of this demand gives gold wider macroeconomic importance, particularly because domestic demand is met largely through imports.

Gold is an internationally traded commodity priced in US dollars; therefore, exchange-rate fluctuations can influence the domestic price of gold in importing countries. Despite India being one of the world's largest consumers of gold, domestic production is negligible in comparison to demand. In FY 2025–26, domestic production stood at only 1.4 tonnes compared with 711 tonnes of imports, making the country heavily dependent on foreign

supply. As a result of this import dependence, changes in international gold prices and the USD/INR exchange rate are important determinants of the price paid by Indian consumers (World Gold Council, 2026). This relationship can be understood through the concept of exchange-rate pass-through, which describes the extent to which exchange-rate movements are transmitted into the domestic prices of imported goods (Carrière-Swallow et al., 2023). Ceteris paribus, a depreciation in the Indian rupee against the US dollar increases the rupee cost of imported gold and is therefore expected to increase domestic gold prices, while an appreciation in the rupee against the dollar is expected to reduce them.

However, the relationship between exchange rates and gold prices in practice is not nearly as straightforward as basic currency conversion. Exchange-rate movements are only one factor influencing domestic gold prices. Fluctuations in the international price of gold are also a major determinant because India relies heavily on imports to satisfy domestic demand (World Gold Council, 2021). International shocks, geopolitical tensions, and periods of global financial uncertainty can increase demand for gold and push up its international price (World Gold Council, 2022). As a result, domestic gold prices may rise because the rupee depreciates, because the international price of gold increases, or because both occur simultaneously. This makes it difficult to explain changes in domestic gold prices through exchange-rate movements alone.

Gold is widely regarded as a safe-haven asset, although its role must be defined carefully. Baur and Lucey (2010) distinguish between a hedge and a safe haven, defining a safe-haven asset as one that is uncorrelated or negatively correlated with another asset during periods of market stress. Previous research therefore suggests that gold may behave differently during periods of financial uncertainty. This distinction is important for the present study because heightened uncertainty may create an additional source of demand for gold beyond the exchange-rate mechanism. During the COVID-19 pandemic, for example, rising global uncertainty and increased investor demand may have influenced international gold prices at the same time as the rupee fluctuated against the US dollar. As a result, movements in domestic Indian gold prices during this period may be more difficult to explain through exchange-rate changes alone.

Previous studies generally identify international gold prices and exchange-rate movements as important determinants of domestic gold prices in India. However, much of the existing research focuses on establishing whether these variables are related. Comparatively less attention has been paid to whether the apparent influence of the exchange rate remains consistent during periods of heightened economic uncertainty, when changes in international gold prices and safe-haven demand may also become especially important. This paper therefore examines whether the relationship between USD/INR fluctuations and domestic gold prices remained consistent between 2015 and 2025.

Methodology

This study was quantitative in nature and used secondary data to investigate the

relationship between the USD/INR exchange rate and domestic gold prices in India. International gold prices were also included to help distinguish movements in gold prices that could not be explained by exchange-rate fluctuations alone.

The study analysed three variables: domestic price of gold (expressed per 10 grams in rupees), the USD/INR exchange rate, and the international price of gold (expressed per 10 grams in rupees). The London gold price was used as a proxy for the international gold price because it is widely recognised as the global benchmark for gold trading. The time period for this study was the financial year 2015 to 2025, and all figures for domestic and gold prices were from an RBI publication. RBI was chosen for this study because it has reliable long-term data and all annual figures are in line with the Indian financial years.

Annual averages were used for all variables because gold prices can fluctuate considerably over short periods. Using annual averages provided a more consistent basis for comparison across financial years. Percentage change in the annual average of domestic gold prices and international gold prices were calculated from the annual data to provide context for the analysis. Coefficients of determination (R^2) were calculated using both percentage and absolute values. Comparing these measures helped assess whether the apparent relationship between exchange rates and domestic gold prices reflected genuine year-to-year variation or simply their shared long-term upward trend. Time periods described in the analysis were chosen largely because they differed from the general trend, such as an exception in 21-22 or the pandemic in 20-21.

Analysis

The data suggests a generally positive relationship between the USD/INR rate and domestic gold prices over the long term. Over the nine years, the exchange rate increased from 65.47 to 84.58, representing a depreciation of over 29% in the value of the Indian rupee. During the same period, domestic gold prices have increased substantially from ₹26,534.26 (₹/10g) to ₹75,841.87 (₹/10g), reflecting approximately a 185.8% appreciation in value. The R^2 between the average yearly exchange rates and domestic gold prices was approximately 0.925, indicating a strong linear relationship in the level data and their overall increase in both exchange rates and domestic gold prices is consistent with economic theory discussed in the background; a weaker rupee means more rupees are required to purchase the same amount of gold sold in US dollars at the same international price, leading to higher domestic prices.

However, as both of these variables generally have an upward trend regardless of one or the other, this value for R^2 likely overstates their relationship. Instead, when taking the R^2 of the % annual change of these variables, it drops to 0.08446, suggesting that exchange-rate movements alone explain only a limited proportion of the year-to-year variation in domestic gold prices. While still a major component of domestic prices, this suggests that other factors have a sizable influence and must be considered as well.

The COVID-19 pandemic provides an important case study for examining whether the relationship between exchange rates and domestic gold prices changed during a period of

heightened economic uncertainty. In FY 2020-21, the US dollar appreciated 4.69% against the rupee; however, domestic gold prices surged over 31.6% in comparison to the previous financial year. This greater increase in domestic gold prices in relation to the exchange rate suggests that gold prices rise during times of market volatility, which is consistent with gold's role as a safe haven asset. In FY 2020-21, international gold prices rose very similarly to domestic prices, with an annual change of 30.57%, suggesting that global gold prices, due to the pandemic, were likely the main driver of the sharp rise in domestic prices.

FY 2021-22 was a notable exception to the overall trend between exchange rates and domestic gold prices. During this period, international gold prices remained virtually constant; however, domestic prices and international prices moved in opposite directions. The US dollar appreciated approximately 0.4%, while domestic gold prices fell over 1.5%. In this case, the effect of the exchange rate was too small to push domestic gold prices upward, especially if there are other factors pushing it in the opposite direction.

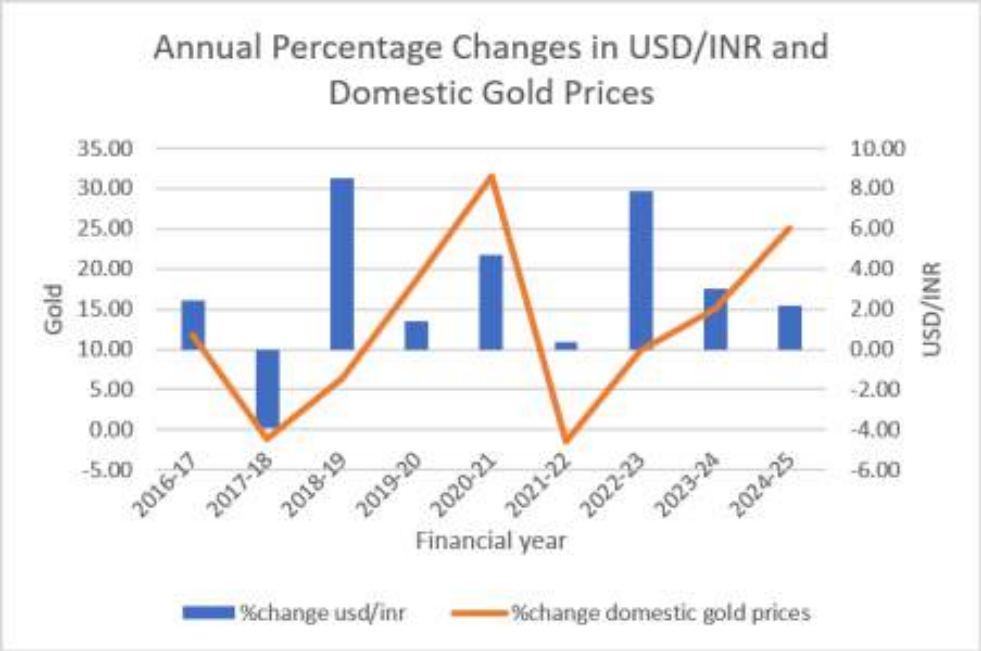
Overall, the data suggests a strong positive relationship over the long term; however, a less consistent one year-on-year. The pandemic during FY 20-21 also highlights gold's role as a safe haven asset, with gold sharply rising due to it being a period of financial uncertainty.

Table 1: Annual averages for the USD/INR exchange rate, domestic gold prices and international gold prices

Year	USD/INR	Annua% change	Avg domestic gold price (₹/10g)	Annual %Change	Avg international gold price (₹/10g)	Annual % change
2015-16	65.47	-	26534.26	-	24231.53	-
2016-17	67.07	2.45	29665.28	11.80	27116.42	11.91
2017-18	64.46	-3.90	29300.08	-1.20	26618.55	-1.84
2018-19	69.92	8.48	31193.41	6.50	28380.36	6.62
2019-20	70.90	1.39	37017.91	18.70	33346.53	17.50
2020-21	74.23	4.69	48723.22	31.60	43541.00	30.57
2021-22	74.50	0.38	47999.25	-1.50	43582.25	0.09
2022-23	80.36	7.87	52730.77	9.90	46605.74	6.94
2023-24	82.79	3.02	60623.95	15.00	52684.42	13.04
2024-25	84.58	2.16	75841.87	25.10	70315.41	33.47

Figure 1: USD/INR vs Domestic gold prices

Figure 2: Annual percentage changes in USD/INR and domestic gold prices



Limitations of The Study

The use of annual averages, while convenient, may have hidden important short-term fluctuations in gold prices, which could have suggested different behaviour. Additionally, the study does not take into account government policy, import duties or the tax implications of importing gold. Lastly, correlation does not necessarily imply causation, meaning that observed

relationships between exchange rates and domestic gold prices should be interpreted with caution.

Conclusion

This study examined the relationship between the USD/INR exchange rate and domestic gold prices in India from FY 15-16 to FY 24-25. It found that the USD/INR exchange rate and domestic gold prices generally moved in the same direction over the study period, supporting the expected positive relationship between the two variables. This produced a high R^2 value of 0.925, which is consistent with the expected long-term relationship.

However, when analysing the variables on a year-to-year basis as a percentage change, the R^2 value fell to 0.084 and suggested that exchange rates, while still significant, explained only a limited amount of the movement in domestic gold prices. This indicated that the strong positive relationship partly may be due to the overall upward trend of domestic gold prices and the USD/INR exchange rate.

The COVID-19 pandemic also highlighted another factor influencing ~~dom~~estic gold prices: gold's role as a safe haven. In FY 2020-2021, domestic gold prices rose far more sharply than the exchange rate, which suggested that gold's role as a safe haven can result in upward pressure on prices during periods of financial uncertainty.

This study had certain limitations; the data analysed here were secondary and analysed on a year-to-year basis. Overall, the findings suggest that exchange-rate fluctuations are an important determinant of domestic gold prices in India, although their influence varies depending on broader international market conditions and periods of economic uncertainty. Future research could examine quarterly or monthly data, as only looking at annual averages may hide aspects of their relationship that could be seen with shorter time periods.

Works Cited

World Gold Council. (2021, October 19). The drivers of Indian gold demand: India gold market series. https://www.gold.org/goldhub/research/drivers-indian-gold-demand-india-gold-market-series

Jadhav R. (2025, February 5). Gold demand in India to cool in 2025 as prices soar, WGC says.

Reuters.https://www.reuters.com/markets/commodities/gold-demand-india-cool-2025-prices soar-wgc-says-2025-02-05/

Business Today Desk. (2026, June 12). Gold is India's second-largest import after crude oil: Why that matters. Business Today. https://www.businesstoday.in/latest/economy/story/gold-is-indias-second-largest-import-after -crude-oil-why-that-matters-536608-2026-06-12

World Gold Council. (2026, March 17). Why gold in 2026? An anchor for Indian portfolios. https://www.gold.org/goldhub/research/why-gold-2026-anchor-indian-portfolios

Carrière-Swallow, Y., Firat, M., Furceri, D., & Jiménez, D. (2023). State-dependent exchange rate pass-through (IMF Working Paper No. 2023/086). International Monetary Fund. https://www.imf.org/-/media/files/publications/wp/2023/english/wpiea2023086-print-pdf .pdf

World Gold Council. (2021, December 9). Bullion trade: India gold market series. https://www.gold.org/goldhub/research/bullion-trade-india-gold-market-series

World Gold Council. (2022, April 28). Gold remains resilient amid heightened global uncertainty. https://www.gold.org/news-and-events/press-releases/gold-remains-resilient-amid-heightened -global-uncertainty

When the Proxy Loses Its Referent
Mileage, Battery Health, and Signal Reconstruction in the Used Electric Vehicle Market
By Jiwon Chung

Abstract

For decades, mileage has been the most useful piece of information in the used vehicle market. A higher odometer reading suggests greater mechanical wear, a shorter remaining life, and higher repair costs. Because mileage is easy to observe and verify, buyers and sellers have come to rely on it as a convenient proxy for a vehicle's condition and value. But the shift to battery electric vehicles (BEV) raises questions about how well this familiar measure still works. Mileage still indicates wear on tires, suspension, seats, and other physical parts. However, it tells us little about the condition of the battery, the most expensive component of an electric vehicle. Battery degradation depends on factors like calendar age, temperature, charging patterns, depth of discharge, driving conditions, and chemistry.

This paper examines whether mileage still carries pricing information in the used electric vehicle market. It also asks whether a reliable measure of battery health could provide information that mileage can no longer convey. Drawing on theories of adverse selection and signaling, the paper develops the concept of proxy decay, where a market signal weakens over time.

Two studies are proposed. The first uses hedonic regression to compare mileage and listing prices between battery-electric and internal-combustion vehicles. The second uses an experimental listing design to examine how consumers respond to listings with verified battery health information, seller claims, or no battery health information. Together, the studies investigate whether mileage has lost value in the used EV market and whether credible battery health information can fill the gap. The broader argument is that used EV pricing may not require a single new metric to replace mileage. Instead, the market needs credible methods to verify battery condition, especially as buyers face greater uncertainty that a component's mileage alone cannot address.

Introduction

Imagine two used electric vehicles, identical in model year, trim level, and mileage (around 40,000 miles). Their interiors are in similar condition, their accident histories are clean, and both are listed at the same price. One battery still holds 91 percent of its original capacity, while the other holds 74 percent – a gap a driver would notice on a cold-weather road trip. The listings don’t reveal which car is which, because the mileage number can’t show that difference. Both cars are priced within a few hundred dollars of each other. That gap, quietly built into an ordinary transaction happening somewhere right now, is close to the entire argument of this paper.

It helps to remember why mileage became the number that mattered in the first place. Mileage became the key metric because a buyer can’t pop the hood and know how many cycles

an engine has endured or how many shifts a transmission has handled. The odometer never precisely measured any of these things. What it offered instead was a decent summary of all of them at once: readable in a second, expensive to falsify without leaving a trail, and standardized enough that mileage on a Honda meant roughly the same as mileage on a Ford. Federal law exists because tampering with mileage was treated as seriously as forging a title.

A battery electric vehicle keeps part of that bundle intact. Mileage still says something honest about tire wear, suspension wear, and interior use, since none of that depended on the engine. What's changed is the most expensive component to replace: the lithium-ion traction pack. It degrades mainly through calendar aging, thermal exposure, and charging behavior, particularly high state-of-charge or fast charging, which research has pinpointed to the electrode and cell levels rather than to odometer readings (Schmuch et al.; Wassiliadis et al.). Two vehicles can leave a factory the same month, rack up an identical 40,000 miles over three years, and arrive on a used lot with meaningfully different battery health, because one was charged gently overnight to eighty percent while the other lived in a hot climate, fast charged most days, and sat at a full charge for long stretches. None of that shows up on an odometer.

This is not a claim that mileage has become worthless as a battery indicator. Driving further means more energy throughput and cycling, so some relationship persists. A recent study of over a thousand vehicles found that mileage explained almost none of the variation in real capacity, while charging behavior explained most of it (Park et al.). Manufacturer-reported state-of-health figures tracked true capacity poorly and inconsistently, and on one major platform, weren't even accessible to outside diagnostic tools. If the factory's own number cannot reliably answer the question, an odometer reading certainly cannot.

This isn't just an academic concern. The National Institute of Standards and Technology is studying how much used electric-vehicle buyers value battery health documentation, suggesting that the information gap is an active policy question (Gore et al.). The research question that follows is straightforward to state and harder to answer:

Does mileage still carry the pricing information it carried in the internal combustion market once the engine is gone, and can a verifiable battery health signal recover what mileage no longer bundles?

The paper makes three contributions. First, it tests empirically whether mileage behaves differently as a pricing factor for battery electric vehicles than for comparable gasoline vehicles, and whether the price variance in mileage, once explained, has vanished or merely shrunk. Second, it develops a theory of proxy decay and a reconstruction principle that outlines the conditions under which a market can recover the lost function of a decaying signal. Third, it introduces a Battery Health Premium Model that distinguishes between the value of good battery news and the value of not being left in the dark.

None of this suggests mileage has stopped mattering. It argues that a familiar number can keep driving prices through habit and familiarity even after what it measures has changed, and that whether the market catches up depends more on trust than on new information.

Literature Review

Adverse selection and the used vehicle market

The theoretical foundation for any information problem in a used goods market lies in Akerlof's analysis of the lemons market. Akerlof showed that when sellers know more about the quality of their goods than buyers, and buyers can't distinguish good units from bad ones before purchase, rational buyers discount every unit toward the average quality. This discount drives owners of genuinely good units out of the market, leaving it stocked with lemons. Genesove tested this logic in the wholesale used-car market, comparing prices for trade-ins sold by new-car dealers to those sold by used-car dealers. Emons and Sheldon took a different empirical route, using Swiss vehicle registration and inspection records to test for adverse selection on the seller side and quality uncertainty on the buyer side simultaneously, and again found support for the lemons framework running in both directions.

What all of this literature shares is a static picture of the asymmetry: sellers know more, buyers know less, and the market prices around that gap in one direction. None of it asks what happens when the object being sold changes in a way that quietly alters what a long-trusted signal actually measures. That question is the gap this paper seeks to address.

Signaling and the difference between a claim and a credible signal

Spence's remedy for adverse selection is that an informed party can reveal private information by taking a costly action a lower-quality type wouldn't imitate, converting something unobservable into something the market can price (Spence). Kirmani and Rao narrowed the question to what makes a given signal credible in a real transaction. They distinguished default signals, which are cheap and can be produced by any seller, from costly signals that only a genuinely high-quality seller would produce.

This distinction applies to battery documentation. A seller who claims the battery is in excellent condition has made a statement that costs nothing and can't be audited. A dashboard reading looks more concrete, but it can still be skewed by ambient temperature, driving style, or a manufacturer's internal calculations, none of which a buyer can verify. A standardized report from an independent diagnostic process sits in a different category, not because the numbers look different, but because of who generated it, how it was generated, and whether it can be reproduced by someone with no stake in the sale. Information and signals are not the same thing, and the used electric vehicle market has plenty of the former without enough of the latter.

Mileage as heuristic, not calculation

Buyers never process mileage with the precision economic theory assumes. Lacetera, Pope, and Sydnor found a pronounced left-digit bias in how buyers read odometers, causing sale prices to drop by $150 to $210 at each 10,000-mile threshold. The finding matters here for two reasons. It confirms that mileage functions in practice as a heuristic rather than a number buyers weigh with any real precision, which supports the idea that buyers will keep leaning on it out of

habit even after its connection to battery condition has weakened. It also shows that the tools needed to detect exactly this kind of proxy behavior, transaction-level regression around salient thresholds, already exist and have been used successfully in an adjacent setting.

Shrouded attributes and the cost buyers do not see

Gabaix and Laibson modeled a related phenomenon: firms can profitably leave a costly transaction attribute hidden from consumers who fail to anticipate it, so long as enough consumers are myopic about the hidden cost and no competitor unravels the practice by first unshrouding it. The framework applies to the total cost of ownership buried inside a used electric vehicle purchase. A buyer who prices a car mainly off mileage and sticker price, while underweighting the risk of a multi-thousand-dollar battery replacement, is behaving like the myopic consumer in the Gabaix and Laibson model. The absence of a mandatory disclosure standard for battery health functions is like an unshrouded add-on cost that no competitive pressure has yet forced into the open.

Battery degradation science

None of the market-design arguments holds up without a physical basis for the claim that battery degradation is weakly tied to distance driven. Schmuch et al. review the electrochemistry and cost structure of automotive lithium-ion cells, laying out the baseline picture: capacity fade results from lithium inventory loss, growth of the solid electrolyte interphase layer, and structural change in the cathode, driven by voltage, temperature, and time at a given state of charge, not by mechanical distance. Wassiliadis et al. confirmed this on a production vehicle, tracking range, efficiency, and pack aging on the Volkswagen ID.3 from component level to full vehicle, finding that usage and calendar factors, not mileage alone, drive the spread in real-world capacity retention.

Two recent sources sharpen the picture. Bilfinger et al. argue that no standardized vehicle-level procedure for measuring state of health exists, meaning the same reported number can mean different things across manufacturers. Park et al. tested this concern empirically, auditing onboard battery management system readings against independently measured capacity across 1,100 vehicles from five manufacturers and finding a correlation between reported and actual health ranging from zero to moderate, with one major platform not exposing a state-of-health figure to outside devices. This complicates the reconstruction question: even a manufacturer's official number is not automatically a verifiable signal. A figure can be printed on a window sticker and still fail the credibility test.

The gap this paper is built to fill

The literature explains most of the pieces but not their interactions. Akerlof explains hidden quality's effect on prices. Spence, Kirmani, and Rao explain what makes a replacement signal credible. Battery science explains why age, mileage, and true condition can diverge. None of it addresses the specific case of a used electric vehicle: a proxy that has stopped tracking what

matters most while keeping its authority. Buyers still ask for mileage first. Dealers still lead with it in advertising. Appraisal software still weighs it heavily. This is a decaying signal, not a missing one, and the existing frameworks were not built with that middle case in mind. We call this proxy decay, and the rest of the paper gives the idea empirical teeth.

Theory: Proxy Decay and the Reconstruction Principle

A proxy earns its place in a market by compressing information a buyer cannot obtain cheaply. Mileage became valuable in the used vehicle market because it condensed several forms of wear into a single number. Moving to a battery powertrain doesn't eliminate mileage's usefulness, but it narrows what it captures, leaving it accurate about some things and silent about the most costly ones.

The reconstruction principle proposed here states that when technological change weakens or removes the underlying function an established proxy used to track, market efficiency depends on whether that function gets reconstructed through a verifiable signal. More data alone doesn't solve the problem; it must arrive in a form buyers trust. Four propositions follow.

Proposition 1: When the lost function is reconstructed through a verifiable signal, price dispersion narrows. As buyers gain a real way to separate higher-quality vehicles from lower-quality ones, the premium demanded to compensate for hidden risk should shrink.

Proposition 2: Where the replacement signal is unverifiable, seller-generated data carries little separating power. A low-quality seller can make the same claim as a high-quality one at the same cost, so trust built on it is thin: one well-publicized false claim can collapse confidence in an entire class of unverified signals.

Proposition 3: A signal's strength is set by how well it satisfies a few verification conditions: third-party or auditable origin, a standardized metric, resistance to manipulation, portability, and a demonstrable connection to an outcome that matters. A signal that satisfies most of these sits near the honest end of the spectrum.

Proposition 4: While a verifiable replacement signal diffuses through the market, buyers don't suspend judgment. They substitute the nearest signal that remains legible and familiar, typically mileage or model year, even after it has lost much of its diagnostic value for the component that matters most. This is the paper's core: it predicts mileage will keep moving used electric vehicle prices at something close to its historical strength even as its link to actual battery condition weakens, because habit and legibility, not accuracy, keep an old proxy in circulation.

One adjacent idea from earlier drafts is deliberately left out. There was an argument that convenience-driven signals become identity markers buyers seek out independent of their informational content, a kind of belonging effect layered on top of the pricing effect. That may be true in some markets, and it might explain some of the stickiness behind Proposition 4, but it would dilute a theory that's supposed to be about verification, not psychology. It is worth its own paper, not a subsection of this one.

What Mileage Bundled and What Changed

Table 1 lays out the orthogonality claim directly, comparing how strongly each major category of vehicle wear tracks mileage on a gasoline vehicle against a battery electric vehicle.

Wear category	Relationship to mileage (Gasoline vehicles)	Relationship to mileage (Battery electric vehicles)
Engine and transmission wear	Strong. Combustion cycles and shift counts scale directly with distance covered.	Not applicable, or greatly reduced. Most BEVs run a single-speed reduction gear with far fewer wearing surfaces.
Tires and suspension	Strong. Wear tracks road contact and distance almost mechanically.	Strong. This mechanism does not depend on what sits under the hood.
Interior and upholstery	Moderate, tied to years of use and distance together.	Moderate, largely unchanged by powertrain.
Traction battery capacity	Not applicable.	Weak and inconsistent. Governed mainly by calendar age, thermal exposure, and charging behavior rather than distance traveled.
Remaining driving utility	Tracks mechanical condition fairly closely.	Tracks battery health and usable range more than it tracks the odometer.

Table 1: Relationship between vehicle wear categories and mileage, by powertrain

The pattern is not that mileage has become useless on an electric vehicle. Mileage still tracks tire wear, suspension wear, and interior condition, categories that don't depend on the engine. But it no longer tracks the traction battery, which dominates purchase decisions and replacement bills.

Candidate replacement signals, and why they are not interchangeable

Several signals have emerged to fill the gap left by mileage, but they're not equal under the verification conditions of Proposition 3. An original equipment manufacturer's state-of-health report should rank highly, but Bilfinger et al. and Park et al. found that, without a standardized measurement procedure and given demonstrated unreliability in real fleets, a manufacturer's number can't be assumed to be reliable. A third-party diagnostic report, generated from telematics data collected over time, ranks close behind or just ahead because it satisfies the third-party origin requirement and uses a standardized metric. A transferable battery warranty with a meaningful remaining term ranks below that, mainly because it shifts financial risk rather than conveying precise health information. Service records rank lower, as they show maintenance discipline but say nothing about the battery.

The dashboard range estimate deserves skepticism, as it looks like hard data but behaves like a claim. It's a number the car generates, visible during a test drive, and updated in real time.

But in Kirmani and Rao's terms, it's a weak signal because it's cheap to make look good: a seller can charge the car fully and drive gently in mild weather the day before a showing, and the figure is confounded by temperature, driving style, and tire pressure. A number that costs a seller nothing to make favorable isn't a costly signal, no matter how precise it looks on a screen. A seller's unverified claim about battery condition ranks last on this list, for the same reason a claim has always ranked last in any signaling hierarchy.

Software support horizon: a real but secondary channel

A second value channel has emerged: how long a manufacturer commits to supporting a vehicle's software, and whether premium features tied to an original owner's account transfer to a new owner's account. This affects resale value, but it's secondary here, since it governs convenience and feature access rather than the vehicle's core function.

The Battery Health Premium Model

The theory above extends a standard hedonic pricing model for used vehicles. Let the baseline price of a used electric vehicle be a function of the conventional hedonic covariates, age, mileage, condition, trim, and region, scaled upward or downward by a documentation term:

$$P_{used} = f(\text{age}, \text{mileage}, \text{condition}, \text{trim}, \text{region}) \times [1 + \beta \cdot SOH_{verified} + \gamma \cdot \text{warranty_remaining} + \delta \cdot \text{software_horizon}] - \lambda \cdot (\text{information absent})$$

This is not a claim that used vehicle prices follow a literal multiplicative formula. It is meant to separate three distinct sources of value that get blended together in ordinary listing data. The first is the conventional hedonic covariates. The second is a set of positive quality signals: β captures the premium associated with a verified state-of-health figure, γ the value of remaining transferable battery warranty, and δ the value of the remaining software support horizon. The third, and the term this paper is actually built around, is λ, the discount applied purely for the absence of documentation, independent of what that documentation would have shown had it existed.

That last distinction is the one existing research has failed to isolate. A hedonic study using listing data alone can estimate the price effect of a disclosed bad state-of-health reading. It cannot, from that same data, separate a buyer's rational discount for expected bad news from a buyer's discount for simply not knowing. A car with a verified health score of 78 percent delivers unwelcome news to a buyer and removes uncertainty. A car with no battery information leaves the buyer unsure, who then prices accordingly. Study 2 is designed to separate those two effects.

Hypotheses

H1, decay. The coefficient on mileage in a log-price regression is significantly smaller in magnitude for battery electric vehicles than for comparable gasoline vehicles, after controlling for age, trim, and region.

H2, residual dispersion. Conditional on standard hedonic covariates, unexplained price variance is higher for battery electric vehicles than for gasoline vehicles, suggesting that the information mileage used to supply has gone missing rather than merely shrinking in magnitude.

H3, reconstruction. Battery electric vehicle listings with independently verifiable state-of-health documentation command a price premium relative to comparable listings without it, with tighter price dispersion around that premium.

H4, verifiability. In a controlled experiment, a third-party state-of-health report increases willingness to pay more than an identical seller-asserted figure of the same value, and willingness to pay based on the third-party report is more resilient to a subsequent disconfirming disclosure.

Method and Success Metric

Study 1: hedonic regression on listing data

The target sample is 800 to 1,500 used vehicle listings, collected as matched pairs on the same nameplate sold in both gasoline and electric configurations: the Hyundai Kona against the Kona Electric, the Ford F-150 against the F-150 Lightning, the BMW 3 Series against the BMW i4. Same-nameplate pairs eliminate confounding variables by holding brand reputation, dealer network, and buyer demographics constant while varying only the powertrain. Exact matching will depend on data availability.

The primary specification is:

$$log(Price) = \beta_0 + \beta_1(Mileage) + \beta_2(BEV) + \beta_3(Mileage \times BEV) + Controls + \varepsilon$$

with controls for age, trim, region, seller type, and observable condition. A secondary specification adds a three-level documentation variable distinguishing listings with independently verifiable battery information, listings with a seller-generated claim only, and listings with no battery information.

The success metric is stated in advance: β_3, the mileage by BEV interaction, should be positive relative to the baseline negative mileage coefficient and significant at $p < .05$, indicating that additional mileage reduces price less for electric vehicles than for gasoline ones; and the verified-documentation coefficient should be positive and significant after controls. A null result on the interaction term would not be a failed study. If mileage retains roughly the same pricing authority for electric vehicles as for gasoline ones despite its weaker diagnostic link to the battery, that would be direct evidence for Proposition 4: that social familiarity can keep a signal's authority intact even as its informational efficiency erodes. The paper will report that outcome as a finding, rather than a disappointment. Study 1 is also the piece of this project most likely to

consume the bulk of the timeline, and listing collection should start immediately rather than waiting on Study 2 to finish.

Study 2: vignette experiment

The second study isolates λ directly. Participants, recruited through a screened panel to a target of roughly 150 to 200, are randomly assigned to view one of three otherwise identical used electric-vehicle listings. In the first condition, the listing carries a standardized battery health report generated by an independent, auditable diagnostic system. In the second, the listing states that the battery is in excellent condition and provides the same numerical figure, but presents it as the seller's own claim. In the third, the listing includes no battery information.

Participants report willingness to pay, purchase intention, perceived seller trustworthiness, and confidence in the vehicle's future reliability. They then receive a standardized disclosure that the battery's actual measured state of health is 78 percent, and the same measures are taken again. The design tests Propositions 1 and 2 directly: willingness to pay should be highest and most stable under the third-party report, lowest and most volatile under no documentation, and intermediate but fragile under the seller's own claim, with trust falling further after the disconfirming disclosure specifically in the seller-claim condition. Study 2 is feasible inside roughly three weeks and provides a direct estimate of the information-absence penalty represented by λ in the pricing model above.

Fallback question

If access to listing data collapses for any reason, a narrower version of this project remains viable using only the vignette design: do buyers discount used electric vehicles that lack battery health documentation, and by roughly how much? This version needs no scraping and can run entirely through the panel experiment, though at the cost of the paper's ability to speak to real market prices rather than stated preferences.

Implications

Managerial implications

For dealers and remarketers, battery documentation should not be treated as marketing collateral for premium listings only. Its value comes from reducing uncertainty: a dealership selling a genuinely healthy battery may get little credit for that quality if a buyer has no way to observe it; a credible diagnostic report converts a hidden advantage into an observable, priceable one. The same logic extends to transferable warranties and software support, but these should be treated as secondary value channels rather than substitutes for battery information. The theory also implies a hard limit on how far manufacturer-issued documentation alone can go: because an OEM both sets its own state of health algorithm and bears the cost of any warranty claim that algorithm triggers, its figures face a credibility ceiling that independent third-party certification does not, a limit the Park et al. audit makes concrete rather than theoretical.

Policy implications

The clearest policy lever is a standardized battery health disclosure requirement at the point of sale, following the precedent set by mandatory accident and title history disclosure. The European Union's Battery Regulation already moves in this direction, requiring a digital battery passport for batteries above a capacity threshold starting in February 2027 (European Parliament and Council). Any disclosure standard built on this paper's findings must include a privacy constraint from the outset: charging and location history that produces reliable independent health readings also discloses a previous owner's location and work habits. A workable standard must certify battery condition without exposing the movement history of its owner, a distinction that the current generation of third-party diagnostic products has not uniformly addressed.

Equity implications

The discount documented in this paper is regressive in a specific, measurable way. Buyers of used vehicles skew toward lower incomes than buyers of new ones, and the buyer least able to absorb a battery replacement bill is exactly the buyer most exposed to the risk an undocumented battery poses. Verification is, relative to that exposure, a genuinely cheap intervention: a third-party diagnostic report costs a small fraction of the replacement bill it protects against, which makes the policy case for mandating or subsidizing disclosure considerably stronger than the case for most consumer protection interventions of comparable cost.

Limitations and Future Research

Several limitations deserve acknowledgment. Listing prices are not transaction prices, and any systematic gap between the two could bias the Study 1 estimates. The underlying technology is moving quickly enough that a single cross-section risks describing a transitional market rather than a settled one; Geotab's own fleet degradation estimates moved from 1.8 percent average annual capacity loss in earlier reporting to 2.3 percent in more recent data (Geotab), which suggests the relationship this paper is trying to measure may itself be shifting year to year. Standardized state-of-health reporting is still taking shape as a regulatory and industry practice, and the verification conditions in Proposition 3 could change in meaning over the next several years. Finally, this project is scoped to a single country, and the used-vehicle market structure, warranty law, and disclosure norms vary sufficiently internationally that these findings should not be assumed to generalize without direct testing elsewhere.

The theory predicts the premium in H3 should shrink over time precisely as verified reporting becomes the norm rather than the exception. The idea of proxy decay, detached from electric vehicles, applies to other markets where a trusted proxy has lost its relevance: home appraisals after renovations alter square footage, credit scores when consumer behavior shifts, or academic credentials as skills outpace certification. Any of these would make a natural companion study to the one proposed here.

Conclusion

Markets rely on simplification because no buyer can inspect every component of an asset before purchase. Mileage became a durable simplification in the auto market, bundling wear into a single trusted number for a long time. Battery electric vehicles have not made that number meaningless. Mileage still tracks tires, suspension, and interior wear about as well as it ever did. However, it no longer accurately captures the largest source of uncertainty in used electric vehicles: battery condition, which is influenced by calendar time, thermal history, and charging behavior.

The asset worth defending here is not a battery health number by itself. It is the credibility of whatever system produced that number. A seller's claim conveys information, but it becomes a signal only if the buyer trusts the number reflects reality. That is why this problem does not resolve itself simply because more battery data becomes technically available. It resolves only if the data arrives through a process buyers have reason to trust, which is a narrower and harder condition than it might first appear. When a familiar proxy continues to command trust after the thing it once measured has moved on without it, the market does not fail loudly. It just keeps pricing uncertainty with a number that no longer means what everyone assumes it means, and that is a quieter, more durable failure than the kind Akerlof originally had in mind.

Works Cited

Akerlof, George A. “The Market for ‘Lemons’: Quality Uncertainty and the Market Mechanism.” Quarterly Journal of Economics, vol. 84, no. 3, 1970, pp. 488–500.

Bilfinger, P., et al. “Why We Need a Standardized State of Health Measurement Procedure for Electric Vehicle Battery Packs: A Proposal for Energy- and Capacity-Based Metrics.” npj Clean Energy, vol. 1, 2025, article 15.

California Air Resources Board. Advanced Clean Cars II: Section 1962.8, Warranty Requirements for Zero-Emission and Plug-In Hybrid Electric Passenger Vehicles, 2026 and Subsequent Model Years. 2022.

Emons, W., and G. Sheldon. “The Market for Used Cars: New Evidence of the Lemons Phenomenon.” Applied Economics, vol. 41, no. 22, 2009, pp. 2867–85.

European Parliament and Council. Regulation (EU) 2023/1542 Concerning Batteries and Waste Batteries. Official Journal of the European Union, L 191/1, 2023.

Gabaix, X., and D. Laibson. “Shrouded Attributes, Consumer Myopia, and Information Suppression in Competitive Markets.” Quarterly Journal of Economics, vol. 121, no. 2, 2006, pp. 505–40.

Genesove, D. “Adverse Selection in the Wholesale Used Car Market.” Journal of Political Economy, vol. 101, no. 4, 1993, pp. 644–65.

Geotab. “EV Battery Health: Key Findings from 22,700 Vehicle Data Analysis.” Geotab Research, 2026.

Gore, C., et al. Data Collection Instrument: Consumer Valuation of Battery Health Information in Used Electric Vehicle Markets. National Institute of Standards and Technology, 2025, doi.org/10.6028/NIST.DCI.008.

Kirmani, Amna, and Akshay R. Rao. “No Pain, No Gain: A Critical Review of the Literature on Signaling Unobservable Product Quality.” Journal of Marketing, vol. 64, no. 2, 2000, pp. 66–79.

Lacetera, N., et al. “Heuristic Thinking and Limited Attention in the Car Market.” American Economic Review, vol. 102, no. 5, 2012, pp. 2206–36.

Park, J., et al. “Battery Health Reporting Fails Independent Validation Across Manufacturers.” arXiv, 2026, arXiv:2603.21592.

Recurrent Auto. “Seattle Startup Recurrent Launches First-Ever Used Electric Vehicle Battery Reports.” Business Wire, Sept. 2020.

Roberson, L., et al. “Battery-Powered Bargains? Assessing Electric Vehicle Resale Value in the United States.” Environmental Research Letters, vol. 19, no. 5, 2024, article 054053.

Schmuch, R., et al. “Performance and Cost of Materials for Lithium-Based Rechargeable Automotive Batteries.” Nature Energy, vol. 3, 2018, pp. 267–78.

Spence, Michael. “Job Market Signaling.” Quarterly Journal of Economics, vol. 87, no. 3, 1973, pp. 355–74.

Wassiliadis, N., et al. “Quantifying the State of the Art of Electric Powertrains in Battery Electric Vehicles: Range, Efficiency, and Lifetime from Component to System Level of the Volkswagen ID.3.” eTransportation, 2022.

Comparative Empirical Descriptive Study of Unemployment Trends in the United States and Germany, 2005–2024 By Zayaan Ali Pallimalil

Abstract

Unemployment is a key macroeconomic variable that indicates the state of the labour market as well as the well-being of people reliant on employment for living standards and economic stability. This paper presents an empirical descriptive comparison of the evolution of unemployment in the United States and Germany between 2005 and 2024 using annual data on the unemployment rate and trend analysis. The analysis covers the general level of unemployment, the volatility of unemployment, the reactions to the global financial crisis in 2008 and to the shock caused by the COVID-19 pandemic, the recovery from these episodes, and episodes of convergence and divergence over time. It is found that unemployment in the United States has shown more dramatic ups and downs in the business cycle, especially in 2009 and 2020. Unemployment in Germany has shown a more gradual long-term fall and smaller upward movements during both the global financial crisis and the COVID-19 shock. Possible reasons for these contrasting experiences are discussed in terms of variations in labour market policies, labour market adjustment with regard to working hours, employment protection, and policy reactions. However, no causal links are attempted to be drawn. Rather, the utility of descriptive cross-country analysis in addition to causal and econometric analysis is shown herein.

1. Introduction

Unemployment represents a key macroeconomic variable that reflects employment and economic stability of individuals as well as the performance of the labour market. This paper provides a comparison of unemployment rates in the United States and Germany from 2005 to 2024 using a descriptive empirical approach. Annual data on unemployment rates are used to examine general levels of unemployment, changes in unemployment rates, responses to the global financial crisis of 2008 and the impact of the COVID-19 pandemic, recovery from these episodes, convergence and divergence, and trends in unemployment rates. The analysis finds that changes in the unemployment rate were higher in the United States in 2009 and 2020 and that long-term unemployment trends were smoother in Germany and lower during the global financial crisis. These findings suggest that such trends may be explained by differences in labour market institutions, levels of employment protection, flexibility to fluctuating hours of work, and the application of policy tools between the two countries. However, inferences about causality cannot be made from this analysis. Instead, the purpose of this analysis is to demonstrate the value of descriptive cross-country analysis for causal and econometric analysis.

Unemployment is an economic problem that directly and indirectly impacts all parties in the economy. For some people, joblessness can impact their lifestyle, job security, and income. Unemployment for a business or a government can signify underutilized production capacity, aggregate demand reduction, a loss of tax revenue, and strain on social welfare programs. But a very low rate of unemployment could be a sign of a shortage of labour and inflationary

conditions. For these reasons, the unemployment rate is an important indicator of the condition of the business cycle and of the labour market.

However, the response to unemployment is different across countries. Countries may experience the same major global crisis, but the effect on the labour market will differ based on employment protection, wage determination, working hours, social insurance, shock policies, and national institutional differences. Although both the United States and Germany are large, highly developed economies, they have a number of differences in their institutions and adjustment mechanisms in the labour market. The U.S. labour market is more reliant on external adjustment by altering employment, while Germany is more reliant on internal adjustment mechanisms that enable firms to adjust working hours without altering employment relationships (Herzog-Stein et al.). The two countries are particularly valuable for a comparative study of the unemployment reaction to similar economic shocks because of these institutional differences.

Analysis of the causes of unemployment within the empirical literature predominantly uses structural simulation-based regression methods. These methods are useful in attempting to capture the link between unemployment and other variables such as output, labour-market matching, job creation, inflation, government policy and employment. However, the results of causality can differ between countries due to the fact that the interactions between institutions and the adjustment process are not fully captured within a single model. A descriptive analysis allows for clarity in terms of timing, volatility, adjustment to a crisis, and recovery between countries. This leaves a gap for a descriptive comparison that uses harmonised unemployment-rate data to examine levels, annual changes, volatility, crisis responses, recovery patterns, and convergence or divergence across the two countries.

This paper investigates the evolution of unemployment in the United States and Germany between 2005 and 2024 in terms of unemployment level, unemployment volatility, crisis adaptation, and recovery. This paper does not make causal inferences, but rather attempts to note and explain trends in annual unemployment data in two institutional settings. The research question guiding this paper is: How did unemployment trends in the United States and Germany differ between 2005 and 2024 in terms of levels, volatility, crisis responses, recovery, and convergence or divergence? This paper proposes that the short-term volatility of the unemployment rate during economic shocks and the overall fluctuations of the unemployment rate will increase because of the relatively flexible labour market in the U.S. However, in Germany, institutional and firm behaviour seems to be more oriented towards keeping employees at the firm and, therefore, towards more stable unemployment trends. In the U.S., a large increase in unemployment will be followed by a relatively faster return of unemployment toward lower levels during an economic expansion.

2. Literature Review

The literature can be grouped around four themes that support the descriptive comparison in this paper: macroeconomic determinants of U.S. unemployment, German internal adjustment

and job retention, direct U.S.-Germany comparisons of labour-market flexibility, and the role of descriptive evidence alongside causal methods.

Using regression analysis, Green et al. explore the economic determinants of unemployment in the United States. Their findings show GDP, S&P 500, inflation, and unemployment insurance to be the key variables correlated with unemployment, with inflation being the most important in their analysis. The study consequently demonstrates the possible links between unemployment and various macroeconomic and financial parameters. Despite its status as an unrefereed preprint, Green et al. serve as an additional source of information for this study since it outlines the findings of their research.

Hutter et al. study the recent recovery in the German labour market using a structural macroeconometric approach grounded in the search and matching framework. They find that increased matching efficiency, increased job creation intensity, an increase in labour supply, and lower job separation propensity have contributed significantly to Germany's success. These findings are of particular interest to this study as they provide evidence that the drop in unemployment in Germany cannot be attributed exclusively to business cycle or GDP movements. Institutional and labour market adjustment also play a role.

Burda and Hunt analyze the Great Recession's exceptional performance of the German labour market. They find a number of contributing mechanisms such as pre-crisis employer hiring restraint, wage moderation, and a reduction in working hours. Working-time accounts enabled companies to free up some of the working-time balances of their employees instead of cutting their workforce. All of these mechanisms have contributed to labour hoarding, and have played a role in stemming the increase in unemployment during the recent recession of 2008-2009.

More recent comparative studies have directly compared differences between Germany and the United States. The paper by Herzog-Stein et al. looks at how the two countries reacted to the COVID-19 shock in the labour market by examining internal labour-market flexibility and external labour-market flexibility. The analysis is especially relevant for the present study, as it offers an institutional framework for the question of how the similar external shocks can lead to divergent unemployment dynamics between the two countries.

Taken together, these studies show that unemployment dynamics can be interpreted through multiple themes rather than through a single explanation: macroeconomic determinants, internal and external flexibility, job-retention institutions, and alternative causal mechanisms. From this evidence, two conclusions can be drawn. Firstly, adjustment to unemployment is conditional. The effect of a major global crisis on the labour market will differ between nations because different adjustment mechanisms are employed. Secondly, both causal and descriptive evidence can be employed. Models can be constructed to estimate the influence of certain variables, while comparisons between trend data can reveal national differences, the speed of change in unemployment rates, and similarities in recoveries. In this paper, the latter is employed in order to provide some context for the adjustment mechanisms identified in the literature.

3. Methodology

An empirical descriptive research design was employed in this study, and comparative research was conducted. Data were collected on the annual rate of unemployment in the United States compared to Germany from 2005 to 2024. The unemployment rate is defined as the number of people in the labour force who are able to work but not employed and seeking employment. This data was selected because it includes two large global events (the global financial crisis in 2008 and the COVID-19 global pandemic) and also includes periods of extended recoveries. Data is used on an annual basis, and all unemployment rate data follows the harmonised labour-force definitions in order to ensure comparability.

Data is measured on an annual basis, and unemployment rates were taken from the World Bank's World Development Indicators series using ILO modelled estimates, which ensures internationally comparable unemployment data (World Bank; see Appendix A). Annual unemployment rates for the United States and Germany are taken from this harmonised series. Monthly data from the U.S. Bureau of Labor Statistics (BLS) are provided merely for illustrative purposes during the early phases of the COVID-19 pandemic, when unemployment surged greatly (U.S. Bureau of Labor Statistics). Other institutional information is taken from academic literature and OECD Employment Outlook 2024 (Organisation for Economic Co-operation and Development).

There are 5 aspects considered in the analysis: the level of unemployment; the volatility of unemployment across years; the reaction to a recession; the pattern and speed of recovery from recession; and the divergence/convergence between the two countries. Annual data are summarized using measures and a graph. For each country, the following measures are calculated: mean, minimum, maximum, range, and the largest annual rise. The standard deviation of the annual changes in the unemployment rate is also calculated to differentiate long-term changes in the unemployment rate from short-term changes. A change in the unemployment rate in a year is the difference between the unemployment rate in the current year t and that in the previous year $t - 1$. The higher the standard deviation, the more the unemployment rates fluctuate from year to year.

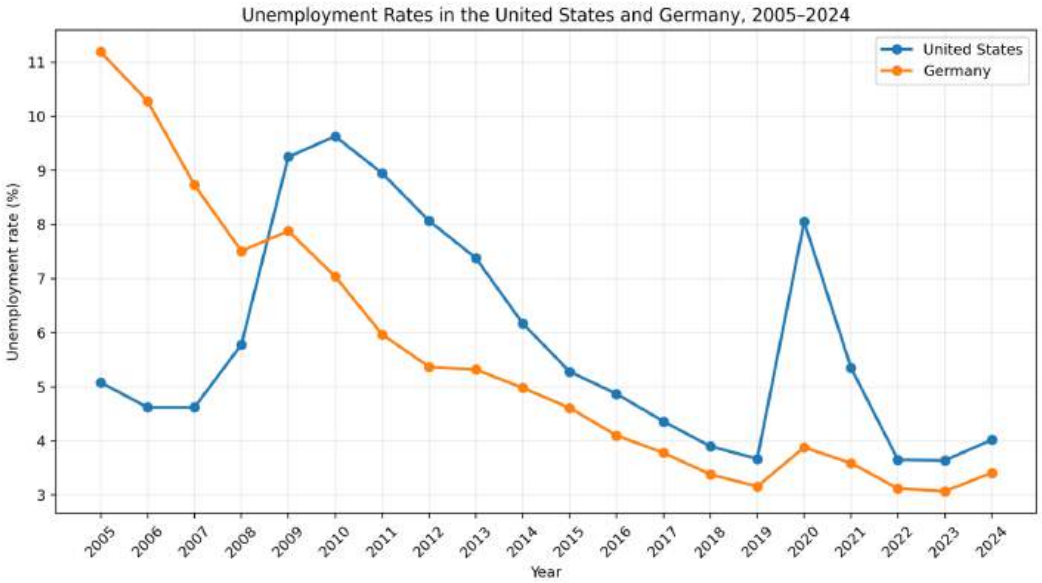
Limitations in the approach include the inability to illustrate cause and effect in relation to a particular reform or institution and unemployment. The approach can illustrate the timing and type of relationship but not cause and effect. Annual data masks any short-term movements that may occur in the month-by-month data (for example, the massive increase in the U.S. unemployment rate in early 2020). Unemployment is also unable to capture all labour market indicators such as labour force participation, hours of work, quality of employment, and underemployment. This is taken into account when interpreting the results.

4. Results

4.1 General unemployment levels

At the beginning of the study period, Germany had a higher level of unemployment compared to the United States. The unemployment rate in Germany was 11.19% in 2005 while in the United States the unemployment rate was 5.08%. Unemployment in Germany has steadily decreased since 2008 (7.51%) and dropped further in 2019 (3.16%). But the United States had a different trend. The unemployment rate remained relatively low in the United States until the financial crisis in 2008, then rose sharply during the crisis before declining significantly over the next 10 years. Germany's average unemployment rate for the period 2005-2024 was 5.52 percent, while the average unemployment rate for the United States was 5.82 percent. These values do not differ very greatly but show very different trends. Germany's initial unemployment rate was higher than the U.S. and Germany's transition to lower unemployment rates occurred over a longer duration of time. The initial unemployment rate in the United States was lower, but there were more cyclical fluctuations in unemployment.

Figure 1. Annual unemployment rates in the United States and Germany, 2005-2024.



Source: World Bank World Development Indicators / ILO modelled estimates; author's calculations.

4.2 Volatility

The overall range of unemployment rates was 5.99 percentage points in the United States, with a minimum of 3.64% in 2023 and a maximum of 9.63% in 2010. The overall range in Germany was 8.12 percentage points, from 11.19% in 2005 to 3.07% in 2023. This is, however, largely due to the fact that Germany's unemployment rate has been in a steady downward trend for the entire period of the study and does not necessarily imply a higher short-term volatility.

The standard deviation of all the annual unemployment rate changes was computed to get a more accurate indicator of the year-to-year volatility. The standard deviation for the U.S. was 1.63 percentage points and for Germany was 0.56 percentage points. This suggests that the

unemployment rate in the U.S. experienced significantly more fluctuation from year to year during this period (2005-2024).

This disparity is also apparent in the largest annual increases in unemployment. The unemployment rate rose by 4.38 percentage points from 3.67% in 2019 to 8.05% in 2020 in the United States, but it only increased by 0.72 percentage points from 3.16% in 2019 to 3.88% in 2020 in Germany. Thus, while the overall range of unemployment rates was larger in Germany, the standard deviation of the annual changes and the size of the largest annual change suggest that there was greater short-term volatility in unemployment in the United States.

The higher volatility found in this study is solely a measure of how volatility has changed over a number of years in the overall unemployment rate and should not be understood as a measure of volatility for all components of the German labour market. Using worker-flow data, Jung and Kuhn conclude that unemployment inflows in Germany are twice as volatile as in the United States, although the average worker flows are much lower in Germany. The separation of the volatility of the unemployment rate from worker flows demonstrates the distinctiveness of the problem of volatility in the unemployment rate.

4.3 Responses to recessions

The global financial recession saw the U.S. unemployment rate rise from 4.62% (2007) to 5.78% (2008) to 9.25% in 2009, 9.63% in 2010, and has subsequently been decreasing. Germany saw unemployment fall from 8.73% (2007) to 7.51% (2008) and then rise slightly to 7.88% in 2009. Hence, the U.S. shows a significantly higher response to the recession in terms of unemployment.

This is more evident during the COVID-19 pandemic. In the U.S., the annual unemployment rate rose from 3.67% (2019) to 8.05% (2020). The annual rate, however, can obscure a brief but intense period of unemployment. The monthly data released by the BLS show an intense spike in U.S. unemployment, reaching a high of 14.8% in April 2020 (U.S. Bureau of Labor Statistics). The annual rate increased slightly in Germany from 3.16% to 3.88%, but failed to reach 4%. Both countries were affected, but significantly more in the U.S.

4.4 Recovery patterns

Since the global financial crisis, unemployment in the U.S. has been steadily improving over time. In 2010, unemployment stood at 9.63% and by 2019 had reduced to 3.67%. The recovery has been long, but the rate of recovery has fallen to a level below pre-crisis rates. Unemployment in Germany has steadily reduced from 7.88% in 2009 to 3.16% in 2019. The recovery in Germany appears smoother because the increase during the initial part of the crisis was small and the fall continued over time.

In the U.S., unemployment reduced from 8.05% in 2020 to 5.35% in 2021, then to 3.65% in 2022. In Germany, unemployment reduced from 3.88% in 2020 to 3.59% in 2021 and then to 3.12% in 2022. The initial impact and the subsequent decline in the level of unemployment were quite high in the United States. Germany had a much smaller rise in unemployment, and thus

there was not as much room for a comparable decline in unemployment. The following post-2020 movements should, therefore, be read in the context of the varying sizes of the initial shock, and not as proof of quicker or slower recovery.

4.5 Convergence and divergence

There were a number of episodes of convergence and divergence in the unemployment rates. In Germany, unemployment was relatively high at the start but has come down markedly since 2005 as the financial crisis developed. In the U.S., unemployment also ended the period below its 2005 level of unemployment, but followed a much more cyclical path, with sharp rises during the two significant crises.

The rates were relatively comparable in 2008. A divergence occurred between 2009-10 due to a greater rise in U.S. unemployment. Both rates then came down and converged during the extended recovery that occurred throughout the 2010s. Another divergence occurred in 2020, and rapid convergence occurred in 2022. Across all shocks, there seems to be a trend of divergence during shocks and convergence during recovery.

This supports this paper's argument that there is greater sensitivity to the business cycle in the U.S. labour market and that Germany shows more stability during downturns in terms of unemployment levels.

5. Discussion

5.1 Labour market flexibility and U.S. volatility

The high unemployment spikes in the U.S. in 2009 and 2020 are consistent with a labour market that allows firms to make significant layoffs in response to decreased demand. In a similar vein, Herzog-Stein et al. describe the contrasting labour-market adjustments of the United States compared to Germany during the COVID-19 crisis as external versus internal flexibility of the labour market, with the United States relying more strongly on external adjustment through employment changes. Such flexibility could cause higher short-term unemployment because some individuals in that unemployment would be laid off relatively quickly in the event of a shock. Flexibility could also allow for a relatively quick recovery of employment during an upswing in demand. Both the size of the shocks and the magnitude of the unemployment decreases in the U.S. are consistent with this interpretation. The U.S. experienced a larger shock and a larger decrease in unemployment than would occur based on the size of the shock. Flexibility is not necessarily inherently good or bad. Such a rapid switching could allow for easier reallocation and lower labour costs for firms but could also transmit the effects of a shock to workers via layoffs. Welfare would then depend on the magnitude and duration of unemployment insurance, speed of employment generation, and the ability of those laid off to find new employment.

The evidence presented here is, however, of a descriptive nature, which cannot demonstrate the link between labour-market flexibility and the observed unemployment dynamics. The results are rather in accord with this institutional interpretation.

However, institutional differences may not fully explain the contrasting unemployment responses during the Great Recession. Hoffmann and Lemieux identify employment fluctuations in the U.S. construction sector associated with the housing boom and bust as an important factor behind the different labour-market experiences of the United States and Germany. They also find that international differences in labour-market performance were broadly consistent with the conventional relationship between output growth and employment. Therefore, the divergence observed in 2008-2009 should not be attributed exclusively to differences in labour-market institutions.

5.2 Employment security and German institutions

Great Recession: Germany's modest response to unemployment during the Great Recession is in line with the evidence of labour hoarding and internal working-time adjustment. Burda and Hunt highlight the importance of working-time reductions, working-time accounts, wage moderation, and pre-crisis employer behaviour for employment performance in Germany in 2008-2009. This meant that firms could tweak labour input to some extent by reducing hours, rather than just by firing workers.

COVID-19: The policy mechanism for the COVID-19 crisis was a little different. In Germany, Kurzarbeit was extensively used to maintain the employment relationship while reducing working hours. Without Kurzarbeit, the unemployment rate in Germany would have risen by an additional around three percentage points at the bottom of the recession, estimate Aiyar and Dao. This evidence shows that the employment retention policies played a role in Germany's comparatively small unemployment response to the pandemic. Kurzarbeit should not therefore be interpreted so much as a precondition that prevailed in Germany during both crises, but rather as an explanation related to the COVID-19 episode.

The German unemployment rate had been falling over time as well, and was falling at the same time as the Hartz labour-market reforms. It is estimated that lower separation rates account for around 76 % of the decrease in unemployment after the reforms (Hartung, Jung, and Kuhn). The changes they analyse are associated with the reform of long-term unemployment benefits, offering one institutional account of the longer-term improvement in the German labour market.

But the descriptive method applied in the present study cannot separate the effect on unemployment of the Hartz reforms from other concurrent developments in the German labour market.

5.3 Interpretation of two crises

Both crises show major global shocks but a varying labour market response to those external shocks. Following the financial crisis, there was almost a doubling of unemployment in the United States and only a small increase in Germany. The rise in unemployment was

significantly higher in the United States than in Germany in both the global financial crisis and during the COVID-19 pandemic. The repeated divergence indicates that a comparable shock may lead to significantly different labour-market adjustments in the two countries. This is similar to the contrasting labour-market experiences during COVID-19 between Germany and the United States (as explored by Herzog-Stein et al.), which is based on the concept of internal and external labour-market flexibility. Regarding recovery, there is some difficulty in making a judgement between stability and flexibility. Germany showed more maintenance of employment during the two shocks whilst in the U.S. the level of unemployment was reduced quite quickly once economic activity began to resume. There is a trade-off between short-term employment stability and labour market turnover. A descriptive comparison could show this trade-off but would not indicate which system would provide the greatest level of welfare.

5.4 Contribution of descriptive analysis

The primary contribution of this paper is not to obtain an estimate of causal effect from a specific policy. Rather, this paper gives a reasonable comparison of the unemployment dynamics in two institutional frameworks. Using regression analysis can help focus on the relations between the policies but may fail to capture the interplay between the labour time policies, dismissal procedures, social insurance, firm responses, and macro-economic shocks. Descriptive analysis can, however, show timing and patterns consistent with adjustment mechanisms identified in the literature and thus help causal analysis.

Furthermore, the analysis has revealed that using the mean unemployment rate is insufficient. Both countries followed a similar overall trend, but with very different trends within that. In Germany, the structural unemployment rate was reduced significantly and became quite low. In the U.S., it was mostly below, but with large cyclical movements. Combining level, volatility, and recovery provides a more comprehensive picture of the labour market condition.

6. Conclusion

This paper examines the difference in unemployment levels between the U.S. and Germany from 2005 to 2024. Germany shows a higher level of unemployment at the beginning of the analysis, but there is a consistent long-term decline. Unemployment in the U.S. is generally lower before periods of crisis, but there is a more visible rise in cyclical unemployment during the global financial crisis and during the COVID-19 pandemic.

The descriptive evidence confirms the main results, namely that short-term unemployment reactions to the two major economic shocks studied were stronger in the U.S. than in Germany, and that aggregate unemployment in Germany rose less during the shocks than in the U.S. and fell smoothly over the long run. The patterns are in line with the literature on the various ways of labour-market adjustment of the two countries, but other and/or alternative explanations are possible in terms of sectoral composition, macroeconomic context and policy interventions. The descriptive design used here cannot identify their respective causal effects.

This study does not explore causation but does show that two highly developed economies can show significantly different trends in unemployment while exposed to the same global shocks. Qualitative cross-country evidence can complement the econometric analysis by highlighting differences, institutional variations, and time periods that merit further exploration.

This analysis has some key limitations, including smoothing of monthly fluctuations using annual data, failure to account for shifts in labour force participation or hours of work when using unemployment rates, and only examining two countries. Further analysis could include the use of harmonized quarterly or monthly data, participation and employment rates, and an explicit examination of the link between unemployment volatility and labour market institutions.

Works Cited

Aiyar, Shekhar, and Mai Chi Dao. “The Effectiveness of Job-Retention Schemes: COVID-19 Evidence from the German States.” International Monetary Fund, IMF Working Paper No. 2021/242, 2021, https://doi.org/10.5089/9781513596174.001.

Burda, Michael C., and Jennifer Hunt. “What Explains the German Labor Market Miracle in the Great Recession?” Brookings Papers on Economic Activity, vol. 2011, no. 1, 2011, pp. 273–319. Project MUSE, https://doi.org/10.1353/eca.2011.0004.

Green, A., et al. “Identifying Economic Factors Affecting Unemployment Rates in the United States.” arXiv, 2024, https://arxiv.org/abs/2411.02374.

Hartung, Benjamin, Philip Jung, and Moritz Kuhn. “What Hides behind the German Labor Market Miracle? Unemployment Insurance Reforms and Labor Market Dynamics.” IZA Discussion Paper, no. 12001, IZA Institute of Labor Economics, 2018.

Herzog-Stein, Alexander, et al. “Germany and the United States in Coronavirus Distress: Internal versus External Labour Market Flexibility.” Journal for Labour Market Research, vol. 56, article 11, 2022, https://doi.org/10.1186/s12651-022-00316-5.

Hoffmann, Florian, and Thomas Lemieux. “Unemployment in the Great Recession: A Comparison of Germany, Canada, and the United States.” Journal of Labor Economics, vol. 34, no. S1, 2016, pp. S95–S139, https://doi.org/10.1086/682424.

Hutter, Christian, et al. “Which Factors Were behind Germany’s Labour Market Upswing? A Data-Driven Approach.” Oxford Bulletin of Economics and Statistics, vol. 84, no. 5, 2022, pp. 1052–1076, https://doi.org/10.1111/obes.12508.

Jung, Philip, and Moritz Kuhn. “Labour Market Institutions and Worker Flows: Comparing Germany and the US.” The Economic Journal, vol. 124, no. 581, 2014, pp. 1317–1342, https://doi.org/10.1111/ecoj.12118.

Organisation for Economic Co-operation and Development. OECD Employment Outlook 2024: Germany Country Note. OECD, 9 July 2024, https://www.oecd.org/en/publications/2024/06/oecd-employment-outlook-2024-country-notes_6910072b/germany_08ee101a.html.

U.S. Bureau of Labor Statistics. “Civilian Unemployment Rate.” U.S. Bureau of Labor Statistics, https://www.bls.gov/charts/employment-situation/civilian-unemployment-rate.htm. Accessed 10 Sept. 2026.

World Bank. “Unemployment, Total (% of Total Labor Force) (Modeled ILO Estimate).” World Development Indicators, World Bank, https://data.worldbank.org/indicator/SL.UEM.TOTL.ZS. Accessed 10 Sept. 2026.

How have Fintech Platforms Changed the Accessibility and Behaviour of Retail Investors in Stock Markets? By Parv Anand

Introduction

The financial sector is evolving significantly, fueled by the intersection of technology and behavioral sensibilities. The emergence of financial technology, or fintech, has transformed the investing experience and the investing environment by creating unprecedented accessibility to financial markets. The primary beneficiaries of this are retail investors—individual asset owners making their own capital allocation decisions without institutional intermediaries. Financial intermediaries, such as mobile trading apps, robo-advisors, and peer-to-peer investing, have lowered investment entry points, allowing countless new retail investors to interact with the stock market more easily and less expensively than ever before.

Meanwhile, the historical perception of investing as a completely rational and analytical exercise is becoming amalgamated with a more recent appreciation for the investor's experience—one that relates to the interplay of emotion, psychology, and behavior. In fact, the COVID-19 pandemic, while leading to widespread economic instability and uncertainty, ultimately stimulated a wave of new retail brokerage accounts and the introduction of new investment opportunities which fostered "meme stocks" and emotionally driven trading patterns. This temporal junction suggested an intrinsic relationship between human behavior and financial decision-making, and it inspired discourse and thought around behavioral finance as a useful lens for understanding investors' capital allocations and systems of beliefs, as well as the potential integration of behavioral finance into fintech platforms.

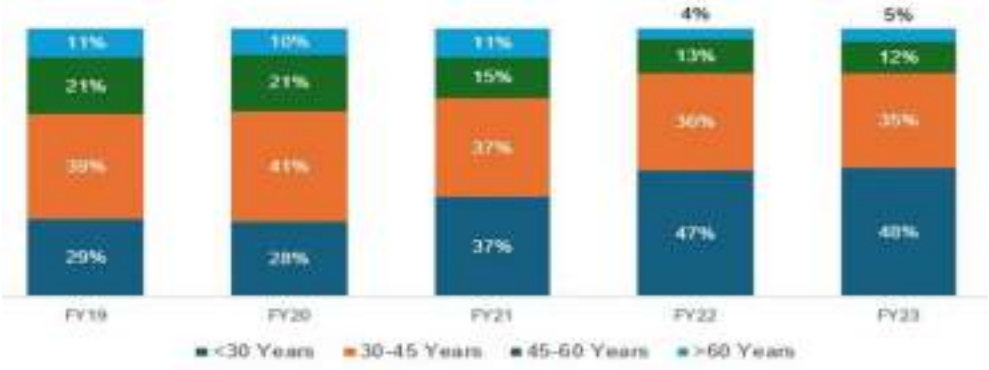
With fintech, accessibility was increased, but so was the way people behaved while investing money. According to the Financial Planning Association, fintech apps and their design can significantly influence the way people behave. This is called behavioral finance. Some examples of this are the notifications warning or highlighting opportunities or the gamified elements like streaks further influencing people to invest more money as well as the social aspect of these platforms that allows others to see what people are doing and get influenced by popular trends.

These dual trends--the rise of fintech innovation and greater awareness of behavioral finance--are combining to create an emergent, changing, and future role of the retail investor. Fintech platforms are not merely providing access to the markets; they are crafting new paradigms of how people make decisions, measure risk, and develop financial habits. Yet this transformation prompts important questions: In what ways has fintech changed access to stock markets for retail investors? And, in what concrete ways has investor behavior changed in response to this transformation?

This paper answers these questions by detailing how fintech platforms are democratizing access to stock markets and influencing the psychology, preferences, and habits of retail investors. In exploring these questions it identifies and explains the emerging necessity for trust, financial literacy, and emotional intelligence to be part of the next global

generation of fintech design and regulation.

Distribution of investors in the capital market across different age groups



Source: Ibef

Investors under 30 have not only grown substantially since FY19, increasing from 29% of all investors to 48% in FY23, but have also shown a willingness to diversify from traditional assets through the stock market and mutual funds. The investor base has increased significantly, while the aforementioned category of investor symbolises a strong basis for future growth in our market. This is the result of the fintech surge during the covid 19 pandemic which made fintech the safest way to continue financial services due to its ability to be used anytime and anywhere. This also brought younger audiences and over time, as shown in the data, these people dominate the current market.

Literature Review

The rise of fintech and its interaction with retail investing has gained significant traction over the last 10 years. In the past, investing was the domain of individuals who typically had a large amount of money to invest, enough financial acumen to consider themselves intelligent, and access to professional institutional-grade tools in order to invest wisely. This paradigm has changed significantly, thanks largely to the advent of fintech platforms.

The World Economic Forum states that there are more retail investors today than ever before. The rise of low-cost platforms and mobile applications with simplified user interfaces allowed this phenomenon to happen. While there are positive developments, important barriers still exist. That is, financial literacy is still limited, and many people still do not have personalized guidance or trust.

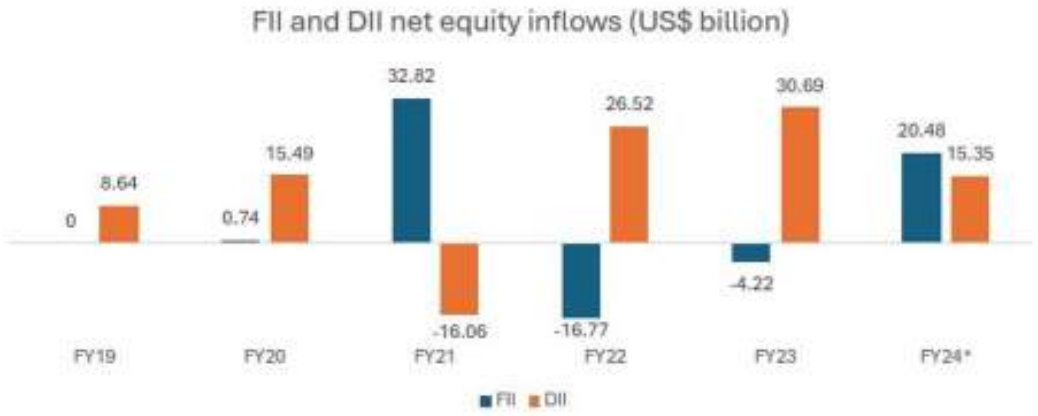
In terms of behavioral influence, yes, fintech is not just an access tool. The report from the Financial Planning Association illustrates that fintechs are functioning in the intersection of finance, and emotion. The pandemic jump-started this trend as millions opened brokerage accounts in times of volatility. Important emotional triggers (FOMO, social-media driven trends such as meme stocks) illustrate the prevalence of psychology in financial behavior. The

ubiquity of behavioral biases like confirmation bias, loss aversion, and anchoring biases must also be factors that fintechs must consider in the design of fintech platforms.

Furthermore, McKinsey's overview of the fintech space illustrates a two-phase evolution. The first phase (2015-2021) was characterized by hypergrowth based on speed, access, and affordability. The second phase (commencing post-2022) is focused on returns, cost restraints, and value creation in the long term. Speed and action is no longer sufficient, and fintechs must best develop viable, financial templates, with ongoing long term embedment models and user centric considerations, especially in emerging or underserved markets (e.g., Africa, Southeast Asia).

Overall, these studies show that fintech marketing not only has democratized access in the industry, but it's also influencing investment behaviors in a substantial way. However, there is still a gap in research related to understanding the longer-term meanings and psychological effects of fintech-enabled investing. Furthermore, there's little evidence of how fintech platforms can reconcile growth with ethical use and emotional support.

Growth in retail investors and domestic funds in India



Fintech and Market Accessibility

One of fintech's most impactful areas has been access to stock markets. Investment traditionally required intermediaries, access to large amounts of capital, and expertise. Today, anyone with a smartphone can start investing from their couch and with little capital by removing intermediaries and using simplified design features.

Not long ago apps like Robinhood and Zerodha eliminated brokerage and trading fees, saved time through the digitization of onboarding, and provided access to real-time price, market data, and news. McKinsey recently reported that fintech investment grew from $19 billion in 2015 to $92 billion by 2021, a demonstration of the global enthusiasm for these services. In markets where conventional banking infrastructure is limited, fintechs like Zelle, Venmo, and mobile wallets provided credible alternatives.

That said, barriers still exist. The World Economic Forum identifies latest challenges like; high minimum investment amounts (for lower-income individuals), minimal advice, and complicated systems. Their answers are increasing financial literacy, providing AI-assisted advice, and simplifying complexities to help new investors access this market.

Fintechs have overcome a lot of historical barriers to market entry - but there is further to go for full access (especially for individuals and households who are historically low-income) that requires more inclusive designs and precise interventions. To achieve this, the collaboration of governments and fintech companies is required in order to make better systems for a wider audience that is fair and safe.

Retail participation has been a significant pillar supporting the positive movement in the stock market. In 2020 there were only 4 crore demat accounts, and now there are 14 crore demat accounts by 2024 - with more than 10 crore new investors entering the capital market during this period.

Fintech and Investor Behavior

Access may have increased, but fintechs have also changed how people behave once they enter the market. More firms are incorporating behavioral finance into fintech design.

The Financial Planning Association has stated that emotion is now the lens through which investment behavior is scrutinized. With the rise of meme stocks, social trading, and the advent of retail investing propelled by pandemic life, we now understand how much investment decisions are the result of emotionally triggered behaviors. The pull of fear, overconfidence, and social approval help explain the behavior of investors who act. Patterns emerge that are more reactive than rational.

Three behavioral biases are worth considering, celebrated in the literature as being especially exceptional:

Confirmation Bias—looking for information that affirms your pre-existing beliefs limits the diversity of your exposure.

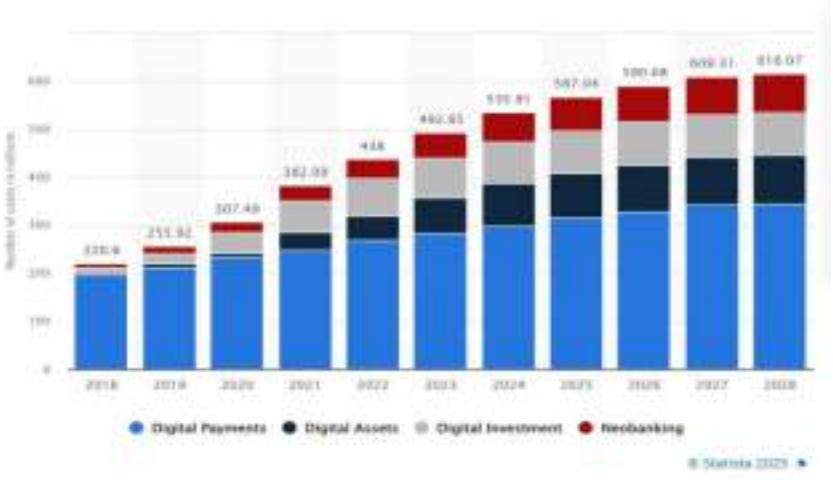
Loss Aversion—ugly, perceived losses outweigh any benefits of possible gains, leading to premature selling, missing profit. It tests the patience and can be extremely difficult to deal with if not approached correctly

Anchoring bias—sticking to an initial piece of information, never adjusting when new information becomes available. This is incredibly harmful as people tend to disregard any sort of suggestion that does not relate well to the initial information they received.

These biases are exacerbated in digital environments, where gamification flourishes through notifications, visual stimulation, and social comparisons. These are also heavily influenced through social media influencers as these influencers achieve something that may not be replicable but because something has been done, multiple people will try it only to fail. Nevertheless, it also provides a positive view by commenting on how AI-enabled personalized fintech apps could develop positive emotional connections with users, improving their decision making and discipline.

In conclusion, fintechs are more than just financial tools. They create behavioral environments that impact risk tolerance and learning curves, and reinforce emotional resilience.

FinTech Growth Statistics



The chart shows a steady rise in global users of digital finance services from 220.6 million in 2018 to a projected 616.07 million in 2028. Digital Payments lead in adoption, while Digital Assets, Investments, and Neobanking also show significant growth, reflecting increasing trust in fintech solutions.

Risks and Challenges

The fintech boom has had its positives, but there are negatives as well. Users are drawn to fintech platforms because of their ease of use all while could also lend itself to rash decision making and over-trading. Many novice investors also have a lack of financial literacy to properly read market trends, or take short-term volatility in stride.

The World Economic Forum warns that improper use and lack of regulatory framework mean new investors will not only be misled by misinformation on the internet, but could also be subject to fraud or a knee jerk emotional mistake in a moment of panic. Trust is easily shattered, and once breached by a technical problem or even ethical dilemma, could erode confidence in a flash.

There is also the structural element of making fintech sustainable. McKinsey has identified economic stresses that fintechs are operating under after their post-COVID boom. Running under the limitations of cost controls, operational efficiencies, and strategic binders or limits to scaling. The danger here laps over the risk of scaling too quickly or compromising the user experience while managing to grow.

Furthermore the biggest risks, by far, could also be the biggest improvements which is the incorporation of AI and high personalisation. AI is still not trusted by everyone and may provide inaccurate information if the model is not trained properly. Also, the AI model can be questioned to lack knowledge if it only recommends certain stocks to all sorts of people. High personalisation could be an issue as it would require more and more data. Data like all of your transaction history is something people would not like to share as a lot more things like location, daily routine etc. can be inferred from this.

Case Examples and Real-World Impacts

The influence of fintech on retail investor behavior can be demonstrated in clear, real world examples, particularly in times of significant market disruption. These examples highlight the way accessibility, emotion, and technology come together to shape investor behavior.

Robinhood and Meme Stock Volatility

There is no one better to highlight than Robinhood itself, a fintech platform known for zero commissions, gamified platform, and at the center of the 2021 meme stock phenomenon. The volatility in trading of stocks like GameStop and AMC were driven primarily by social media and emotional elements such as FOMO - fear of missing out. As retail investors entered into these trades in significant numbers, most being first-time investors, they did not do so because of fundamental and financial analysis. Rather, most entered these trades based off memes on social media and emotional hysteria. The fintech platform then, tied to social influence, generated the potential for emotionally based decision making and leveraged high-risk trading behavior which is what dominates the current market.

COVID-19 and the Growth of Brokerage Accounts

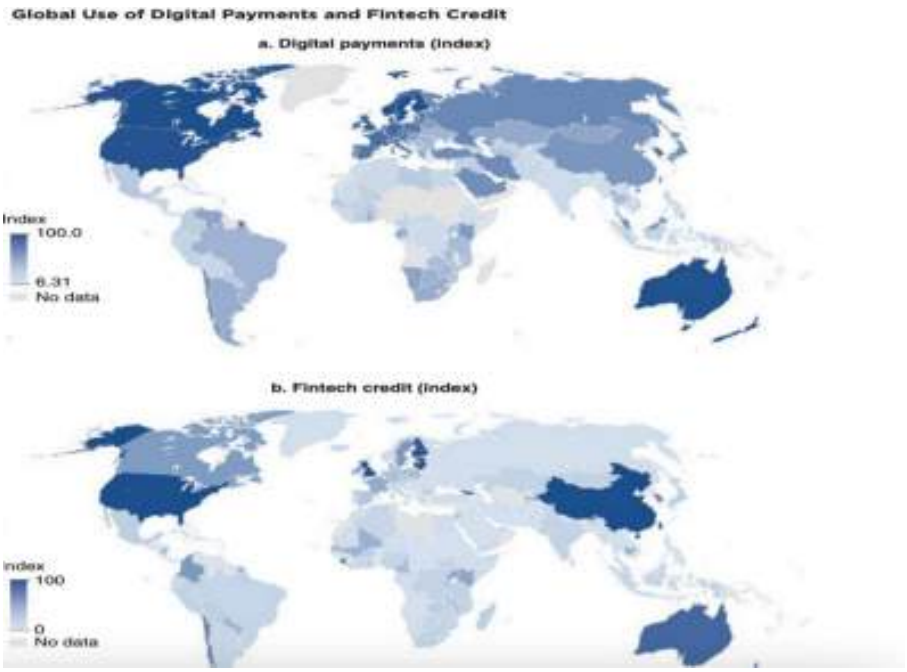
As detailed in the article by the Financial Planning Association, during the pandemic—an unusual time of volatility in global markets—over 10 million brokerage accounts were opened. This time also saw a sudden wave of first-time investors entering the market and many decided to use fintech platforms to invest for the first time. Investors are often

paralyzed by uncertainty, but when they feel empowered by accessibility and the potential for wealth generation—there is no barrier to entry. The rapid increase of participation depicts fintech as a major catalyst in the creation of a new group of retail investing—heavily motivated by the emotional experience drawn from hope, fear, and exhilaration. This fear was also heavily influenced by the pandemic itself as

AI-Powered Personalization in Fintech Apps

While fintech platforms are starting to include AI tools that, adapt to users, the Financial Planning Association's study shows, the idea that apps can build an "emotional connection" with users by understanding personally relevant interests, previously held beliefs, and behavioral pathways is very meaningful for future formulations. The intent of these AI tools is to not only facilitate investing but also offer navigational and emotional guidance, to diminish the impact of behavioral biases (like loss aversion and confirmation bias).

Global Expansion and Inclusion



McKinsey has documented the movement of fintech's to new regions such as Africa, Southeast Asia, and Latin America, where the full banking infrastructure is not available. Given the penetration of mobile and the use of the internet in these regions, at a minimum, fintechs are offering access to payment and investment tools as well as fuel for financial inclusion, meaning that people can store money, send payments, and even invest without directly relying on the traditional banking system.

In each of the scenarios above, we can see how the fintech innovation has effectively decreased the barriers to entry along with changing behaviors in investing; not

only who can invest, but how they behave in financial markets.

Discussion

The confluence of fintech and behavioral finance is proving to be a source of powerful change in retail investing. The research illustrates that fintech has been successful in increasing accessibility to investing, however it is important to note, there can be both pros and cons to accessibility. Accessibility to the market is being affected positively by mobile-first platforms and low-cost services. The accessibility created by these platforms Financial Planning Association articulates in research note that there are many more people wanting to invest, however there also appears to be a lack of confidence, trust and guidance in order for them to be successful. This is where financial literacy and emotional awareness becomes useful and robust financial literacy and financial emotional literacy is still a work in progress for many current fintech platforms.

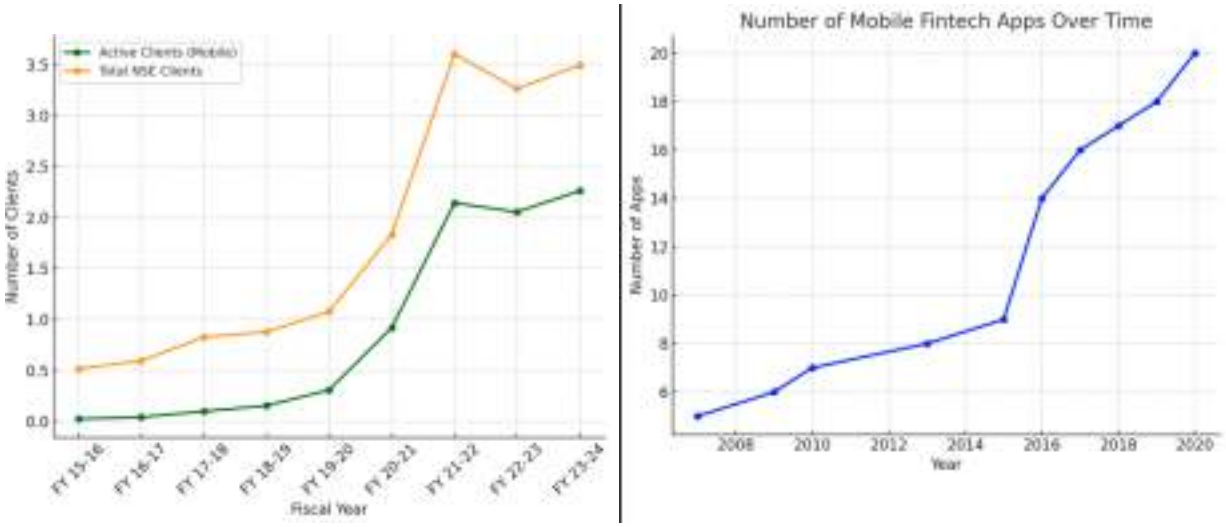
The Financial Planning Association cites in their research notes, how emotion is at the center of how retail investors make decisions, especially new investors. Loss aversion, confirmation bias, and anchoring, are all examples of biases that influence how users make investing decisions, and more often than not, biases can be exaggerated through platform design features. Users are borrowing investments from one another through emotive factors. The meme stock activity provided increased visibility into the emotional aspect of investing, users engaged based on perceived opportunity and fear of missing out and online momentum.

Meanwhile, McKinsey's analysis highlights a shift from hypergrowth towards long-term sustainability in the fintech world. Everybody wants a unicorn, but as fintechs start reaching sustainability, they're pressured to clean up costs and become profitable. While this could be a risk to the user experience if not balanced with the needs of low-income users who rely on accessible and intuitive platforms, many fintechs are coming to realize that emotional intelligence and behavioral support features are just as important—if not more important—than the technology stack they choose.

Some more insight comes from recent fintech trends identified in 2025, most of which highlight a strong connection between fintech and behavioural finance. For example, the rise of fraud by 25% since 2023 with about 12 billion dollars in losses has led to a lot of fear. This growing fear could lead to a lot of irrational and impulsive decisions as well as lead to a delay of actions, which can cause a lot of potential damage to the people in the market. This simply reflects how important emotional stability is when utilising fintechs and being a member of the market as with even the slightest of fear, anything and everything could go wrong. This is also an aspect that most fintechs are working towards as they progress towards a sustainable and safe environment for retail investors.

Parallelly, the use of AI-powered personalisation comes with its own set of behavioural considerations. Sure, it does provide you with saving nudges tailored to your convenience, saving you from randomly investing without discipline, but at the same time, they could be

feeding your confirmation bias and limiting you to a very specific area. Over-personalisation can severely limit exposure that could potentially help you understand the market better and also cause you to stay uncertain when hearing other thoughts, especially on social media through influencers and other personalities.



View of a study of impact of Financial Technology (Fintech) on retail investors in trading securities. (n.d.). https://www.eelet.org.uk/index.php/journal/article/view/1836/1636

The study by Dr. Riya Nathani on the impact of fintech or retail investors in trading securities in India highlights another aspect, which is that the more fintech apps available, the more people engage with them, which is a likely result of habitual engagement and trust that was built over time. We can see that over the years the number of fintechs grew from around 5 to 20 in India and the number of Active clients went from around 1 lakh to about 2.2 crore in 2024 With more and more apps, the gaps are being filled, and fintech industry is capitalizing on any sort of opportunity that it visible.

Finally, the popularity of direct payments, which facilitate trillions of dollars worth of transactions, indicates the need for speed, accessibility, and simplicity by the people. These services have low friction but may make the user feel that they are much safer and simpler than they actually might be. If we consider behavioural implications, impulsive and overconfidence is promoted through the overuse of this app which could lead to careless decisions.

Together, through all of these ideas, we do see that behaviour and accessibility are a part of fintech and how it functions. From its appearance to its features, each comes with its own personality. As it continues to grow with more and more advanced features, Fintechs do need to capture the most behavioural appeal to maximise clients and users as well as gain a better understanding of this new market.

Conclusion

Fintech platforms have truly changed the landscape of the stock market by making investing more accessible, affordable and user-friendly. Accessibility has allowed for growing economies to be capitalised by fintech companies as growing opportunities. It has also helped fintech make its maximum impact as well as stand where it is today. However the influence of fintech goes way beyond making services easier and accessible. It has completely changed the way people invest. Even the smallest of details like notifications, gamified elements, AI assistance, personalisation etc. impact and shape the psychology of a user. Biases and emotions like fear, greed can be built or destroyed with these elements and this is just the beginning.

Fintech is no longer a financial tool but a behavioural ecosystem full of deception. You never know when or where you are being influenced if you do not look at things from a rational view which is extremely difficult in a fluctuation based market. Even the success of fintechs in the future depends on how they present themselves as well as use the behavioural knowledge about their users to their advantage by giving them a sense of safety, stability and building on their trust as time goes on. This is the reason why frauds have increased within fintechs as people are turning more prone to being manipulated simply because of the introduction of emotion. Furthermore, one simply cannot disregard behavior patterns that can be observed in fintech simply because of the sheer quantity of them. From social media influence to even a conversation with a friend will shape your thinking and will affect your decision. In conclusion, fintech has not only opened the door for millions of new investors but changed the way investors think, act and function in the market as a whole. It has led to a new era of investing that is more connected, more empowered but more vulnerable to emotional influences than ever before.

Works Cited

- Feyen, E., Natarajan, H., & Saal, M**.** (2021). Fintech and the future of finance: Market and policy implications (No. 122787, World Bank Policy Research Working Paper). World Bank. https://doi.org/10.1596/1813-9450-9692
- Frankle, D., & Bizzarri, E. (2025, May 9). More people want to invest than ever before. Here's how to help them. World Economic Forum. https://www.weforum.org/stories/2025/05/retail-investment-participation-empowerment
- McKinsey & Company. (2023, October 24). Fintechs: A new paradigm of growth. https://www.mckinsey.com/industries/financial-services/our-insights/fintechs-a-new paradigm-of-growth
- McKinsey & Company. (2024, January 16). What is fintech (financial technology)? https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-fintech • Sullivan, T. (2024, February 7). 10 fintech trends that are defining the industry's future. Plaid. https://plaid.com/resources/fintech/fintech-trends/
- Financial Planning Association. (2024). The intersection of fintech and behavioral finance: A new paradigm for investment. Journal of Financial Planning. • Guillemette, M. A., & Rasure, E**.** (2024). The intersection of fintech and behavioral finance: A new paradigm in investment decision-making. Journal of Financial Planning, 37(1), 44–53. https://www.financialplanningassociation.org/learning/publications/journal/JAN24 - intersection-fintech-and-behavioral-finance-new-paradigm-investment-

Denim Wars: How Celebrity Endorsement Influences People and Shapes Advertising and Brand Culture By Mariia Tarasiuk

ABSTRACT

Celebrity endorsement is a widespread marketing strategy that is used by companies to enhance brand-consumer connection, credibility, and build a sustainable brand perception and image that aligns with a company's mission and values. If done properly, celebrity endorsement is mutually beneficial for both the brand and the endorser. This paper examines the impact of this technique on the advertising industry through an anthropological lens: how society perceives certain brands, and how public buying behavior and a sense of cultural identity are maintained or reshaped as a result of this tactic. By analyzing three recent marketing campaigns that include Beyoncé's collaboration with Levi's, Sydney Sweeney's partnership with American Eagle, and KATSEYE's commercial for GAP, this research shows the influence that celebrity endorsements have on brand image, and provides valuable insights for creating effective, culturally sustainable advertising strategies. Studying these cases and the way they influence public buying behavior demonstrates how consumers are affected by brand partnerships, and how their brand perception is shaped by celebrities promoting certain products or services due to the emotional connection and the associations this tactic builds. While this technique can be used in an effective and sustainable manner, it is important to find an endorser that aligns with the brand's values and supports the company's message; otherwise, if used carelessly, celebrity partnership can harm the brand's reputation and long-term success. A critical part of building a strong brand identity is understanding cultural nuances and fashion preferences in businesses created for certain demographic groups. By following the well-chosen patterns that are approved of by brands' target audiences, companies can create strong and long-lasting identities. Considering the rapidly changing business landscape in the fashion advertising industry, this research, based on recent and relevant marketing cases, can be a useful asset for creating a strong, consistent, long-lasting brand.

INTRODUCTION

Celebrity endorsement is an advertising strategy that uses famous people, such as actors, musicians, or athletes, to promote products or services with the intention of increasing brand image and visibility, while forging a distinct identity. Celebrity endorsement has been actively used in the marketing industry in recent times due to its originality and the emotional brand-consumer connection that it builds. It has become an incredibly powerful tool of advertising in terms of brand sustainability and brand perception due to enhanced credibility, brand differentiation, and target audience alignment.

While celebrity endorsement is a powerful marketing strategy that increases consumer attitude, brand image, and purchase intention, it can also lead to higher levels of skepticism or disgust when exposed to the public too frequently. As this tactic becomes more common, consumers lose their trust and begin to perceive celebrity-endorsed brands with hostility. So, ads

that are created in the commercial endorsement contexts with globally known performers or athletes might often be perceived negatively. However, more natural contexts enhance the consumers' para-social interactions (one-sided interactions, where a person invests their energy into a media figure who is unaware of their existence) with a celebrity and improve brand-related responses (Zhang and Hung 1). In terms of brand sustainability, it is incredibly important to align the brand's values with the celebrity's public image, as well as to build long-term partnerships that eventually enhance brand equity and consumer loyalty, while also emphasizing the role of crisis management in navigating endorsement controversies (Zhang 3,4). In these cases, cultural sustainability can be significantly improved by the celebrity endorsement implementation.

As high fashion celebrity endorsement that heavily relies on respect for market and culture is not a well-researched topic today, many marketers have to rely on the previously studied topics of brand equity in mass production. To be successful, celebrity endorsement should be carefully planned and well executed, even though the majority of consumers claim that this advertising technique doesn't affect their purchase intentions (Jafar et al. 79). This viewpoint, combined with the positive results of endorser-marketing strategies, suggests the idea that most people do not acknowledge being influenced by the ads. However, their subconscious affects their brand perception in a positive, or sometimes negative, way.

Studying celebrity endorsement through a cultural lens is important because it reveals the strong connection between marketing strategies and social values and the formation of consumers' identity. Anthropological analysis helps explain why certain celebrities resonate with specific audiences, and how various cultural symbols are used to build brand authenticity and shape brand perception through both global trends and local cultural contexts. This approach highlights cultural sustainability issues in endorsements, such as overconsumption and harmful stereotypes, as well as potential benefits, like the improvement of social norms. By acknowledging and understanding these dynamics, marketing specialists can accurately predict the long-term cultural impacts of celebrity endorsements, making sure that the branding strategies prioritize not only the short-term success and profitability but also ethical considerations and responsibilities.

This paper will analyze studies from various academic fields to answer the question of how celebrity endorsements as a tactic influence the marketing industry in terms of brand perception and sustainability. Aside from examining how advertisers use celebrity endorsers to promote products, this paper will discuss which brands benefit most from this strategy and why celebrity endorsement can be such a powerful marketing tool. Furthermore, this paper will explore this technique's influence on both the brand and the endorser, and their independent success outside of this collaboration.

While celebrity endorsement is often mutually beneficial for both the brand and the endorser, the objective of this paper is to examine the impact of this technique on the advertising industry, particularly how society perceives certain brands, as well as people's sense of identity and the ways in which traditions and heritage are maintained or reshaped as a result of this tactic. This study reviews relevant articles to examine how celebrity endorsements influence brand

image and to provide valuable insights for developing effective, culturally sustainable marketing strategies.

As this paper explores how advertising techniques influence people's perceptions of brands, I want to briefly explain why I'm particularly interested in this topic. While I've always been a fan of a good marketing campaign – whether it was a makeup, supplement, or a sportswear brand – I wondered what the science behind all these processes was. I've always wanted to see what was going on behind the scenes of famous advertisements that had catchy songs and popular athletes/performers in them. Translating these advertising strategies into real ideas and calculations was a big dream of mine, which served as a reason to do this particular research today.

BACKGROUND

Various business practices and marketing strategies actively shape cultural identity, especially through celebrity endorsements and the popularization of unique traits, origins, ages, and beauty types. Celebrity endorsement is a tactic, in which advertisers use personalities that are well-known in public, whether due to their credibility or attractiveness, to heighten the credibility and increase the effectiveness of commercials (Mberia and Zipporah 162,164). By picking a celebrity of a certain race, ethnicity, age, gender, etc., brands become responsible for building their brand-endorser association and the reputation of the celebrity. While successful marketing campaigns can improve society's view of somebody's language, education, socioeconomic status, and even religion, scandalous ads can make these associations negative, causing stereotypes, anger, and disagreements between different cultural groups.

While celebrity endorsers don't always have to be internationally recognized actors or athletes, it is crucial that they are extra well-known within the brand's target audience. Celebrities assist ads to become more popular, stand out, and increase communicative ability. Moreover, they can help improve and polish the company's image. Not only do celebrities create new interest and excitement in consumers, but they can breathe life into a falling brand. Moreover, they also can create a higher level of recall and make the advertisement memorable, causing the trust level to go up (Adam and Hussain 85,88). This leads to the conclusion that carefully-chosen endorsers can be incredibly beneficial for the brand's short-term and long-term success.

It is also noteworthy that this tactic depends on the country in which the brand is based. Having studied cultural identity, researchers discovered that there's a really wide variety of preferences when it comes to picking a celebrity of a certain nationality. For instance, while most responders from Finland and the U.S. wanted international celebrities to promote high fashion brands, the majority of Indian responders preferred the endorser to be a local celebrity. (Jafar et al. 79) Examples like this show the clear cultural impact on this particular industry, while also emphasizing the importance of geographical nuances in advertising.

A brand is a visual, emotional, rational, and cultural image that is associated with a certain company or a product. A positive or a negative brand perception directly influences the

company's sales and market success. Advertisements create feelings and expectations in a customer and influence consumer associations with the brand when they encounter the brand name. Celebrity endorsements give a brand a touch of glamour and get customers excited by showing them a known face (Das and Padhy 85). Therefore, perceptions of brands are highly influenced by celebrity endorsers and the messages they convey. Brand perception plays a crucial role in the profitability of a company. Hence, it's one of the most significant components of marketing strategies. Owned by consumers, brand perception builds customer relationships, where all companies aim to be seen as positive. Positive brand perception influences consumers to give priority to choosing that brand, while negative perception pushes consumers away. As a brand isn't just considered to be the name of the company, but also the illustration of culture, vision, value, mission, and personality, all of this data combined turns into a mental impression that influences the product's quality and reputation in the eyes of customers (Guliyev 137,139). Overall, brand perception is an essential element of advertising, that consists of multiple factors, and reflects the company's message and views. Taking culture, mission, and vision into consideration when developing a strong marketing campaign can be a strong enhancer of a brand's success and longevity, giving it an advantage over rival companies and boosting its progress and growth.

This topic is studied by a wide range of students in many fields, including marketing, psychology, sociology, and anthropology, as it has elements of culture, media, and consumer behavior. Since no single field can fully explain how celebrity influence works, studying it requires combining viewpoints on human behavior, culture, and the market. Researchers in business institutions examine how celebrity endorsements influence consumer decision-making through psychology and neuroscience, showing that celebrities can increase trust and affect consumer choices (Basiouny). At the same time, sociology students research factors like celebrity-brand fit, cultural relevance, social identity, etc., to learn how they shape consumers' attitudes and purchasing behavior (McCormick). There is also interdisciplinary research in digital media and data science that analyzes the influencer culture and its ethical and cultural impacts, especially in global markets, driven by social media (Goanta et al. 1,2). Overall, these scholars usually study this topic not only to understand how celebrity endorsements influence sales, but also how they shape long-term cultural values and perceptions of brands in society. They use a mix of qualitative and quantitative data to capture both statistics and deeper cultural details. Quantitative approaches mostly include surveys, experiments, and data analysis to measure the subtle changes in consumer behavior and trust after exposure to celebrity endorsements. Qualitative methods, like interviews, focus groups, etc., give researchers a more human perspective on the problem, helping them understand how people actually feel, and what the opinions within various cultural groups are.

To make the brand last longer, sustainability should be considered when choosing a spokesperson for it. The celebrity should match the brand development stage and be consistent with brand tonality and long-term stability to enhance trust. By studying and implementing the sustainable endorsement mechanism, companies can increase their brand equity and learn

effective crisis management techniques (Zhang 5). Brand sustainability has been an active area of research, which includes meeting customer needs and achieving sustainable consumption to build a consistent brand with a stable, successful future. While economic development causes an increase in consumer base, it also results in various challenges, surrounding cultural forces in achieving sustainable market practices (Afzal et al. 1,3). Overall, brand sustainability increases brand longevity, the company-consumer relationship, and trust, resulting in an improved brand reputation and a steady, successful brand progression.

This paper will look at three recent advertising campaigns and analyze how celebrity endorsement influenced brand perception, sustainability, as well as short-term and long-term sales. We will compare the immediate success and later effects of the widely publicized 2024–2025 Levi's campaign featuring Beyoncé, the scandalous 2025 Sydney Sweeney-American Eagle collaboration, and, finally, the viral 2025 Katseye commercial with Gap.

Starting with a nostalgic, 1980s-1990s themed 4-part Levi's commercial, featuring globally recognized musical icon Beyoncé, the company brought a lot of attention to their brand. Three months after Beyoncé dropped her new album, Levi's dropped the first part of their series of advertisements. From September 2024 to August 2025, the brand shared four chapters of the story: "Launderette," "Pool Hall," "Drive-In," and, finally, "Denim Cowboy." The last one sparked the most speculation, with the biggest connection to the "COWBOY CARTER" album, as well as some bits of foreshadowing that some fans believe to be hinting at a rock-styled Act III album, expected in the spring of 2026. A combination of these factors, combined with a "COWBOY CARTER" music hit "Levii's Jeans" created a perfect environment to keep both jean lovers and music fans engaged. While some critics considered the new series of commercials overly-sexualized, most audiences recognized the message behind the story as being about female empowerment, and not female objectification (D'Ambrosio). As older generations resonated with the old-style atmosphere, and the younger audience appreciated the feminist values behind the ad, Beyoncé has helped Levi's earn $65 million in media value, lifting the company's women's business sales by 12%. This partnership exceeded the brand's expectations, gathering 4.3 billion impressions by April of 2025 (Dorisca). In conclusion, this commercial was a great move for both parties: Levi's is actively using the positive views associated with Beyoncé to boost their brand perception and remain relevant for younger generations of consumers, and Beyoncé is benefiting from additional social media engagement and a heightened interest in her music (D'Ambrosio).

Following the Levi's campaign, American Eagle released their new commercial in July of 2025, starring Sydney Sweeney in it. The campaign's tagline – "Sydney Sweeney has great jeans" has sparked huge controversy over the real message behind the ad. In the video, the blonde-haired, blue-eyed, 27-year-old actress says: "Genes are passed down from parents to offspring, often determining traits like hair color, personality and even eye color… My jeans are blue." Viewers found it easy to notice the word play with "jeans" and "genes", which led to harsh criticism that the ad could be interpreted as referencing ideas about genetic traits, race, and conventional beauty standards. On top of that, the audience found similarities between this

commercial and the 1980 Brooke Shields' Calvin Klein ad, which was heavily criticized for sexualizing the 15-year-old actress (Manuel and Inskeep). However, despite global criticism and an unconventional brand message, American Eagle's collaboration with Sydney Sweeney boosted the company's shares by 10% and added an estimated $400 million in market value, while also increasing the clothing retailer's web traffic by 100%. Instead of plummeting, the brand's sales suddenly increased as a result of controversy that led to sustained attention and interest (Naushad). At the same time, while the brand didn't suffer any adverse consequences as a result of the ad, Sydney Sweeney, the brand endorser, faced significant challenges with the sales of her movies. After critics called the ad racist and blamed the young movie star for promoting eugenics, her new film "Americana" came in 16th at the box office and only brought in $500,000 during its opening weekend (Mitchell). So, while the long-term impacts of the controversial American Eagle campaign on their brand perception and image may still be uncertain, this case is a vivid example of how public backlash can expand beyond advertising and undermine the success of an endorser's other professional projects and work.

Lastly, we're going to analyze the 2025 KATSEYE collaboration with GAP, which is often compared to the scandalous American Eagle commercial. The campaign was created with the intention to enhance culturally relevant voices and celebrate individuality, as the global girl group combined sharp choreography with the 2003 hit "Milkshake." In just three months, with a mix of 2000 nostalgia and a rising multinational pop-group popularity, the "Milkshake" clip on GAP's YouTube accumulated over 50 million views and became the most liked and shared TikTok video in the company's history. On Instagram, it also broke the brand's previous 24-hour record and tripled it, reaching 15 million users in its first day (Vaziri). With more than eight billion impressions and 500 million views, the "Better in Denim" campaign attracted a lot of highly-engaged Gen-Z, as well as high-income consumers. The brand's sales grew 7% compared to the previous year, and GAP Inc., which includes Old Navy, Banana Republic, and Athleta, experienced a 5% sales growth (Vlamis). By working with an inclusive Gen-Z girl group, the brand's perception improved, making it look more trendy and bolder. The dynamic, catchy advertisement gave GAP an opportunity to completely reposition itself and make the brand relevant, fun, and exciting again (Xiao).

ANALYSIS

All three campaigns were a huge success in the denim industry, increasing brands' sales, impressions, and overall engagement. They demonstrate that celebrity endorsements influence not only consumer purchasing decisions but also the cultural meanings that consumers attach to brands, ultimately shaping long-term brand identity and perception. Beyoncé and Levi's collaboration was a premium, strong, well-aligned campaign; Sydney Sweeney's endorsement for American Eagle was highly controversial, however, the most revenue-boosting; and KATSEYE's partnership with GAP created a huge cultural moment, praised for its inclusivity and diversity.

Personally, I find GAP's approach to be the most influential on the entire jean industry. While not being controversial, but rather promoting diversity, the KATSEYE campaign helped the brand not only gain huge sales and market success, but also strengthen its long-term perception. By outperforming their rivals in engagement during the "denim ad war", GAP had the biggest cultural impact and completely redesigned its own image to be more relevant and engaging. The company's marketers did a wonderful job combining the high-stakes conflict with a positive, clear, and sustainable message. While not immediately gaining as much profit as American Eagle, GAP managed to enhance its brand-customer connection and credibility to secure a successful, long-lasting future.

While writing and analyzing this research, I've learned that, when chosen well, celebrity endorsements can be a highly effective tactic for almost any company or brand. By creating a sustainable campaign that is well-aligned with both the brand's and the endorser's values and public image, both sides can benefit from this collaboration process a lot. Celebrity endorsements can boost engagement, sales, and absolutely reshape the perception of certain brands. They can be an extremely effective tool to promote certain products or services with sustainable, long-lasting results.

Ultimately, this paper emphasizes the importance of having a clear, distinctive brand identity. In the United States, in particular, audiences heavily rely on the brand's value and cultural mission. By finding a company that aligns with their views and beliefs, customers actively boost the brand's sales and engage with its posts, helping products and services gain more recognition and market success. By using celebrity endorsements to shape brand identity, increase credibility, and promote their cultural views, companies can attract the right audiences that will actively support and assist them in a long-term perspective in a consistent, sustainable way. Similar techniques accelerate consumer decision-making and increase the emotional connection between the customer and the brand, making audiences more likely to purchase certain products and pay more for them. By purchasing products or services tied to celebrities people would like to associate with, buyers feel a deeper sense of fulfillment and satisfaction that leads to them having a stronger desire to purchase these products again.

CONCLUSION

This paper answered the question of how celebrity endorsement as a tactic impacts the marketing industry in terms of brand perception and brand sustainability. Through analyzing Beyoncé's campaign for Levi's, Sydney Sweeney's collaboration with American Eagle, and KATSEYE's advertisement for GAP, it was discovered that celebrity endorsement is an extremely powerful advertising tool for creating a long-lasting brand image that directly influences brand perception and sustainability. The findings of this study suggest that celebrity endorsement is most effective when it extends beyond simple product promotion and becomes part of a broader cultural narrative. Consumers increasingly expect brands to reflect values and identities that resonate with their own experiences, making authenticity and cultural relevance essential components of successful marketing campaigns. As celebrity endorsement can create a stronger sense of credibility and a deeper connection between a customer and a brand, it serves

as a useful tool to strengthen the para-social interactions with consumers all over the world. In a society where social media plays a big part in people's everyday lives, using brand partnerships and similar marketing strategies is a smart step to enhance brand engagement, trust, and global notability of a company.

For future research, I suggest studying an endorser's perspective on brand partnerships and how they affect a celebrity's reputation and further career success. Studying celebrity endorsements from a different perspective could help marketers and advertisers better understand this technique and develop more effective, sustainable ways of using it for their brands. As social media continues to blur the boundaries between celebrity culture and consumer behavior, the importance of strategically aligned endorsements is likely to increase, making celebrity partnerships an even more influential component of modern brand-building.

Works Cited

Adam, Muhammad Amir, and Nazish Hussain. “IMPACT OF CELEBRITY ENDORSEMENTS ON CONSUMER BUYING BEHAVIOR.” British Journal of Marketing Studies, 2017. European Centre for Research Training and Development UK, https://www.researchgate.net/profile/Muhammad-Amir-Adam/publication/319529018_Impact_of_Celebrity_Endorsement_on_Consumer_Buying_Behavior/links/59b14300a6fdcc3f888dd462/Impact-of-Celebrity-Endorsement-on-Consumer-Buying-Behavior.pdf.

Afzal, Farman, et al. “Market Sustainability: A Globalization and Consumer Culture Perspective in the Chinese Retail Market.” MDPI, 2019. MDPI, https://www.mdpi.com/2071-1050/11/3/575.

Basiouny, Angie. “The Marketing Psychology Behind Celebrity Endorsements - Knowledge at Wharton.” Knowledge at Wharton, 30 May 2023, https://knowledge.wharton.upenn.edu/article/the-marketing-psychology-behind-celebrity-endorsements/. Accessed 18 April 2026.

D'Ambrosio, Dorina. “Levi’s in-jean-ious Beyoncé ads are a masterclass in acceptable sexualization.” The Drum, 6 March 2025.

Das, Saumendra, and Prasant Kumar Padhy. “Brand Perception by Celebrity Endorsement.” International Journal of Advanced Research in Management and Social Sciences, vol. 1, no. 2, 2012. International Journal of Advanced Research in Management and Social Sciences, https://download.ssrn.com/15/04/09/ssrn_id2592412_code1391045.pdf?response-content-disposition=inline&X-Amz-Security-Token=IQoJb3JpZ2luX2VjEND%2F%2F%2F%2F%2F%2F%2F%2F%2F%2FwEaCXVzLWVhc3QtMSJHMEUCIEby3aHulEPdTvmtjFV48WbMutxLz5Ke9ka8Ap%2BPtDOKAiEAkKra3zj1jvs.

Dorisca, Samantha. “The Beyoncé Effect Strikes Again With Levi's Partnership Earning $65M In Media Value And Contributing 12% Increase In Its Women's Business In Q3.” AFROTECH, 8 December 2025, https://afrotech.com/beyonce-partnership-levis-led-to-65m-in-media-value.

Goanta, Catalina, et al. “Computational Studies in Influencer Marketing: A Systematic Literature Review.” arXiv, 2025. Cornell University, https://arxiv.org/abs/2506.14602?utm_source=chatgpt.com.

Guliyev, Seymur M. “THE IMPACT OF BRAND PERCEPTION AND BRAND IMAGE ON CONSUMER PURCHASING BEHAVIOR IN AZERBAIJAN.” Science, Education and Innovations in the context of modern problems., vol. 6, no. 1, 2023. IMCRA, https://imcra-az.org/uploads/public_files/2023-02/article_seymur_the-impact-of-brand-perception-and-brand-image-on-consumer-purchasing-behavior-in-azerbaijan-1.pdf.

Jafar, Amir, et al. “Celebrity endorsements vis-á-vis brand equity of high fashion products.” ProQuest, 2011, p. 79, https://www.proquest.com/docview/2622618473?fromopenview=true&pq-origsite=gscholar&sourcetype=Scholarly%20Journals.

Manuel, Obed, and Steve Inskeep. “Does Sydney Sweeney's American Eagle 'great jeans' campaign mark a shift for ads?” NPR, 29 July 2025, https://www.npr.org/2025/07/29/nx-s1-5482837/american-eagle-sydney-sweeney-jeans-ad#:~:text=The%20campaign's%20tagline%20—%22Sydney%20Sweeney,personality%20and%20even%20eye%20color.

Mberia, Hellen K., and Mwendwa Mildred Zipporah. “The Effect of Celebrity Endorsement in Adverisements.” Open Access Journal, vol. 3, no. 5, 2014. International Journal of Academic Research in Economics & Management Sciences, https://knowledgewords.com/index.php/ijarems/article/view/1320/1336.

McCormick, Karla. “Celebrity endorsements: Influence of a product-endorser match on Millennials attitudes and purchase intentions.” Journal of Retailing and Consumer Services, vol. 32, 2016. ScienceDirect, https://www.sciencedirect.com/science/article/abs/pii/S0969698916300212.

Mitchell, Taiyler S. “Sydney Sweeney's New Niche Film Flops After American Eagle Ad Controversy.” HuffPost, 18 August 2025, https://www.huffpost.com/entry/sydney-sweeney-new-film-box-office-american-eagle-ad_n_68a2642ce4b0858c7642bcfa. Accessed 26 April 2026.

Naushad, Abdul Azim. “Sydney Sweeney's Jeans Ad Added $400M to American Eagle — Report.” Mandatory, 24 November 2025, https://www.realitytea.com/2025/11/24/sydney-sweeney-jeans-ad-american-eagle-sales/. Accessed 26 April 2026.

Vaziri, Aidin. “Gap's viral Katseye 'Milkshake' ad lifts sales and revives SF brand.” San Francisco Chronicle, 20 November 2025, https://www.sfchronicle.com/entertainment/article/gap-earnings-katseye-milkshake-21199561.php. Accessed 26 April 2026.

Vlamis, Kelsey. “Gap CEO says its viral denim campaign wasn't just a hit online — it drove double-digit growth in sales, fueled by Gen Z.” BUSINESS INSIDER, 20 November 2025, https://www.businessinsider.com/gap-ceo-viral-denim-campaign-sales-growth-gen-z-2025-11.

Xiao, Yi. “Why I Walked Into GAP After Years — And Left With Nothing.” Medium, 19 September 2025, https://medium.com/marketing-in-the-age-of-digital/why-i-walked-into-gap-after-years-and-left-with-nothing-c44b7efea815.

Zhang, Ke, and Kineta Hung. “The Effect of Natural Celebrity-Brand Association and Para-Social Interaction in Advertising Endorsement for Sustainable Marketing.” MDPI, 2020. MDPI, https://www.mdpi.com/2071-1050/12/15/6215.

Zhang, Ruotong. “Sustainable Celebrity Endorsement Strategies: Long-term Impacts on Market Development Stages.” SHS Web of Conferences, 2025. SHS Web of Conferences, https://www.shs-conferences.org/articles/shsconf/pdf/2025/11/shsconf_iclrc2025_01007.pdf.

Beyond Translation: Barriers to Tax Compliance Among Limited English Proficient Immigrant Taxpayers By Isaac Kwun

ABSTRACT

Federal tax compliance is often difficult even for native English-fluent Americans. It is often more difficult for immigrant taxpayers with limited English proficiency (LEP). This literature review examines four types of barriers facing LEP immigrant taxpayers: linguistic, psychological, administrative, and technological. This review argues that these four types of barriers can reinforce one another and make literal translation of existing materials insufficient. The review also identifies a gap in the literature: no study identified in this review has directly tested simplified bilingual tax materials against standard IRS translations. Closing this gap may have a meaningful impact on unintentional noncompliance and forfeited tax benefits among immigrant taxpayers, especially those who are LEP and undocumented.

Keywords Limited English Proficiency (LEP), Tax Literacy, Undocumented Immigrants, Linguistic Barriers, Psychological Barriers, Administrative Barriers, Technological Barriers

INTRODUCTION

Paying federal taxes in the United States is a difficult process, even for native-level English-speaking, non-immigrant Americans. A 2024 national survey conducted by the Tax Foundation and Public Policy Polling showed that only two percent of respondents demonstrated "proficient" tax knowledge. More than half of taxpayers with a postgraduate degree fell into the "beginner" category (Callaway et al.). Experts, including economists, also acknowledge the confusing nature of the American tax system, especially when compared to that of other nations. This complexity requires taxpayers to spend more time filing returns than residents of comparable nations (Towe). For instance, the U.S. government does not pre-fill returns using employer and bank data, unlike twenty-eight other OECD countries ("What Are Prepopulated Tax Returns?").

This difficulty increases sharply for taxpayers who lack full proficiency in English. The Migration Policy Institute states that 28.9 million people in the United States have limited English proficiency. The majority of them, 23.5 million, are immigrants (Batalova). Nonetheless, many of these immigrant taxpayers who are not fluent in English are full participants in the tax system. For instance, in 2022, undocumented immigrants alone paid an estimated $96.7 billion in federal, state, and local taxes (Davis et al.).

The systems that serve them, however, are not keeping pace. The National Taxpayer Advocate reports that roughly one in five taxpayers eligible for the Earned Income Tax Credit fails to claim it ("Tax and Financial Literacy" 106). Rather than expanding language access to support undocumented immigrants and individuals with limited English proficiency, the Internal Revenue Service recently reduced the number of languages regarding offered written and web-based materials from twenty to seven, after a March 2025 executive order (Cohn).

The practical consequences outside of statistics are also visible by the choices immigrants make regarding tax refunds. Many immigrants choose not to claim refunds they are legally owed. Law professor Francine Lipman describes filing a tax return as a "staggering exercise" for undocumented immigrants, who must disclose their mailing address and employer to the federal government they fear could use that information against them. The Internal Revenue Service treats a mismatched Social Security number on an Individual Taxpayer Identification Number return as an implicit admission of unauthorized work. Visiting an IRS office or any government building can also be "very intimidating" for undocumented immigrants. Together, these factors lead many immigrants to forgo money that they are owed for fear of contact with a federal agency (Lipman 120). Lipman argues that there is an unofficial "undocumented immigrant tax," a burden that undocumented immigrants bear in part through fear rather than statute.

The status quo described above, in which immigrant taxpayers fall into noncompliance or forfeit tax benefits that they are legally entitled to, is caused by multiple factors. The following sections examine each type of barrier in turn and show how these barriers frequently reinforce one another. These factors are largely composed of two levels, individual and systemic. Individual level barriers are linguistic and psychological, while systemic level barriers are administrative and technological. The following sections examine each of these four barriers, and then examine the more difficult problem when a single LEP taxpayer encounters several of these problems at once.

1. INDIVIDUAL LEVEL BARRIERS

For taxpayers with limited English proficiency (LEP), difficulty navigating the U.S. tax system often begins before calculations occur. Taxpayers have to go through several steps such as interpreting specialized terms, determining whether rules apply to their own circumstances, and translating written instructions into specific decisions. Such tasks create two closely related individual-level barriers for LEP taxpayers: linguistic difficulty in understanding tax information and psychological uncertainty regarding the understanding and application of it.

1.1 Linguistic Barriers

Language access has been a constant issue for LEP taxpayers. Lee argues that taxpayers with limited English proficiency face inherent barriers to exercising their rights under U.S. tax law (791). The study also examines the IRS's responsibility to make the tax system accessible to LEP taxpayers. Although the IRS has developed language-access policies, the study finds that converting those language-access policies into meaningful access is a different issue. While tax information may be available in different languages, this does not necessarily lead to all taxpayers being able to understand and use it.

The linguistic problem is broader than whether tax information is provided in English or in other languages. This is because tax forms contain tax language that is often specialized and therefore difficult to interpret through regular basic vocabulary understanding. For instance, Manandise analyzed 234 IRS forms, schedules, and worksheets from the 2017 tax year and

identified recurring linguistic features that often complicated interpretation. Around 43.6 percent of the text segments in these documents examined had at least one domain-specific multiword term (Manandise 37). The study also found that the tabular structure of these documents often compresses information into noun phrases and fragments, making relationships between tax concepts implicit.

This distinction is important because tax comprehension by nature requires more than simply recognizing translated words. A taxpayer may understand the ordinary meaning of each word in a phrase within a tax document. However, this is often not enough; they must also know the phrase's meaning within the realms of tax law and calculations. Manandise illustrates this issue regarding calculation instructions, where individual words appear to be relatively straightforward, but combined operations of these words require the reader to identify conditions, apply percentages, and perform calculations in the correct order. An example of a sequence of calculation instructions within a form analyzed in this study had a Flesch-Kincaid grade level of 19.5, which is a highly skilled reading level (Manandise 40). Therefore, the issue of underlying conceptual and procedural complexity remains, even though a direct translation may reduce one layer of difficulty.

Blank and Osofsky provide a related explanation through their concept of simplexity (193). They suggest that the IRS does frequently convert complex tax law into plain-language guidance so that taxpayers can more easily understand it. However, this simplifying process does not eliminate the complexity of the underlying law. Plain-language explanations may omit exceptions, present legally uncertain questions as straightforward rules, or otherwise reduce some legal precision for the sake of accessibility. Their analysis therefore shows a distinction between simplifying the law itself and simplifying its explanation.

Together, these studies suggest an important distinction between translation and comprehension. While translation can improve language access by making information available in a taxpayer's preferred language, it does not necessarily reduce the technical vocabulary, implicit relationships, conditional reasoning, or calculations contained in the original tax material. For LEP taxpayers, translated IRS materials may therefore remove part of the language barrier without fully resolving the difficulty of interpreting and applying the underlying tax rule. Existing literature shows the importance of language access and the complexity of tax communication but provides much less evidence about whether a simplified bilingual explanation leads to better comprehension than conventional translation.

1.2 Psychological Barrier

While the introduction describes one psychological barrier, the fear of contact with a federal agency which leads to forgoing owed refunds, as Lipman argues, there are other psychological barriers that do not depend on immigration status alone.

Difficulty understanding tax information may also create a second individual-level barrier: uncertainty about whether the taxpayer is completing the return correctly. Tax

compliance goes beyond financial and time costs; there are subjective costs as well. The experience of trying to comprehend and correctly complete the process raises the subjective cost.

Choi and Jurow Kleiman directly examine this subjective cost of tax compliance through a survey that measures how much taxpayers value reductions compared to the potential unpleasantness regarding filing a return. Their findings show that respondents were more concerned about making unintentional mistakes than about the time they spent completing their taxes (1283). The authors therefore argue that policies aimed at reducing tax-compliance burdens should focus on simplifications that reduce the possibility or perceived possibility of mistakes, in addition to saving time (1314).

These findings are especially relevant to taxpayers who are uncertain about whether they have correctly interpreted tax terminology and instructions. If the primary concern is not simply the amount of time needed to complete a form but whether an unnoticed misunderstanding could lead to an incorrect return, linguistic difficulty becomes a psychological burden as well. A taxpayer who is unsure about the meaning of tax terms such as a qualifying dependent, filing status, or eligibility requirement not only determines an answer but also decides whether they trust their own interpretation. For LEP taxpayers, this uncertainty may increase when the tax rule is being processed in a language they are less familiar with or through translated material that retains the structure and terminology of the original tax document.

Research on simplified tax drafting provides additional evidence that the way tax rules are shown can affect taxpayers' psychological experiences. Woellner et al. compared traditional and simplified versions of Australian tax legislation using a mixed-method design that included problem-solving tasks, observations of non-verbal behavior, self-reports, and focus groups. Both novice and expert participants reported more positive experiences when working with the simplified version. However, the researchers found no statistically significant improvement in participants' ability to apply the law correctly or at the time required to apply it.

This finding introduces an important distinction. While simplification can improve how taxpayers feel about navigating tax information, it does not necessarily improve whether they apply it correctly. Conversely, an intervention designed to improve tax comprehension should not be assumed to reduce psychological burden unless outcomes regarding both aspects are measured. Psychological comfort, comprehension, and filing accuracy are related yet distinct outcomes.

The existing literature therefore leaves an important gap regarding the intersection of language and psychological burden. Studies of LEP taxpayers establish that language access can restrict effective participation in the tax system, while studies about tax language suggest that technical complexity persists even when information is written in apparently clear language. Separate studies show that taxpayers are particularly concerned about unnoticed filing errors and that simplified drafting can reduce some psychological burden. Yet all of these studies are rarely examined together, particularly regarding limited-English-proficient immigrant taxpayers in the United States.

This gap raises a more specific question than whether translated tax materials should exist. It asks whether explanations that combine a taxpayer's preferred language with deliberate simplification can improve the taxpayer's ability to understand and correctly apply tax rules compared to merely providing a translated version of the same underlying complexity. Examining comprehension and simulated filing accuracy separately can therefore help determine whether simplified bilingual explanations address barriers that translation alone may leave unresolved.

2. SYSTEMIC LEVEL BARRIERS

Unlike the individual-level barriers described above, there are also several structural-level barriers that have to do with the system itself. These are also often called administrative burdens by public administration scholars. Moynihan, Herd, and Harvey divide them into three kinds: the learning costs of discovering what a program requires, the compliance costs of satisfying it, and the psychological costs of dealing with the state. They claim that such burdens are chosen, rather than accidental. Immigrant taxpayers face the complexity described in the introduction and, in many cases, a second set of procedures built only for them.

2.1 Administrative Barriers

Many immigrant taxpayers are unable to obtain a Social Security number, and file taxes using an Individual Taxpayer Identification Number, or ITIN, instead. These ITIN applications are still processed by the Internal Revenue Service entirely on paper, with no online option. Such processes lead applications, even if they have no problems, to be processed for seven to eleven weeks ("ITIN Processing").

Within these applications, the applicants' identity must be proved. There are currently thirteen types of documents that are accepted, which is a cut number from forty during the 2003 revisions. The original documents must be mailed unless the applicant visits a Certifying Acceptance Agent. These agents are sparsely distributed across the country. In 2023, only 81 such agents worked at free tax-preparation sites in seventeen states ("IRS Processing"). So, even if an immigrant taxpayer who speaks English fluently prepares everything error-free, the whole process may take eleven weeks, and they would have to mail away a passport.

The issue is that these long, complicated processes screen people out as well, in addition to delaying. The Treasury Inspector General for Tax Administration found that the IRS received 1,120,047 ITIN applications in 2020 and rejected 225,850 of them, roughly one in five ("Administration of the ITIN Program"). The situation especially becomes more severe when there are rule changes. For instance, in 2012, the number of applications received fell by 21.6 percent from the same period a year earlier, while rejected applications rose 46 percent after the IRS's tightening of document requirements in June (2012 Annual Report). Changes in government processes impacting applicants are also shown in an example regarding closing Social Security field offices; disability enrollment was cut by sixteen percent, concentrated

among applicants with the least education (Deshpande and Li 215). This occurred, despite the eligibility rule remaining the same.

ITIN applicants are often excluded from certain rules outright, compared to SSN holders. For example, the 2017 Tax Cuts and Jobs Act limited the Child Tax Credit to children who have Social Security numbers. Roughly one million children who had previously qualified lost eligibility (Fano; Vimo 177). This also applies to the Earned Income Tax Credit. Suro and Findling write, "ITIN filers have the same income tax due as Social Security filers, but they do not receive the same credits" (251). These changes are reflected in filing behavior as well. ITINs claimed for dependents fell from 3.7 million in 2016 to 0.8 million in 2022 ("Most Serious Problem #7" 96). Clearer writing may reduce informational barriers, but it cannot change statutory eligibility restrictions.

2.2 Technological Barriers

As mentioned previously, those who hold a Social Security Number have access to digital tools that may help the tax filing process. So, first-time applicants with ITINs cannot file a return electronically using an ITIN during the calendar year the ITIN is assigned. The instructions for Form W-7 direct first-time applicants to attach the form to a paper return and mail it to another office in Austin, Texas. (*Instructions for Form W-7* 5). A similar case is shown in Norway, where child benefits are awarded automatically from registry data. Those whose records do not fit must apply by hand, causing low income households to carry the heaviest burden (Larsson). Automation may help a lot of people, but those left out of it suffer the most from it.

The availability of digital tools does not eliminate the digital divide. There are gaps regarding both physical access and the skills required to utilize these digital tools, as distinguished by Van Deursen and van Dijk (893). The Pew Research Center claims that 68 percent of Hispanic adults had home broadband, which is 13 points less than that of White adults (McClain and Bishop). Ono and Zavodny found a positive association between English ability and technology access and use among immigrants of the United States (1455). Therefore, although digital online tools are not completely unhelpful, the filers who are least equipped to absorb it carry the heaviest burden.

Section Conclusion

Neither of these barriers reflects a deficit of the taxpayer. The issues mainly arise from how the system is built separately for ITIN applications and how gaps regarding technology exist. Thomson et al. show that these barriers are shown not only by procedure but by outcomes as well, by showing that limited English proficiency predicted lower Earned Income Tax Credit receipt amongst income-eligible families (1725).

COMPOUNDING EFFECTS AND THE LIMITS OF TRANSLATION-ONLY INTERVENTIONS

The various barriers mentioned above are often studied separately. However, these barriers often occur one after another or compound in practice. Many taxpayers with limited English proficiency struggle to understand English tax forms and notices and depend on paid preparers. Around eighty percent of ITIN-related returns were completed by paid preparers between 2018 and 2023, according to the National Taxpayer Advocate ("IRS Processing" 234). ITIN applications, as mentioned before, are paper-based, so this causes other barriers regarding long waiting times and mailing of original identity documents ("Most Serious Problem #7" 89, 94). A rejection during this process will additionally require the taxpayer to understand the IRS notice to correct the issue.

Administrative problems become even more difficult when language barriers are involved. Although fluent English speakers are more likely to understand what went wrong with their rejected application, LEP taxpayers often struggle to do so. The IRS has sent 969,863 ITIN renewal notices in English but only once in Spanish in 2019 ("Multilingual Notices" 69 n. 2). Spanish notices were around a third of a percent of all IRS notices that year, even though it was estimated by the Taxpayer Advocate that almost four percent of working-age taxpayers were Spanish speakers with limited English fluency (“Multilingual Notices” 73). Although research reviews steps separately instead of reviewing a full chain of events, the possibility of barriers compounding together into a more serious problem exists.

DISCUSSION

As the literature reviewed for this paper demonstrates, the assumption that simplified bilingual tax instructions are more likely to increase the comprehension of limited-English-proficient immigrant taxpayers compared with the IRS's own translated materials is not fully supported. Simplified instructions could potentially increase comprehension through a mechanism distinct from the translation of official tax documents. While language-access studies confirm the value of providing tax information to limited-English-proficient taxpayers in the language they prefer, linguistic analyses of IRS materials illustrate that tax jargon, compressed instructions, conditional language, and the multistep nature of calculations present additional barriers to comprehension (Lee; Manandise). The concept of tax simplexity developed by Blank and Osofsky indicates that presenting tax law information in a way accessible to non-experts is itself a complex task. This suggests that simplification may address a different set of communication challenges than translation alone. Simplified bilingual instructions, then, may serve as an additional tool for increasing taxpayer comprehension, but the evidence presented in the literature reviewed for this paper does not demonstrate that the method has an advantage over the alternatives currently used by the IRS.

Any potential benefit must also be considered across different outcomes. Language access, comprehension, confidence, correct application of a tax rule, and simulated filing accuracy should not be conflated. Taxpayers may find explanations easier to understand without applying the tax rule more accurately. For example, Woellner et al. found that participants reacted positively to the proposed simpler versions of the Australian tax code, but simplification

had no statistically significant effect on the accuracy of applying the tax code. At the same time, Choi and Jurow Kleiman demonstrated that concern about making inadvertent mistakes is a significant psychological cost of tax compliance. These findings suggest that simplification may alleviate some psychological burden or lead to more positive experiences without necessarily changing objective outcomes. Therefore, it is important to assess comprehension and simulated filing accuracy separately and to consider confidence or uncertainty as distinct outcomes.

The systemic barriers identified in the literature reviewed in this paper present limitations to an informational intervention. Processing delays, administrative errors, documentation requirements, statutory restrictions on tax credits, and barriers to accessing digital tools cannot be eliminated simply by providing more accessible explanations of tax rules. A taxpayer who understands the subtleties and complexities of specific rules may still face a systemic issue that is unrelated to information dissemination. Systemic and individual barriers might also combine in ways that complicate interactions with the tax system. One issue, such as a taxpayer's application being rejected, might require further action on the taxpayer's part to interpret the reason for the rejection and determine how to respond. The combination of systemic issues that can disproportionately affect immigrant taxpayers and individual barriers to tax compliance may further complicate their ability to complete their filings. Interactions with the tax system may become even more difficult for LEP taxpayers when instructions are presented using technically complex or primarily English-language terms. The creation of simplified bilingual versions of tax materials should therefore be viewed as an attempt to mitigate one informational barrier to tax compliance rather than as a solution to the broader systemic barriers immigrant taxpayers may face.

This distinction also makes clear that simpler bilingual explanations are better seen as supplements to IRS guidance rather than substitutes, since simplification, while potentially making the text more accessible, can also conceal nuances and exceptions (Blank and Osofsky 221). Supplementary explanations can make official guidance more comprehensible while still preserving taxpayers' access to the authoritative material, thereby addressing both accessibility and legal precision. However, the literature reviewed cannot confirm the effectiveness of this approach, since none of the studies attempted to directly measure such an effect. The reviewed articles address language barriers to tax compliance, difficulties caused by complex tax-specific language, associated psychological costs, procedural barriers, and attempts at simplification in various ways and across different populations, including non-U.S. populations. However, none of the articles attempt to compare the effectiveness of simplified bilingual explanations as supplements to existing IRS guidance for immigrant taxpayers with limited English proficiency while separately measuring comprehension and simulated filing accuracy. The literature offers ample justification for such a study, but none of the research reviewed attempted this specific comparison. A direct empirical comparison would therefore be useful in determining whether such an approach offers benefits beyond existing translated guidance and whether improved comprehension translates into better performance on simulated filing tasks.

LIMITATIONS, CONCLUSIONS, AND SUGGESTIONS

Because this is a literature review, it synthesizes existing studies rather than collecting new empirical data. It describes what researchers have found regarding these four barriers and what gaps could be found in the existing literature. This limits the review's ability to establish whether one specific barrier causes noncompliance or whether fixing a barrier would reduce issues.

Treating LEP taxpayers as a single group is also an issue. As in most groups, there are many real differences between members within the LEP taxpayer group. These differences include language, reading ability in the taxpayer's first language, immigration status, income, education levels, etc. A large portion of the research also mainly deals with Spanish-speaking LEP taxpayers, which may be limiting in terms of describing the experience of LEP taxpayers.

This paper also includes administrative evidence from the United States government, including the IRS, Treasury Inspector General, and the National Taxpayer Advocate reports. Policies and numbers change and fluctuate every year, and past patterns cannot necessarily predict future patterns regarding government action. Therefore, the relevance of this paper may go out of date as time passes. The review also mainly covers federal filing, rather than state or local taxes. This literature review concludes that tax compliance regarding LEP immigrant taxpayers is more than a language problem. The literature points out four types of barriers operating at two levels.

At the individual level, tax language is specialized and compressed. This creates doubt regarding whether a tax return is correct, which is a cost LEP immigrant taxpayers often pay. At the systemic level, the ITIN process involves paper filing, document requirements, delays, rejections, and rules that rule out ITIN holders. This is an issue beyond the power of clearer writing. The digital tools that are often used to make filing easier are less available to filers who may need them the most.

Translation and comprehension are not the same thing; giving a taxpayer a form in their language may remove one layer of difficulty, but other issues remain. These issues include technical tax vocabulary, unstated links between concepts, and conditional rules. The aforementioned barriers also often compound and don't arrive one at a time. Therefore, simplified bilingual explanations are a supplement and may address an informational barrier but fail to address the other structural barriers.

The clearest next step this review follows from the gaps found in this review. Researchers should test simplified bilingual explanations against standard IRS translations with LEP taxpayers. Comprehension and simulated filing accuracy should be measured separately. Participants may find simplification more comfortable to use, but this does not necessarily mean that they applied the tax rules more accurately.

More research is needed regarding the administrative burden and the technological burden, which specifically deals with aspects such as the ITIN process and how the digital divide affects tax filing. Further research should also include a more linguistically diverse group of immigrants, as current studies mainly target Spanish-speaking immigrants.

Works Cited

Batalova, Jeanne. “Frequently Requested Statistics on Immigrants and Immigration in the United States.” Migration Information Source, Migration Policy Institute, 12 Mar. 2026, migrationpolicy.org/article/frequently-requested-statistics-immigrants-and-immigration-united-states.

Blank, Joshua D., and Leigh Osofsky. "Simplexity: Plain Language and the Tax Law." Emory Law Journal, vol. 66, no. 2, 2017, pp. 189–263. scholarlycommons.law.emory.edu/elj/vol66/iss2/1/.

Callaway, Zoe, et al. "National Tax Literacy Poll: Assessing Taxpayer Knowledge and Perceptions of the US Tax System." Tax Foundation, 18 Dec. 2024, taxfoundation.org/research/all/federal/us-tax-literacy-poll-knowledge-perceptions/.

Choi, Jonathan H., and Ariel Jurow Kleiman. "Subjective Costs of Tax Compliance." Minnesota Law Review, vol. 108, no. 3, 2024, pp. 1255–1334. minnesotalawreview.org/wp-content/uploads/2024/02/3.2_Choi-Jurow-Kleiman.pdf.

Cohn, Michael. "IRS Reduced Multilingual Services after Trump Order." Accounting Today, 3 Aug. 2026, accountingtoday.com/news/irs-reduced-multilingual-services-after-trump-order.

Davis, Carl, et al. "Tax Payments by Undocumented Immigrants." Institute on Taxation and Economic Policy, 30 July 2024, itep.org/undocumented-immigrants-taxes-2024/.

Deshpande, Manasi, and Yue Li. "Who Is Screened Out? Application Costs and the Targeting of Disability Programs." American Economic Journal: Economic Policy, vol. 11, no. 4, 2019, pp. 213–48. https://doi.org/10.1257/pol.20180076.

Fano, Arianna. "Barriers to Tax Credit Education and Take-Up Among Immigrant Communities." Bipartisan Policy Center, 15 Dec. 2023, bipartisanpolicy.org/article/barriers-to-tax-credit-education-and-take-up-among-immigrant-communities/.

Internal Revenue Service. Instructions for Form W-7: Application for IRS Individual Taxpayer Identification Number. Rev. Dec. 2024, Department of the Treasury, irs.gov/pub/irs-pdf/iw7.pdf.

Larsson, Karl Kristian. "Digitization or Equality: When Government Automation Covers Some, but Not All Citizens." Government Information Quarterly, vol. 38, no. 1, 2021, article 101547. https://doi.org/10.1016/j.giq.2020.101547.

Lee, Jennifer J. "Operationalizing Language Access Rights for Limited English Proficient Taxpayers." Temple Law Review, vol. 91, no. 4, 2019, pp. 791–811. templelawreview.org/lawreview/assets/uploads/2019/10/Lee_91-Temp.-L.-Rev.-791.pdf.

Lipman, Francine J. "The 'Illegal' Tax." Connecticut Public Interest Law Journal, vol. 11, no. 1, 2012, pp. 94–123. cpilj.law.uconn.edu/wp-content/uploads/sites/2515/2018/10/11.1-The-ILLEGAL-Tax-by-Francine-J.-Lipman.pdf.

Manandise, Esmé. "Towards Unlocking the Narrative of the United States Income Tax Forms." Proceedings of the Second Financial Narrative Processing Workshop, Linköping University Electronic Press, 2019, pp. 33–41. aclanthology.org/W19-6405.pdf.

McClain, Colleen, and William Bishop. "What We Know about Internet Use, Smartphone Ownership and Digital Divides in the U.S." Pew Research Center, 8 Jan. 2026, pewresearch.org/short-reads/2026/01/08/internet-use-smartphone-ownership-digital-divides-in-u-s/.

Moynihan, Donald P., et al. "Administrative Burden: Learning, Psychological, and Compliance Costs in Citizen-State Interactions." Journal of Public Administration Research and Theory, vol. 25, no. 1, 2015, pp. 43–69. https://doi.org/10.1093/jopart/muu009.

National Taxpayer Advocate. "IRS Processing of Individual Taxpayer Identification Numbers." 2024 Annual Report to Congress, Taxpayer Advocate Service, Dec. 2024, pp. 224–40, taxpayeradvocate.irs.gov/wp-content/uploads/2024/12/ARC24_RR_Research_3.pdf.

National Taxpayer Advocate. "Most Serious Problem #7 — Individual Taxpayer Identification Number Processing: IRS Dependence on Paper Forms and Manual Document Review Is Causing Delays, Mistakes, and Potential Security Risks." 2024 Annual Report to Congress, Taxpayer Advocate Service, Dec. 2024, pp. 88–103, taxpayeradvocate.irs.gov/wp-content/uploads/2024/12/ARC24_MSP_07_ITIN-Processing.pdf.

National Taxpayer Advocate. "MSP #9 — The IRS's Handling of ITIN Applications Imposes an Onerous Burden on ITIN Applicants, Discourages Compliance, and Negatively Affects the IRS's Ability to Detect and Deter Fraud." 2012 Annual Report to Congress, Taxpayer Advocate Service, 2012, pp. 154–79, taxpayeradvocate.irs.gov/wp-content/uploads/2020/08/Most-Serious-Problems-ITIN-Individual-Taxpayer-Identification-Number.pdf.

National Taxpayer Advocate. "Multilingual Notices: The IRS Undermines Taxpayer Rights When It Does Not Provide Notices in Foreign Languages." 2019 Annual Report to Congress, vol. 1, Taxpayer Advocate Service, 2019, pp. 69–79, taxpayeradvocate.irs.gov/wp-content/uploads/2020/08/ARC19_Volume1_MSP_08_MULTILINGUALNOTICES.pdf.

National Taxpayer Advocate. "Tax and Financial Literacy." 2024 Annual Report to Congress, Taxpayer Advocate Service, Dec. 2024, pp. 104–17, taxpayeradvocate.irs.gov/wp-content/uploads/2024/12/ARC24_MSP_08_Literacy.pdf.

Ono, Hiroshi, and Madeline Zavodny. "Immigrants, English Ability and the Digital Divide." Social Forces, vol. 86, no. 4, 2008, pp. 1455–79. https://doi.org/10.1353/sof.0.0052.

Suro, Roberto, and Hannah Findling. "Tax Equality for Immigrants: The Indispensable Ingredient for Remedying Child Poverty in the United States." Journal on Migration and Human Security, vol. 9, no. 4, 2021, pp. 251–70. https://doi.org/10.1177/23315024211034836.

Thomson, Dana, et al. "Earned Income Tax Credit Receipt by Hispanic Families with Children: State Outreach and Demographic Factors." Health Affairs, vol. 41, no. 12, 2022, pp. 1725–34. https://doi.org/10.1377/hlthaff.2022.00725.

Towe, Christopher. "How Does the U.S. Tax System Stack Up?" Econofact, 3 Mar. 2026, econofact.org/how-does-the-u-s-tax-system-stack-up.

Treasury Inspector General for Tax Administration. Administration of the Individual Taxpayer Identification Number Program. Report no. 2022-40-013, 26 Jan. 2022, oversight.gov/sites/default/files/documents/reports/2022-02/202240013fr.pdf.

Van Deursen, Alexander J. A. M., and Jan A. G. M. van Dijk. "Internet Skills and the Digital Divide." New Media & Society, vol. 13, no. 6, 2011, pp. 893–911. https://doi.org/10.1177/1461444810386774.

Vimo, Jackie. "The History and Future of ITINs and Taxigration Policy for Undocumented Taxpayers in the United States." Pittsburgh Tax Review, vol. 21, no. 2, 2024, pp. 169–85. https://doi.org/10.5195/taxreview.2024.234.

"What Are Prepopulated Tax Returns?" Tax Policy Center Briefing Book, Tax Policy Center, Urban Institute and Brookings Institution, Jan. 2024, taxpolicycenter.org/briefing-book/what-are-prepopulated-tax-returns.

Woellner, Robin, et al. "Can Simplified Legal Drafting Reduce the Psychological Costs of Tax Compliance? An Australian Perspective." British Tax Review, no. 6, 2007, pp. 717–34.

AI-Driven Recommendations and Consumer Financial Decision-Making: A Literature Review and Survey-Based Investigation By Aarna Bafna

Abstract

Artificial intelligence is now used in many areas of personal finance, and it may affect how consumers make financial decisions. This study pairs a narrative literature review with a survey of 67 respondents to examine how AI relates to financial decisions, purchasing intentions and attitudes towards automated recommendations. In the survey, willingness to purchase AI recommended products was associated with how often respondents used AI, how highly they rated its recommendations and their comfort with AI collecting their personal data. The survey recorded stated intentions from a convenience sample, so these associations do not show causation or changes in actual behaviour. The literature also raises privacy and bias concerns and warns against overreliance on automated advice. AI may help consumers make better-informed financial decisions, provided it is used alongside their own judgement.

Introduction

The transition towards cashless societies has altered how consumers manage, save, and spend their money. As digital payments, mobile banking, and online financial systems generate larger troves of consumer data, AI begins to play an increasingly important role in analysing consumer behaviour and providing personalised guidance. Tools for budgeting, expense tracking, saving, investing, and goal-setting have consequently become more accessible.

At the same time, the use of these technologies raises important concerns, including privacy risks, algorithmic bias, limited transparency, and the possibility of consumer overreliance on automated advice. This study therefore examines the potential role of AI-driven recommendations in budgeting, saving and spending, alongside the factors associated with consumers' willingness to act on these recommendations. Figure 1 illustrates the rapid expansion of generative AI within the financial sector.

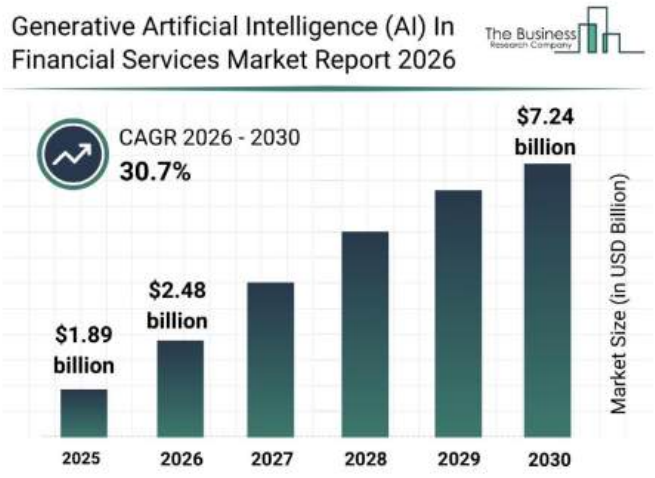


Figure 1. Projected growth of generative AI in financial services, 2025–2030. Adapted from The Business Research Company.

The market is projected to grow from $1.89 billion in 2025 to $7.24 billion by 2030, representing a compound annual growth rate of 30.7%. This rate of growth suggests that AI is moving beyond a specialised or experimental role and becoming increasingly embedded in financial-service operations. This projected growth suggests increasing commercial interest in several applications, including personalised financial recommendations, fraud detection, risk assessment, customer service, and investment analysis.

More broadly, this trajectory makes the present inquiry particularly relevant because AI-powered tools are shaping consumers' financial decisions as their role in the sector continues to expand. Beyond its practical uses, AI may shift personal finance from a reactive process to a more anticipatory one. By identifying financial risks and behavioural patterns before they intensify, AI-based tools can help consumers anticipate cash-flow shortages, identify patterns of impulsive spending, and adjust their financial strategies while there is still time to act.

These capabilities suggest that AI may improve how consumers monitor and plan their finances, although evidence of lasting improvements in financial outcomes remains limited. The study draws on two sources of evidence. The first is a narrative review of the academic literature on AI-driven financial guidance, which sets out the benefits and risks that researchers have identified for personal financial management. The second is a questionnaire completed by 67 respondents, which records their attitudes towards automated recommendations and their stated willingness to buy AI-recommended products. The literature gives the background against which the survey results are read. The survey shows how purchasing intentions vary with AI usage, views on recommendation quality and comfort with personal data collection. The results are then compared with the literature to see what they suggest about personal financial decisions. The survey recorded intentions from a convenience sample and not actual behaviour, so it is used to describe patterns and not to test causes.

The survey data were collected during an earlier research project at PricewaterhouseCoopers (PwC) and reanalysed here. The central research question is: how do consumers' AI usage, their views on the quality of AI recommendations and their attitudes to personal data collection relate to their willingness to buy AI-recommended products, and what does this suggest for personal financial decision-making?

Research Methodology

Narrative Literature review

This narrative literature review draws on peer-reviewed journal articles, conference papers, working papers, institutional reports, and preprints published primarily between 2018 and 2026. Sources were identified through Google Scholar, publisher databases, SSRN, arXiv, and institutional repositories using combinations of the terms "artificial intelligence," "personal finance," "budgeting," "saving," "spending," "robo-advice," "financial literacy," "algorithmic bias," and "explainable AI." Sources were included when they directly examined consumer financial decision-making, evaluated a relevant AI or digital intervention, or addressed privacy,

fairness, transparency, trust, or autonomy in financial services. Because evidence specific to generative AI remains limited, closely related studies of robo-advice, digital financial feedback, and payment-interface design were included where their relevance and limitations could be stated explicitly. The review is qualitative rather than systematic and therefore does not claim exhaustive coverage or permit meta-analytic conclusions.

Primary Research Survey and reanalysis

Alongside the literature review, the study uses primary data from an earlier research project I carried out at PricewaterhouseCoopers (PwC). The data came from a structured questionnaire completed by 67 respondents. I recruited them through personal contacts and friends, so the sample is a non-probability convenience sample. Participation was voluntary and respondents consented to their data being collected. Responses were anonymous.

The questionnaire asked how often respondents use AI, how they judge the quality of AI-generated recommendations, how comfortable they are with AI collecting personal data and how willing they are to buy AI-recommended products. I analysed the responses with descriptive statistics and cross-tabulations. I used spreadsheet software and AI-assisted tools to organise the survey responses, run the statistical calculations and chart the cross-tabulated results. I chose which variables to compare and interpreted the findings myself.

Frequency of AI usage (Q4), comfort with data collection (Q12) and perceived recommendation quality (Q13) were each compared with willingness to buy AI-recommended products (Q9). I then compared the results with the academic literature. Q10, on preferred AI-powered shopping features, was also examined descriptively. Of the 67 respondents, 65 answered Q10 and 66 answered Q13. To compare purchasing intentions across groups, I combined some response categories and calculated within-group percentages.

A PwC professional guided the original research, helped me develop the questionnaire and gave feedback during the project. I carried out the data collection, analysis and report on my own. I ran the survey myself, chose the variables for comparison, directed the research and interpreted the findings. The PwC professional gave guidance and feedback during the original project.

The sample is small and was not selected at random, and the survey recorded self-reported intentions. The results therefore cannot show that any of these attitudes cause purchasing decisions, and they cannot show changes in actual financial behaviour.

AI-Driven Technologies in Personal Financial Management

The integration of AI into personal finance is redefining how consumers make financial decisions. The growing use of AI-powered budgeting and expense-tracking tools, automated savings recommendations and goal-setting systems, and personalised spending guidance is reshaping how individuals manage and allocate their money. Manual expense tracking and budgeting can be time-consuming and prone to error.

AI can automate the collection and organisation of financial data, thereby reducing the effort required for manual record-keeping. In addition, by analysing historical expenditure, transaction amounts, purchase categories, and transaction frequency, AI systems can identify spending patterns and anticipate future expenses and budgetary needs for specific periods ("AI-Powered Personal Finance Management Applications"). A 50-person study conducted by Aishwarya and Hemalatha concluded that smart expense tracking using machine learning facilitated more effective financial management (Aishwarya and Hemalatha). Beyond budgeting and expense tracking, AI-based saving recommendations and automated goal-setting systems can use reinforcement learning to process large and complex data sets and adapt to changing financial conditions. This enables a more personalised approach to financial goal-setting than uniform saving recommendations (Avacharmal et al.). Vanguard, for example, uses reinforcement learning to identify optimal saving rates across multiple goals and income sources (Mohammed et al.). This framework takes into account a consumer's assets, income, cash flow, liabilities, risk tolerance, and investment allocations, thereby identifying saving and investment decisions that maximise the probability of achieving multiple financial goals.

AI-enabled personalisation of financial decisions also uses pattern recognition and multi-model integration to facilitate the exchange and analysis of consumers' personalised financial data (Shuryhin and Zinovatna). Predictive and behavioural analysis of large volumes of data can reveal hidden patterns and support more accurate recommendations for budgeting, spending, and financial planning. Furthermore, combining models such as Isolation Forest, ARIMA, LSTM models, and large language models (LLMs) can provide a more comprehensive analysis of a consumer's financial behaviour and potential anomalies. Such systems may reduce financial ambiguity and irrational behaviour by identifying and addressing cognitive biases (Nahidi), thereby improving users' decision-making ability.

Empirical evidence further illustrates these capabilities. MIT researchers asked 1,000 participants to write prompts about their financial situations, saving, spending, and investing, and then applied advice generated by ChatGPT and Gemini to simulated financial lives from ages 22 to 90 (Choukhmane et al.). The study found that the models encouraged higher savings during consumers' working years and recommended the strategic drawdown of savings during retirement. The models also helped most simulated individuals over age 30 establish substantial savings buffers and aligned saving and spending behaviour more closely with standard life-cycle economic theories.

Notably, saving was introduced in 76% of responses even though only 20% of users mentioned it, while liquidity considerations appeared in 83% of responses despite being raised by only 6% of users. In numerous simulations, retirement wealth exceeded $1 million, indicating substantial wealth-accumulation potential under the modelled scenarios. However, these outcomes depended significantly on users' financial literacy and the quality of the prompts provided.

The Impact of AI on Personal Financial Decisions

The effect of AI on personal financial decisions is complex. As discussed previously, LLMs can support automatic categorisation of expenses, personalised budgeting, and predictive warnings. They may also democratise access to financial guidance by providing structured, on-demand advice to consumers who cannot afford professional financial advisers.

Pak's survey of 2,170 adults found that 67.8% had used an LLM for at least one financial task, including 47.6% for budgeting and expense tracking. Importantly, adoption did not vary substantially by income or wealth, suggesting that AI-based tools can reach consumers across different financial groups (Pak). AI can also support cash-flow management. Machine-learning models can forecast cash flows using historical transactions, receivables, and recurring payments (Małkus et al.). This can allow earlier identification of shortages or liquidity gaps and help businesses respond by arranging resources or postponing non-essential expenditure (Salas-Molina and Rodríguez-Aguilar). However, these benefits must be considered alongside important limitations, including algorithmic bias, an incomplete understanding of contextual information, and concerns relating to privacy and cybersecurity.

In the context of saving, AI can analyse a consumer's income, recurring expenditure, existing savings, and financial objectives to determine a saving rate tailored to the individual's circumstances rather than relying on generic advice (Bauman et al.). For example, if a consumer has only a broad goal such as "save for university," AI can translate that objective into a specific monthly savings target by working backwards from the deadline, the amount required, expected returns, and the consumer's available monthly surplus (Choi et al.). Because static financial plans can quickly become outdated when income or expenditure changes, AI can continually adjust contribution levels, revise deadlines, or reallocate funds among competing goals. In this way, AI may support more efficient financial goal-setting through contextually relevant saving recommendations. AI's technical capacity to generate financial recommendations, together with its increasing adoption among consumers, is reasonably well documented.

However, whether these recommendations produce measurable and lasting increases in saving behaviour remains unclear. Establishing this would require further longitudinal and experimental research that tracks actual saving outcomes rather than relying primarily on stated intentions or short-term usage data. Beyond saving and budgeting, AI may also influence spending behaviour and impulse control by making previously unnoticed spending patterns more visible, introducing friction into purchasing decisions, and providing personalised budgeting recommendations.

Consumers often underestimate frequent, low-value purchases because each transaction appears small in isolation. AI systems can aggregate these purchases and display their cumulative monthly or annual cost, thereby increasing the psychological visibility of everyday spending. Field et al.'s "Banking on Transparency for the Poor: Experimental Evidence from India" examined automated voice notifications sent to women in India who had recently opened bank accounts.

Participants' awareness of their account balances and trust in banking agents improved, but these effects weakened after the notifications stopped (Field et al.). This suggests that

continuous financial feedback may raise awareness in the short term without necessarily producing lasting behavioural change. Impulse purchases are also easier to make when payment is immediate and requires little effort. AI tools may counter this by introducing friction into the process (Dev et al.). For instance, they can require confirmation before a non-essential purchase is completed or temporarily restrict spending once a predetermined limit is reached.

The literature therefore presents AI as a dual-use technology: its impact depends partly on whether a system is designed to protect consumers' financial goals or to maximise engagement and sales. Most available causal evidence concerns digital notifications, disclosures, and payment design generally rather than AI-driven interventions specifically. Consequently, it remains uncertain whether AI can produce sustained reductions in impulsive spending rather than short-term effects.

AI may also affect financial literacy and consumer trust. Generative AI can simplify concepts such as compound interest, credit scores, diversification, and inflation, while adapting its explanations in response to follow-up questions. Arndt's study of ChatGPT as a financial-learning tool surveyed 102 participants before and after use.

The results showed a positive shift in perceptions and suggested potential for AI as a personalised learning aid. However, two important limitations emerged: the accuracy of the information provided and users' ability to formulate effective prompts (Arndt). A deeper concern is that AI can provide conclusions without requiring users to understand the reasoning behind them, meaning that confidence may increase more quickly than actual financial literacy. Trust appears to follow a similar pattern.

Consumers tend to trust financial AI more when its recommendations seem accurate and relevant to their circumstances, and this trust may build through repeated use even in the absence of a personal relationship with the tool. Financial decisions are not purely rational, however; they can carry significant emotional weight, and many consumers continue to believe that a human adviser can understand their circumstances more fully than an algorithm (Northey et al.).

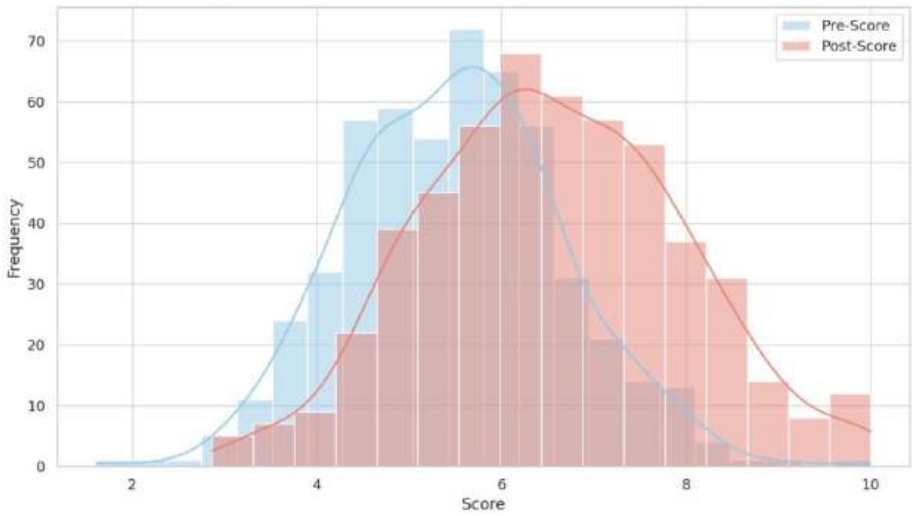


Figure 2. Distribution of financial-literacy scores before and after AI-tool use. Adapted from Kulshrestha et al.

Figure 2 shows a rightward shift in financial-literacy scores following engagement with AI-based financial tools. Kulshrestha et al. found that mean scores rose from 5.5 before AI use to 6.6 afterwards, a statistically significant increase ($p < 0.01$). Post-use scores were also more concentrated towards the higher end of the scale, indicating that AI tools may support users' understanding of financial concepts by providing guidance that is accessible, personalised, and immediate.

These findings offer some empirical support for the argument that AI can complement conventional financial education and narrow existing gaps in financial literacy. A pre-post comparison alone, however, cannot establish that AI caused this improvement, particularly in the absence of a non-AI control group. Further research over a longer time horizon is needed to determine whether the knowledge gained is retained and subsequently translates into changes in financial behaviour.

Primary Survey Findings and Comparative Analysis

The survey asked about consumers' attitudes to AI, mainly their purchasing intentions, their views of AI-generated recommendations and their concerns about personal data collection. Five questions were chosen for analysis because they bear on AI-driven financial decision-making: Q4 (frequency of AI usage), Q9, Q10, Q12 and Q13. Q9, Q10, Q12 and Q13 were analysed descriptively, and Q4 was added in the cross-tabulations to examine how usage relates to purchasing intentions.

1. ***Willingness to purchase AI-recommended products (Q9)***

Q9 asked how likely respondents were to buy a product recommended by an AI system. The largest group, 41.8%, was neutral. A further 34.3% said they were likely to buy and 7.5% said very likely, while 13.4% said they were unlikely, and finally approximately 3.0% selected very unlikely.

Responses were mixed. Many respondents were open to considering an AI-recommended product, but the large neutral group suggests that a good number are still unsure about relying on AI when they buy. This raises the question of whether familiarity with Al necessarily translates into confidence in its recommendations.

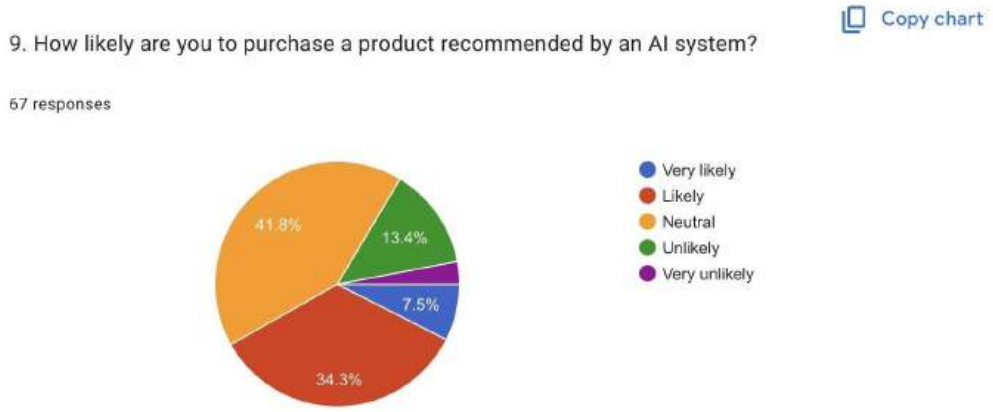


Figure 3. Willingness to purchase AI-recommended products (Q9)

2. *Preferences for AI-powered shopping features (Q10)*

Q10 asked which AI shopping features respondents preferred. Price comparison tools were chosen most often (55.4%), followed by product review summaries (46.2%) and personalised recommendations (43.1%). Virtual shopping assistants (23.1%) and automated customer support (13.8%) drew less interest.

Respondents favoured tools that give them information to work with. Price comparison and review summaries let a consumer weigh alternatives before buying, whereas personalised recommendations put forward options the system has chosen. The difference matters for the question of whether AI mainly helps people make informed financial choices or steers them towards particular products.

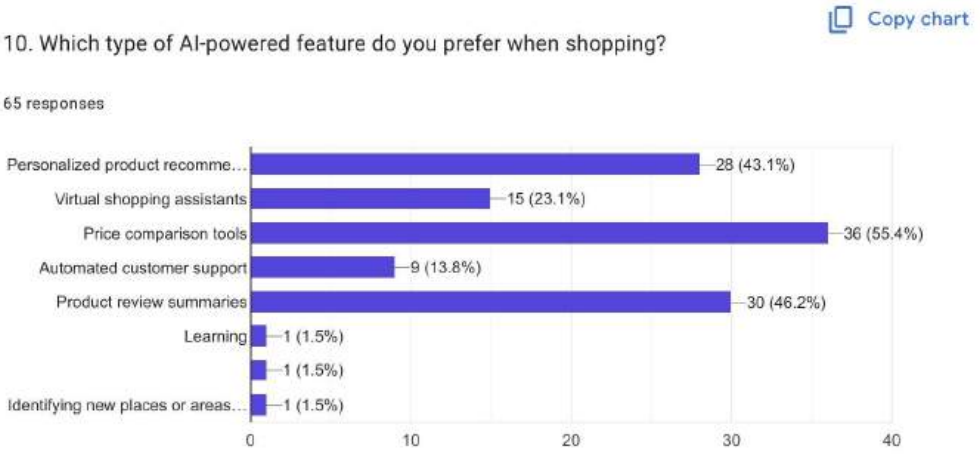


Figure 4. Preferred AI-powered shopping features (Q10)

3. *Attitudes towards data collection and privacy (Q12)*

Q12 asked how comfortable respondents were with AI collecting personal data for personalised services. The largest group, 34.3%, was neutral. Another 28.4% were comfortable and 10.4% very comfortable, while 17.9% were uncomfortable and 9.0% very uncomfortable.

Attitudes differed. Some respondents seemed open to personalisation based on their data, and others had reservations about sharing information with AI. This matters for AI-powered financial applications, which often need sensitive financial data to personalise their advice. The question, however, covered AI data collection in general and did not ask about sharing financial data specifically.

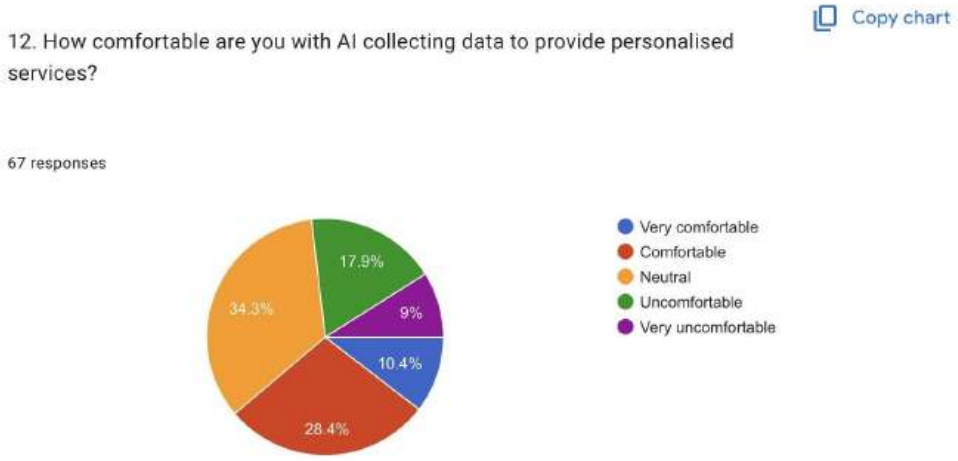


Figure 5. Comfort with AI personal data collection (Q12)

4. AI recommendations compared with human recommendations (Q13)

Q13 asked whether respondents believed AI provided better recommendations than humans in most situations. Of the sample, 39.4% agreed and 12.1% strongly agreed, 33.3% were neutral and 15.2% disagreed.

Respondents generally saw AI recommendations as useful, but a third had not decided whether they were better than human ones. In personal financial management, automated guidance can offer efficiency and easier access, but it cannot always take account of the context that a human adviser would understand.

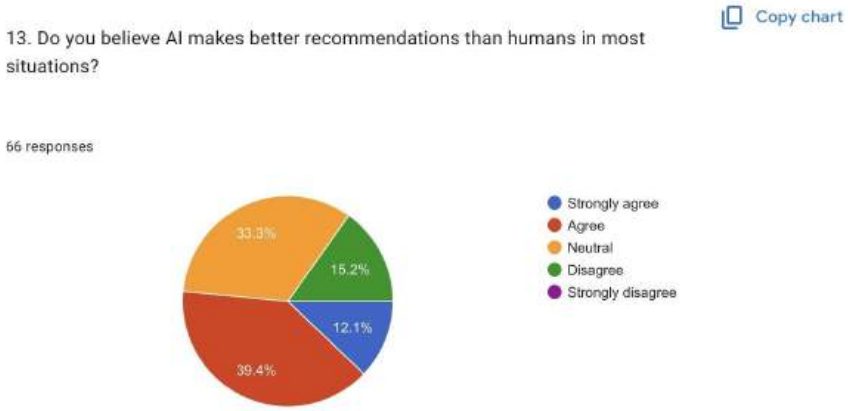


Figure 6. Perceived superiority of AI recommendations over human recommendations (Q13)

Further statistical and comparative analysis

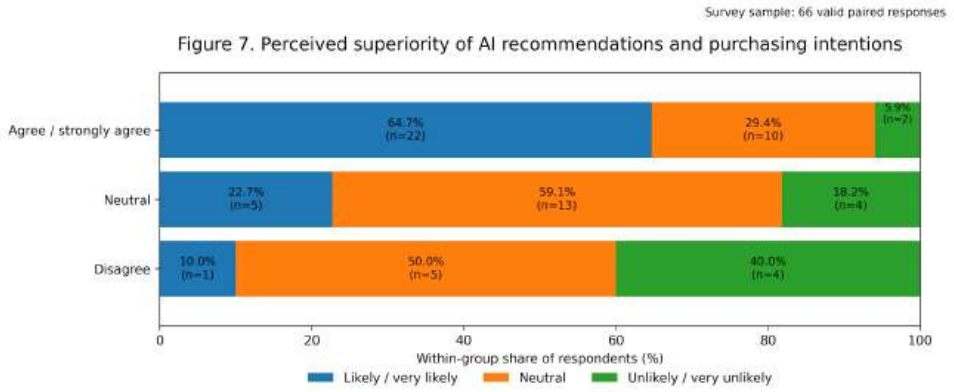
I cross-tabulated three variables against stated willingness to buy AI-recommended products (Q9): perceived quality of AI recommendations (Q13), comfort with personal data collection (Q12) and frequency of AI usage (Q4). The tables give counts and the text gives within-group percentages. Some adjacent categories were combined for comparison, for example "Likely" and "Very likely" as positive purchasing intentions. The five Q9 response categories are kept separate in the tables.

1. Perceived superiority of AI recommendations and purchasing intentions (Q13 × Q9)

The first cross-tabulation tested whether respondents who thought AI gave better recommendations than humans were also more willing to buy what AI recommended.

Table 1. Perceived superiority of AI recommendations and purchasing intentions

Q13 response	Very likely	Likely	Neutral	Unlikely	Very unlikely	Total
Agree / Strongly agree	5	17	10	1	1	34
Neutral	0	5	13	4	0	22
Disagree	0	1	5	3	1	10
Total	5	23	28	8	2	66



Respondents who rated AI recommendations more highly were more willing to buy. Of the 34 who agreed or strongly agreed that AI recommends better than humans, 22 (64.7%) were likely or very likely to buy an AI-recommended product. The share was 22.7% (5 of 22) among neutral respondents and 10% (1 of 10) among those who disagreed. The gap between the agree and disagree groups was 54.7 percentage points.

A favourable view did not always lead to a purchase intention. Ten of the 34 who agreed remained neutral about buying, and two were unlikely or very unlikely to buy. Price, personal preference, the financial consequences of the purchase and how relevant the recommendation seemed to the individual could each play a part, but the questionnaire did not test any of them.

This relates to the literature on consumer trust in automated financial advice. Consumers may recognise what AI recommendations can do and still be reluctant to give up control of decisions involving their money. Q13 asked whether respondents considered AI recommendations better than human ones. It did not ask whether they trusted AI, so the result should not be read as a measure of trust.

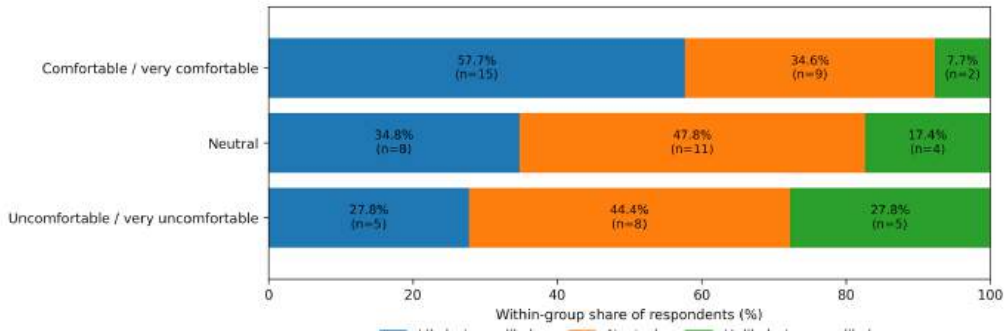
2. Comfort with personal data collection and purchasing intentions (Q12 × Q9)

The second cross-tabulation tested whether comfort with AI collecting personal data was associated with willingness to buy AI-recommended products.

Table 2. Comfort with AI data collection and purchasing intentions

Q12 response	Very likely	Likely	Neutral	Unlikely	Very unlikely	Total
Comfortable / Very comfortable	4	11	9	1	1	26
Neutral	1	7	11	4	0	23
Uncomfortable / Very uncomfortable	0	5	8	4	1	18
Total	5	23	28	9	2	67

Figure 8. Data-sharing comfort and purchasing intentions



Of the respondents who were comfortable or very comfortable with data collection, 57.7% (15 of 26) had positive purchasing intentions. Among those who were uncomfortable or very uncomfortable, the share was 27.8% (5 of 18). The neutral group fell between the two at 34.8% (8 of 23). Respondents who accept personalisation may be more receptive to recommendations built on their interests and preferences. The difference between the comfortable and uncomfortable groups was 29.9 percentage points, smaller than the 54.7 points found for perceived recommendation quality.

Five respondents who were uncomfortable with data collection were nevertheless likely to buy an AI-recommended product, so concern about data and willingness to consider AI suggestions can coexist. A consumer might hesitate to share personal information and still find a

recommendation useful when comparing products. The survey cannot show whether convenience, perceived accuracy, price or some other factor explains these cases.
The result matters for AI-driven personal financial management, where personalised budgeting and saving advice may need sensitive financial information. Q12 asked about AI data collection in general and not about sharing banking or financial records. The results therefore describe attitudes to AI personalisation and cannot show how consumers would respond to a request for financial data.

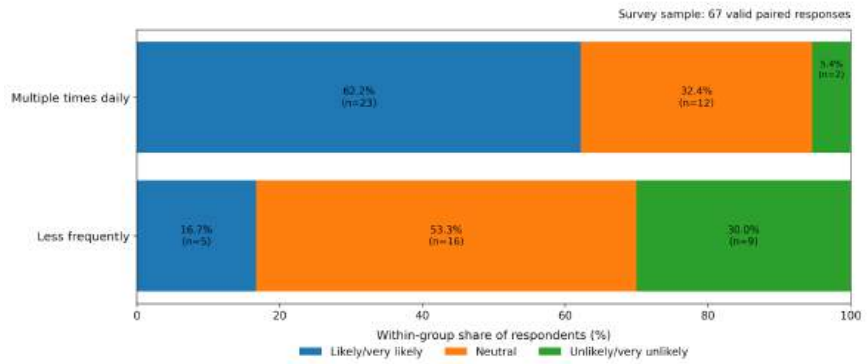
3. Frequency of AI usage and purchasing intentions (Q4 × Q9)

The third cross-tabulation tested whether frequent users of AI tools were more willing than less frequent users to buy what AI recommended. Respondents who used AI multiple times a day were placed in one group, and those who used it once a day, a few times a week, rarely or never were placed in the other.

Table 3. AI usage frequency and purchasing intentions

Q4 response	Very likely	Likely	Neutral	Unlikely	Very unlikely	Total
Multiple times a day	5	18	12	1	1	37
Less frequently	0	5	16	8	1	30
Total	5	23	28	9	2	67

Figure 9. AI-use frequency and purchasing intention



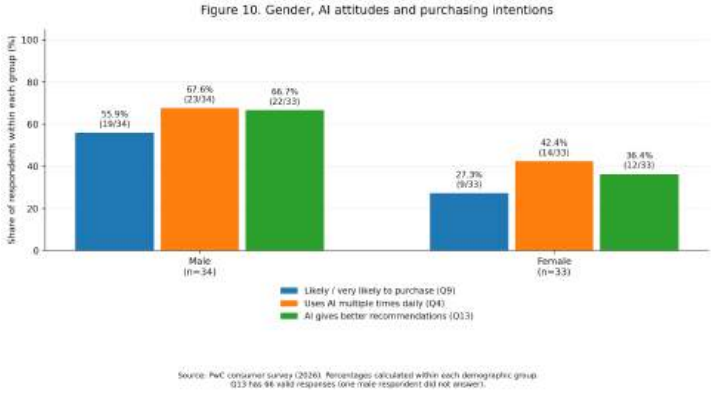
Purchasing intentions differed considerably by usage frequency. Of the 37 respondents who used AI multiple times a day, 23 (62.2%) were likely or very likely to buy AI-recommended products, compared with 5 of 30 (16.7%) among less frequent users. The gap was 45.5 percentage points. Negative intentions followed the same pattern: 5.4% of frequent users chose "Unlikely" or "Very unlikely", against 30% of less frequent users.
Repeated use may make automated suggestions familiar and so make consumers readier to use them when assessing a purchase. The survey cannot establish the direction of the relationship. People who are already enthusiastic about AI may be more likely both to use it often and to accept its recommendations.

Frequent use did not guarantee willingness to buy. Twelve frequent users were neutral and two were unlikely or very unlikely to buy. Familiarity with the technology and reliance on its recommendations are separate matters. A person may use AI heavily for study or work and still not rely on it to guide purchases.

Demographic differences in AI usage and purchasing intentions

To check whether responses differed by demographic characteristics, the survey data were analysed by gender, age and employment status. Demographic groups were compared on purchasing intentions (Q9) and frequency of AI usage (Q4). Perceived quality of AI recommendations (Q13) was also examined by gender. The results add to the cross-tabulations reported above. The small sample and the recruitment method limit what can be concluded from them.

1. Gender and purchasing intentions



The sample had 34 male and 33 female respondents. Their willingness to buy AI-recommended products differed. Among men, 19 of 34 (55.9%) were likely or very likely to buy one, compared with 9 of 33 women (27.3%). Neutral answers were more common among women (51.5%) than men (32.4%). Within this sample, men reported greater willingness to buy.

Usage frequency differed in the same direction. Of the men, 67.6% used AI multiple times a day, against 42.4% of the women. Differences in familiarity may therefore account for part of the gap in purchasing intentions.

The pattern held within each gender. Among women who used AI multiple times a day, 8 of 14 (57.1%) had positive purchasing intentions, against 1 of 19 (5.3%) among less frequent users. Among men, the figures were 65.2% for daily users and 36.4% for less frequent users. Usage frequency was associated with purchasing intentions in both groups. It cannot be the only factor, since other characteristics may affect both how often someone uses AI and whether they would buy what it recommends.

Perceived recommendation quality also differed. Among the 33 male respondents who answered Q13, 22 (66.7%) agreed or strongly agreed that AI provides better recommendations than humans, compared with 12 of the 33 female respondents (36.4%). This may help explain the

difference in purchasing intentions, but the survey cannot show that it does. The gender analysis shows differences in usage, perceptions and purchasing intentions within the sample. It gives no basis for saying that gender determines trust or behaviour.

2. Age and purchasing intentions

Differences by age were smaller and less consistent. Respondents under 25 had the highest share with positive purchasing intentions (53.8%), followed by those aged 25–35 (42.9%), 36–55 (38.5%) and over 55 (37.5%). The ordering points to slightly more receptiveness among younger respondents, but it did not carry through to AI usage. The 25–35 group had the highest share using AI multiple times a day (85.7%) without having the highest share of positive purchasing intentions. Frequent use of AI therefore did not always go with equal willingness to buy across age groups.

Group sizes limit interpretation. The 36–55 group had 39 respondents and the 25–35 group only seven, so one person's answer moves the percentages in the smaller groups considerably. All respondents under 25 were students, which means age and student status coincide in that group and their effects cannot be separated. The data indicate possible differences between age groups. They do not support a conclusion that younger consumers are generally more willing to act on AI purchasing recommendations.

3. Employment status and purchasing intentions

Students had the highest share with positive purchasing intentions (53.8%), followed by employed respondents (45.5%). Only one of the six homemakers and none of the four retired respondents reported positive intentions. Usage followed a similar order. Of the employed respondents, 63.6% used AI multiple times a day, as did 61.5% of the students, but only one homemaker and no retired respondent did. In this sample, students and employed respondents used AI more and were more receptive to its purchase recommendations than the homemaker and retired groups.

These figures need careful reading. With six homemakers and four retired respondents, a single changed answer would shift the percentages sharply. Employment status also overlaps with age and with characteristics the survey did not measure, such as income and financial literacy, and the analysis did not control for them. Employment status may therefore be relevant, but the data do not establish it as an explanation for differences in purchasing intentions.

Demographic patterns and AI in financial decision-making

The demographic results add to the main cross-tabulations. Frequent AI use was associated with positive purchasing intentions in both gender groups. Men and women also differed in how they rated AI recommendations, which suggests that attitudes toward AI vary between demographic groups. The data cannot show whether this difference is linked to gender itself or to other factors, such as how often respondents use AI.

Frequent use of AI and confidence in acting on its recommendations need to be kept apart. Heavy users may be more receptive to automated suggestions, but that does not mean they follow every recommendation or make better financial decisions. For personal financial management the distinction matters. AI budgeting and saving tools may give useful personalised guidance, while shopping recommendation systems may lead consumers to spend more. The questionnaire did not record actual purchases, spending or changes in saving, so the demographic differences in purchasing intentions say nothing about whether financial outcomes improved or worsened.

The differences reported here are descriptive and may reflect overlapping influences. They should not be read as independent demographic effects. A larger and more representative sample would be needed to test whether they persist when age, gender, employment status, financial literacy and prior experience with AI are considered together.

Risks and Ethical Considerations

Overall, AI has the potential to improve financial decision-making, but its adoption warrants caution because of important ethical, practical, and behavioural limitations. One concern relates to privacy and data security. Personal-finance applications typically connect to bank accounts through data aggregators such as Plaid, Yodlee, MX, and Finicity rather than linking directly to the bank itself.

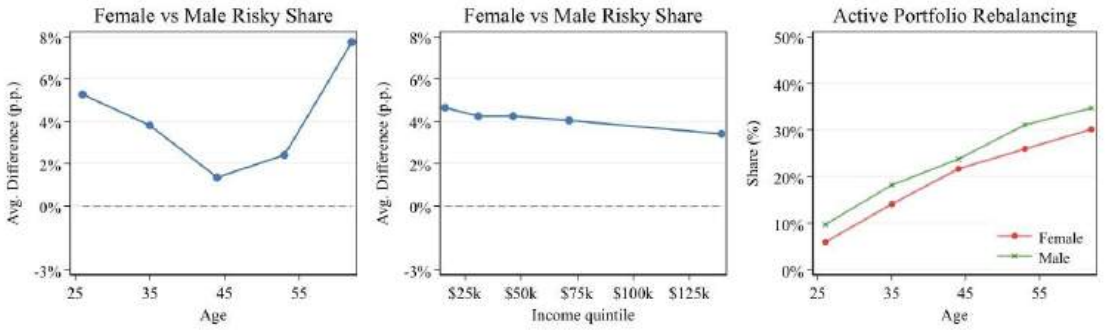
U.S. consumers increasingly rely on fintech applications to manage budgets, payments, and loans, and this functionality depends on open-banking systems in which data aggregators act as intermediaries, a role that remains imperfectly understood (Alcazar and Hayashi). When AI models process financial data directly, additional attack vectors may emerge. Researchers developing financial language models identify membership-inference and model-inversion attacks as genuine risks to private financial data and note that differential privacy for on-device financial LLMs remains underexplored (Zhu et al.). Beyond privacy concerns, algorithmic bias presents another significant risk. AI systems may reproduce discrimination already embedded in financial data.

Because algorithms learn from historical lending, investing, and spending patterns, data that reflects unequal treatment may cause AI to perpetuate rather than correct those inequalities. Morse and Pence identify several points at which bias can enter financial algorithms, including training data, model design, proxy variables, and developer choices. Consumers may consequently receive more expensive loans, lower credit limits, or less suitable products, with those who have lower financial literacy potentially being particularly exposed.

A separate study found that participants with lower financial literacy provided less detailed prompts, resulting in less personalized recommendations; simulations suggested that this could leave such consumers nearly $50,000 worse off by age 60 (Choukhmane et al.). This finding highlights a paradox: consumers who most need financial support may be least equipped to provide the information that AI requires to generate suitable advice. AI may also fail to account adequately for financial vulnerability.

Zeka and Johns tested ChatGPT, Claude, and Perplexity on financial-planning scenarios involving vulnerable consumers and found that vulnerability was handled inconsistently across all three systems: ChatGPT under-addressed it, Perplexity often redirected users without sufficient detail, and Claude required users to complete substantial self-guided planning. Advice generated under these conditions may appear personalised while overlooking debt pressure, unstable income, caregiving responsibilities, or limited financial knowledge. For these reasons, AI-generated financial recommendations should support human judgement rather than replace it entirely.

Providers should regularly test outcomes across different demographic groups, explain the main factors behind recommendations, gather sufficient information from consumers, and ensure that human review is available for vulnerable users or decisions with major financial consequences. Algorithmic bias can also lead AI systems to generate different financial recommendations based on demographic characteristics such as gender. Figure 11 compares equity allocations and portfolio-rebalancing advice provided to male and female users and illustrates a consistent gender gap across these AI-generated recommendations.



Notes: This figure examines the gender gap in LLM-recommended equity shares by age and income. The left panel plots the average difference in recommended risky asset shares (men minus women) by age. The middle panel plots the same difference by income quintile. The right panel plots the share of individuals whose LLM-recommended portfolio change constitutes active rebalancing (rather than passive drift), separately for men and women by age.

Figure 11. Gender gaps in recommended equity shares and portfolio rebalancing. Adapted from Choukhmane et al.

The algorithm recommended higher risky-asset allocations to men than to women across every age and income category, and men were also more likely to receive advice involving active portfolio rebalancing. These findings suggest that AI systems may reproduce gender-based assumptions about risk tolerance, potentially leaving women with advice that is overly conservative and that restricts long-term wealth accumulation. A further concern involves the opacity of AI-generated financial guidance.

Systems may recommend a particular budget, investment, or saving strategy without disclosing the reasoning behind that recommendation. A systematic review covering 138 studies found that such opacity can make financial decisions harder to justify after the fact, thereby undermining both user trust and the ability to assess risk accurately (Černevičienė and Kabašinskas). Accountability becomes further complicated when AI-generated advice causes financial harm. Because responsibility may be distributed across developers, financial

institutions, data providers, and end users, no single party is clearly positioned to bear responsibility for an adverse outcome.

Some scholars have described this as a diffusion of agency, arguing that it obstructs attempts to establish liability when an AI-generated recommendation results in loss (Lioliou et al.). Even when providers attempt to address opacity by offering explanations, the results may remain unsatisfactory. Providing model statistics or confidence scores does not, on its own, create meaningful transparency for non-expert users. A review of 150 peer-reviewed studies found that developers struggle to reconcile comprehensibility with accuracy: explanations simplified enough for consumers to understand may omit technical detail required by regulators, while explanations detailed enough to satisfy regulatory standards are often difficult for consumers to interpret (Khan et al.). Transparency and accountability therefore remain central challenges for AI-driven financial advice.

AI-driven financial recommendations can create a trade-off between efficiency and accountability. Although AI can provide fast, personalized guidance, opaque decision-making may prevent consumers from understanding why a recommendation was made or whether it serves their interests. This becomes especially significant when unsuitable advice results in financial loss because responsibility may be divided among developers, institutions, and users.

Without understandable explanations and appropriate human oversight, trust in these systems risks becoming blind reliance rather than informed confidence. Finally, consumer autonomy is another important ethical consideration. Consumer autonomy, understood here as an individual's capacity to make independent and informed financial decisions without excessive dependence on AI recommendations, is complicated by two related risks.

The first concerns over reliance. Consumers do not always recognise when AI-generated advice is incorrect. Research on what has been termed "appropriate reliance" indicates that individuals often struggle to disregard incorrect AI recommendations, raising the possibility that users may defer to automated output rather than exercising independent judgement.

Sustained use of such tools may also erode confidence in one's own financial competence and reinforce dependence over time (Schemmer et al.). The second risk concerns framing. AI systems do not present information neutrally; the way in which a recommendation is delivered can influence consumer choice. In an experimental study involving 334 participants, approximately 67% accepted robo-advice, with acceptance rates varying according to how performance information was presented and whether outcomes were framed in terms of gains or losses (Mahmud et al.). Under these conditions, consumers may experience their decisions as autonomous even when the framing of the advice has substantially shaped the outcome.

Taken together, the evidence suggests that AI's effect on consumer autonomy is neither straightforwardly positive nor negative. On the one hand, these tools can broaden access to personalised financial information and support more informed decision-making. On the other hand, overreliance and susceptibility to framing indicate that autonomy can be undermined even when consumers retain the formal capacity to choose.

AI is therefore best positioned as a support for financial decision-making rather than as a substitute for it, a distinction that has important implications for both system design and regulation.

Integrated discussion: survey findings and existing literature

The literature review and the survey approach AI in consumer finance from different directions. The literature describes how AI can help with budgeting, saving, financial literacy and personalised guidance. The survey asks how willing consumers are to act on AI-generated purchasing recommendations. A consumer's acceptance of a recommendation says little about whether acting on it improves their finances, and both sources leave that question open.

Perceived recommendation quality and consumer reliance

Perceived recommendation quality produced the widest gap in purchasing intentions of the three variables examined. Of the respondents who agreed or strongly agreed that AI gives better recommendations than humans, 64.7% had positive purchasing intentions. The figure was 22.7% for neutral respondents and 10% for those who disagreed.

Northey et al. studied how AI in banking affects consumers' beliefs about financial advice, and Mahmud et al. show that acceptance of robo-advice can vary with how the information is presented. Both studies fit the survey result: how consumers judge a recommendation appears to bear on whether they act on it when money is involved.

Acceptance of a recommendation does not amount to informed trust. Schemmer et al. describe the difficulty people may have in recognising and rejecting incorrect AI advice. A consumer who rates AI recommendations highly may receive good guidance more readily and may also be more exposed to unsuitable suggestions. The survey cannot distinguish these cases. It recorded perceived quality and stated intentions, and it did not record the accuracy of any recommendation or any actual purchase. Q13 also did not measure trust. The result is an association between favourable evaluations and purchasing intentions. It is not evidence of overreliance.

AI usage frequency and purchasing decisions

Respondents who used AI multiple times a day were far more likely to report positive purchasing intentions than less frequent users: 62.2% against 16.7%. Pak documents the use of generative AI for personal financial tasks such as budgeting and expense tracking. The two studies measure different things. Pak examines financial applications, and the PwC questionnaire asked about AI use in general.

Familiarity may make automated recommendations seem more useful. The causal order could also run the other way, because people who already think well of AI may use it more often and be more willing to follow it. Heavy use does not imply financial reliance either. Twelve frequent users were neutral about buying AI-recommended products, so daily use did not settle the question for them.

Personalisation, privacy and consumer acceptance

Respondents who were comfortable with AI collecting personal data were more willing to buy AI-recommended products than those who were uncomfortable: 57.7% against 27.8% had positive purchasing intentions. People who accept data-driven personalisation may find the resulting recommendations easier to accept.

Personalised financial advice usually requires data on income, spending, assets and financial goals. Such data can make advice fit a person's situation better, and collecting it carries privacy risk. Alcazar and Hayashi discuss the role of data aggregators in financial services, and Zhu et al. identify privacy risks in financial language models.

Concern about privacy and receptiveness to recommendations coexisted in some respondents. Five who were uncomfortable with data collection still had positive purchasing intentions, and the survey did not ask why. Q12 covered AI data collection in general and did not ask about access to financial records, so it gives no direct evidence on how consumers would respond to a request to share banking information.

AI as a financial support tool and as a driver of consumption

The literature and the survey differ in what the recommendations are meant to do. The literature describes AI that monitors spending and suggests how much to save. Field et al. studied automated financial notifications and suggested that digital feedback can improve financial awareness, although the effects weakened after the intervention ended. Work on payment-interface design suggests that added friction can lead consumers to reconsider unnecessary purchases.

The PwC survey instead looks at receptiveness to recommendations that may lead to a purchase. In Q10, 55.4% of respondents chose price comparison tools, 46.2% chose product review summaries and 43.1% chose personalized recommendations. Respondents could choose more than one feature, and the top and bottom of these three shares differ by about 12 percentage points. The data therefore show only slightly more interest in tools that help evaluate a purchase than in recommendations chosen by the system.

An AI budgeting application may help a user find spending to cut, while a shopping recommendation system may lead the same user to consider more products. Which of these happens may depend on how the system is designed and what it is built to do. The survey does not show that AI recommendations increased spending, and the reviewed literature does not show that AI reliably produces lasting improvements in saving.

Demographic differences and alternative explanations

In the survey, 55.9% of men had positive purchasing intentions, compared with 27.3% of women. Men also used AI more often: 67.6% used it multiple times a day, against 42.4% of women. Some of the gaps in purchasing intentions may therefore reflect differences in familiarity with AI.

The literature review separately raises the risk of biased AI financial advice, including the gender differences in investment recommendations that Choukhmane et al. examine. That concerns what AI systems produce. The survey concerns what respondents say they would do, and differences between respondents' attitudes do not demonstrate algorithmic discrimination. The age and employment subgroups are small, and all respondents under 25 were students, so age and student status cannot be separated. These limits restrict what can be said about demographic influences.

Limitations and implications

Each of the three attitudes examined was associated with higher stated willingness to buy AI-recommended products: more frequent use, a higher rating of recommendation quality and greater comfort with data collection. The literature likewise finds that consumers' perceptions and the presentation of automated guidance affect financial decisions. Neither source shows that greater willingness to follow recommendations leads to better financial outcomes.

The survey used a convenience sample of 67 respondents and recorded self-reported intentions. It has no data on actual purchases or on changes in saving and spending. The analysis is descriptive and does not control for age, employment status, financial literacy or prior experience with AI. A larger and more representative survey, combined with experimental or longitudinal evidence on real financial behaviour, would address these gaps.

The survey bears on whether consumers are willing to follow AI recommendations. It says nothing about whether following them improves their finances, and the literature reviewed does not settle that question either. AI may help consumers manage their money when its recommendations are accurate and fit their goals. Consumers who follow recommendations readily are more exposed when those recommendations are poor, so transparency and privacy protection carry more weight. Assessing AI in personal finance therefore requires evidence on the quality of the decisions it leads to, in addition to evidence on acceptance.

Conclusion

AI has potential to improve personal financial management through automation of budgeting and recognition of patterns in spending, and can generate personalized advice for saving and financial planning. Whether it does so depends on the capability of the systems and on the willingness of consumers to act on their recommendations. This study used a narrative literature review combined with survey data from 67 respondents to investigate what is associated with acceptance of AI-generated recommendations and the implications this has for financial decisions.

The survey yielded three main results. Respondents who used AI multiple times a day were more receptive to buying AI-recommended products than infrequent users: 62.2% of frequent users had positive purchasing intentions vs. 16.7% of the latter. Of respondents who thought AI provided better recommendations than humans, 64.7% had positive purchasing intentions vs. 10% of those who disagreed. Respondents comfortable with AI collecting personal

data were also more receptive to purchasing (57.7%) than those who were uncomfortable (27.8%). Respondents who were familiar with AI, who evaluated its recommendations highly, or who accepted data sharing were generally more receptive to automated guidance.

In the context of the literature, the results highlight the distinction between accepting AI recommendations and demonstrating improvements in financial outcomes. While AI may enable consumers to make more informed decisions, a higher willingness to act on its advice does not indicate an improvement in budgeting, saving, or spending behaviours. The literature also highlights unresolved risks in the form of algorithmic bias, limited transparency, privacy violations and overreliance, which are most pertinent in the context of consequential financial decisions.

The study's sample size was small and non-random, and the survey measured self-reported intentions rather than observed behaviours. The associations identified cannot indicate causation and do not apply more broadly to the population. Subsequent research should use larger and more representative samples and employ longitudinal or experimental study designs to establish whether acceptance of AI recommendations actually results in measurable improvements to consumers' financial well-being.

The results suggest the value of AI that supports consumers in their own judgement rather than replacing it, with transparency, data protection, and human oversight maintained. The benefits of AI in financial decision-making depend on the validity of its recommendations and the ability of consumers to recognise this, which willingness to follow its advice does not necessarily indicate.

Works Cited

"AI-Powered Personal Finance Management Applications." IEEE Xplore, ieeexplore.ieee.org/document/11199144/. Accessed 6 Sept. 2026.

Aishwarya, S., and M. Hemalatha. "Smart Expense Tracking System Using Machine Learning." Proceedings of the 15th International Conference on Agents and Artificial Intelligence, 2023, www.scitepress.org/Papers/2023/126139/126139.pdf. Accessed 6 Sept. 2026.

Akhtar, Farida, et al. "Evolution of Robo-Advisors: A Literature Review and Future Research Agenda." International Journal of Consumer Studies, vol. 49, no. 6, 2025, article e70131. doi.org/10.1111/ijcs.70131.

Alcazar, Julian, and Fumiko Hayashi. "Data Aggregators: The Connective Tissue for Open Banking." Payments System Research Briefing, Federal Reserve Bank of Kansas City, 24 Aug. 2022, www.kansascityfed.org/documents/9012/PaymentsSystemResearchBriefing22AlcazarHayashi0824.pdf.

Arndt, Holger. "Economic and Financial Learning with Artificial Intelligence: A Mixed-Methods Study on ChatGPT." arXiv, 23 Feb. 2024, doi.org/10.48550/arXiv.2402.15278.

Avacharmal, Rajiv, et al. "Leveraging Reinforcement Learning for Advanced Financial Planning for Effective Personalization in Economic Forecasting and Savings Strategies." 2024 15th International Conference on Computing Communication and Networking Technologies, IEEE, 2024, pp. 1–4. doi.org/10.1109/ICCCNT61001.2024.10725494.

Bauman, Tessa, et al. "Deep Reinforcement Learning for Robust Goal-Based Wealth Management." arXiv, 25 July 2023, doi.org/10.48550/arXiv.2307.13501.

Ben David, Daniel, et al. "Explainable AI and Adoption of Financial Algorithmic Advisors: An Experimental Study." arXiv, 5 Jan. 2021, revised 9 June 2021, doi.org/10.48550/arXiv.2101.02555.

Choi, James J., et al. Small Cues Change Savings Choices. National Bureau of Economic Research, Feb. 2012. NBER Working Paper no. 17843, doi.org/10.3386/w17843.

Choukhmane, Taha, et al. "AI Financial Advice: Supply, Demand, and Life Cycle Implications." SSRN, 29 July 2026, doi.org/10.2139/ssrn.6446286.

Dev, Harshal, et al. "From Cash to Cashless: UPI's Impact on Spending Behavior Among Indian Users and Prototyping Financially Responsible Interfaces." arXiv, 18 Jan. 2024, doi.org/10.48550/arXiv.2401.09937.

Field, Erica M., et al. Banking on Transparency for the Poor: Experimental Evidence from India. National Bureau of Economic Research, July 2022. NBER Working Paper no. 30289, doi.org/10.3386/w30289.

Khan, Farhina Sardar, et al. "Model-Agnostic Explainable Artificial Intelligence Methods in Finance: A Systematic Review, Recent Developments, Limitations, Challenges and Future Directions." Artificial Intelligence Review, vol. 58, 2025, article 232, doi.org/10.1007/s10462-025-11215-9.

Kulshrestha, Nitin, et al. “AI Based Enhancing Financial Literacy and Personal Investment Decisions.” Advances in Consumer Research, vol. 2, no. 4, 2025, pp. 2214–28.

Lioliou, Eleni, et al. “Governance under Algorithmic Opacity: How Financial Firms Construct Accountability and Control around AI in Risk Disclosures.” Information Systems Frontiers, 17 June 2026, doi.org/10.1007/s10796-026-10762-y.

Mahmud, Hasan, et al. “Human-Robo-Advisor Collaboration in Decision-Making: Evidence from a Multiphase Mixed Methods Experimental Study.” Decision Support Systems, vol. 198, 2025, article 114541, doi.org/10.1016/j.dss.2025.114541.

Małkus, Bartłomiej, et al. “Financial Management System for SMEs: Real-World Deployment of Accounts Receivable and Cash Flow Prediction.” arXiv, 5 Nov. 2025, doi.org/10.48550/arXiv.2511.03631.

Mohammed, Shareefuddin, et al. “Embracing Advanced AI/ML to Help Investors Achieve Success: Vanguard Reinforcement Learning for Financial Goal Planning.” arXiv, Oct. 2021, doi.org/10.48550/arXiv.2110.12003.

Morse, Adair, and Karen Pence. “Technological Innovation and Discrimination in Household Finance.” Finance and Economics Discussion Series, Federal Reserve Board, 2020, doi.org/10.17016/FEDS.2020.018.

Nahidi, Narmin. “AI in Personalized Financial Decision-Making.” AI, FinTech, and the Future of Robo-Advisory: Risk Management and Ethical Considerations, edited by Narmin Nahidi and Alex Zarifis, Springer, 2026, pp. 75–94, doi.org/10.1007/978-3-032-18109-1_4.

Northey, Gavin, et al. “Man vs Machine: How Artificial Intelligence in Banking Influences Consumer Belief in Financial Advice.” International Journal of Bank Marketing, vol. 40, no. 6, 2022, pp. 1182–99, doi.org/10.1108/IJBM-09-2021-0439.

Pak, Tae-Young. “How Individuals Use Generative AI for Personal Financial Management.” Journal of Behavioral and Experimental Finance, vol. 49, 2026, article 101145, doi.org/10.1016/j.jbef.2026.101145.

Salas-Molina, Francisco, and Juan A. Rodríguez-Aguilar. “Data-Driven Multiobjective Decision-Making in Cash Management.” EURO Journal on Decision Processes, vol. 6, 2018, pp. 77–91, doi.org/10.1007/s40070-017-0075-y.

Schemmer, Max, et al. “Should I Follow AI-Based Advice? Measuring Appropriate Reliance in Human-AI Decision-Making.” arXiv, 14 Apr. 2022, doi.org/10.48550/arXiv.2204.06916.

Shuryhin, Kostiantyn A., and Svitlana L. Zinovatna. “Recommendation System for Financial Decision-Making Using Artificial Intelligence.” Applied Aspects of Information Technology, vol. 7, no. 4, Dec. 2024, pp. 348–58, doi.org/10.15276/aait.07.2024.24.

The Business Research Company. Generative Artificial Intelligence in Financial Services Global Market Report 2026. Sept. 2026, www.thebusinessresearchcompany.com/report/generative-artificial-intelligence-ai-in-financial-services-global-market-report. Accessed 6 Sept. 2026.

Zeka, Bomikazi, and Raechel Johns. “What Could Possibly Go Wrong? A Consumer-Centric Approach to Exploring Vulnerability and Algorithmic Bias in AI-Driven Financial Planning.” Financial Planning Review, 2026, article e70040, doi.org/10.1002/cfp2.70040.

Zhu, Sichen, et al. “When FinTech Meets Privacy: Securing Financial LLMs with Differential Private Fine-Tuning.” arXiv, 10 Sept. 2025, doi.org/10.48550/arXiv.2509.08995.

Černevičienė, Jurgita, and Audrius Kabašinskas. “Explainable Artificial Intelligence in Finance: A Systematic Literature Review.” Artificial Intelligence Review, vol. 57, 2024, article 216, doi.org/10.1007/s10462-024-10854-8.

A Comparative Analysis of ATP Masters 1000 Event Marketing in China and the United States By Isaac Zhang

Abstract

Professional tennis tournaments sell more than matches. They also sell a way of experiencing the sport: through luxury, local identity, player narratives, or a sense of community. This paper compares the social-media positioning of the Rolex Shanghai Masters with three ATP Masters 1000 events in the United States: the BNP Paribas Open at Indian Wells, the Miami Open, and the Cincinnati Open. A qualitative comparative case study was combined with a descriptive content analysis of 80 Instagram posts, 20 from each tournament's official account. Posts were coded for prominent players or celebrities and for broader lifestyle, venue, and event-experience content. The cases reveal distinct brand logics. Shanghai relies heavily on premium hospitality, celebrity visibility, and city branding; Indian Wells emphasizes the established identity of "Tennis Paradise"; Cincinnati presents itself as the communal "Heart of Tennis"; and Miami blends elite sport with food, fashion, music, and a festival atmosphere. Interpreted through glocalization, the experience economy, fan identification, and the Psychological Continuum Model, the findings show that international tournaments benefit from adapting a global sports product to local consumer cultures while retaining player-centered stories that deepen attachment. The paper concludes with realistic, measurable recommendations for each event.

Introduction

The ATP Tour is global, but tennis is not marketed in the same way everywhere. A Masters 1000 tournament in California can draw on decades of local habit, family traditions, and familiar rivalries. A tournament in Shanghai may need to introduce the sport to newer audiences while also competing with the city's other premium leisure experiences. The matches follow the same rules, and many of the same players appear, yet the meaning built around the event can be quite different.

The Rolex Shanghai Masters makes this contrast especially useful. Established in 2009, it is the only ATP Masters 1000 event in Asia and has become a major stop on the tour's Asian swing ("ATP Masters 1000 Shanghai"). The United States, by comparison, hosts three Masters 1000 events: Indian Wells, Miami, and Cincinnati. Each has a recognizable identity. Indian Wells calls itself "Tennis Paradise" and promotes both elite competition and a destination experience (*BNP Paribas Open*). Cincinnati's official brand centers on the "Heart of Tennis," presenting the tournament as a shared connection among place, people, and players ("Our Brand"). Miami openly combines world-class tennis with dining, entertainment, and themed programming; its free Opening Day, for example, includes practices, qualifying matches, campus activities, and a culinary passport ("2026 Miami Open").

These differences raise the central question of this study: how do the Shanghai Masters in China and U.S.-based Masters 1000 tournaments use social media and brand positioning to

attract and engage fans? The argument is not that China has one consumer culture and the United States another. Four tournaments cannot support a claim that broad. Rather, the comparison shows how specific events respond to their histories, locations, audiences, and commercial partnerships. Shanghai's content is more strongly organized around luxury, celebrities, and the image of the city, whereas the three U.S. events rely more heavily on player stories, established tournament identities, and fan experience. Across the cases, however, recognizable players and public figures remain powerful points of attention.

Methodology

This study uses a qualitative comparative case-study design supported by a small descriptive content analysis. The cases are the official Instagram accounts of the Rolex Shanghai Masters, Cincinnati Open, Miami Open, and BNP Paribas Open. The dataset contains 20 posts from each account, for a total sample of 80 posts. Each post was coded according to whether a prominent player, celebrity, or other recognizable public figure was central to the content. Posts without such a figure were treated as general tournament content, including venue imagery, hospitality, food, fashion, city branding, and event promotion. The analysis also reviewed official tournament websites and brand pages to identify how each event describes itself.

For each account, the analysis recorded how many posts focused on prominent players or celebrities, the percentage of the 20-post sample represented by those posts, the mean number of likes across the full sample, and the mean for the featured subset. Standard deviations were also calculated because engagement varied considerably from one post to another. These figures are used only to describe patterns within each account. They should not be interpreted as proof that featuring a famous person caused higher engagement, since Instagram distribution, posting time, account size, match results, paid promotion, and the popularity of a particular athlete may all affect the number of likes.

The qualitative stage identified repeated themes in the captions and visuals, including lifestyle positioning, city identity, player access, rivalry and achievement narratives, hospitality, entertainment, and community. These patterns were then interpreted through four ideas from marketing and sport research. Pine and Gilmore's experience-economy framework explains how organizations create value by staging memorable experiences around a core product (98-99). Robertson's concept of glocalization helps explain how a global property is adapted to local culture (28-30). Wann and Branscombe's work on team identification examines the degree to which spectators identify with their team (1-2). Finally, Funk and James's Psychological Continuum Model describes movement from awareness to attraction, attachment, and allegiance (121-22).

Findings

Rolex Shanghai Masters

Content across the Shanghai account emphasized lifestyle positioning through luxury hospitality packages and tournament experiences. City branding also emerged through images of Shanghai's iconic architecture and luxury sponsors, presenting the event as a symbol of prestige. The tournament's website reinforces this approach by placing VIP packages and ATP Experiences alongside match information ("Home"). In the 20-post sample (N = 20), seven posts featured celebrities or other notable figures (n = 7; 35%). Across the full sample, the mean number of likes was 10,877.1 (SD = 13,878). The seven featured posts averaged 16,652.6 likes (SD = 20,031.1), which was 5,775.5 likes higher than the overall sample mean.

These observations suggest that Shanghai uses the tournament to promote an image of tennis as part of a premium urban lifestyle. Luxury hospitality and city imagery may appeal to visitors who are initially drawn to the occasion rather than to player rankings or rivalries. The marketing opportunity is to build on that interest with stories that give casual visitors a closer connection to the sport.

Cincinnati Open

The Cincinnati Open's content was largely organized around its "Heart of Tennis" brand. Player experiences and fan responses helped create a friendly atmosphere in which everyone appeared welcome. Meeting players and enjoying tournament traditions were recurring themes, while lifestyle content was less prominent. The tournament's official brand page similarly emphasizes the connection among the place, the people, and the players ("Our Brand"). In the 20-post sample (N = 20), 12 posts focused on professional players (n = 12; 60%), while eight focused on the venue, tournament branding, or other experiences (n = 8; 40%). The full sample averaged 6,930 likes (SD = 13,216). The 12 player-focused posts averaged 10,582 likes (SD = 16,253), or 3,652 more likes than the overall sample mean.

The higher average for player-focused posts fits Cincinnati's emphasis on personal connection. Its welcoming identity is expressed through opportunities to meet players, watch them practice, and share in the tournament's traditions. These themes offer a basis for stories that make players feel approachable while keeping the event's tennis identity central.

Miami Open

Miami Open content showed a hybrid strategy, combining tennis with food, fashion, and music. Food and entertainment were presented as part of the tournament experience, creating the closest U.S. parallel in this comparison to Shanghai's lifestyle positioning. Its 2026 Opening Day, for example, offered free admission to qualifying matches, practice sessions, and campus programming, with a culinary passport available as a paid add-on ("2026 Miami Open"). In the 20-post sample (N = 20), 14 posts featured well-known professional players (n = 14; 70%), while six focused on the wider tournament experience (n = 6; 30%). The mean for the full sample was 19,550 likes (SD = 22,816). The 14 player-focused posts averaged 27,421 likes (SD = 23,218), which was 7,871 likes above the full-sample mean.

The Opening Day event illustrates an effort to broaden participation by inviting visitors to experience more than the matches. At the same time, the player-focused posts received a higher average number of likes than the full sample. Together, these observations suggest that Miami can use lifestyle activities to attract interest while keeping players and tennis at the center of its communication.

BNP Paribas Open at Indian Wells

Indian Wells content consistently centered on the "Tennis Paradise" identity. Player accomplishments and stories were central to the sampled content. Premium and lifestyle experiences were available, but they generally supported the tennis experience rather than taking the spotlight. Featured players in the sample included Aryna Sabalenka, Jannik Sinner, and Alex Eala. The official website also promotes world-class tennis, off-court amenities, and access to practice courts (*BNP Paribas Open*). In the 20-post sample (N = 20), 13 posts featured prominent players (n = 13; 65%), and the remaining seven focused on other aspects of the event (n = 7; 35%). The full sample averaged 16,568 likes (SD = 21,629). The 13 player-focused posts averaged 22,123 likes (SD = 25,004), a difference of 5,555 likes from the overall sample mean.

The large standard deviations indicate that responses varied considerably across posts. Player stories attracted a higher average number of likes, but the tournament's broader identity also provides material for communication. Its history, atmosphere, and traditions could help explain what "Tennis Paradise" means to visitors, especially those who are still becoming familiar with the event.

Cross-Case Interpretation

The experience economy offers a way to interpret how these tournaments create value around the matches. Pine and Gilmore argue that businesses can make an offering more valuable by providing memorable, personally engaging experiences (98-99). Miami's combination of tennis, food, and music illustrates this approach. Its Opening Day programming invites visitors to experience the wider tournament environment, suggesting that lifestyle partnerships could form a more coherent part of the event's identity rather than a collection of separate attractions ("2026 Miami Open").

Glocalization offers a second explanation. Robertson uses the term to describe the interaction between global standardization and local difference (28-30). Every Masters 1000 event belongs to the same ATP category, yet each tournament translates that status into local cultural language. Shanghai links a global sport with metropolitan prestige and premium consumption. Miami connects tennis to the city's entertainment culture. Cincinnati draws on regional warmth and continuity, while Indian Wells turns its desert setting into a destination. These differences are not merely cosmetic; they shape which audiences can recognize themselves in the event.

The engagement figures also invite a distinction between attention and identification. Wann and Branscombe examine how strongly spectators identify with their team (1-2). For

individual athletes, Carlson and Donavan find an association between athlete identification and emotional attachment (193, 202). Applying these ideas to tennis, player-centered posts may give followers a person to admire, understand, or discuss. Funk and James's Psychological Continuum Model distinguishes awareness, attraction, attachment, and allegiance (121-22). On this interpretation, lifestyle content may attract initial interest, while player narratives may help cultivate attachment. The present study does not directly measure either process.

Taken together, the four samples show a consistent pattern: posts featuring recognizable players or celebrities had a higher mean number of likes than the full 20-post sample for the same account. This finding suggests a practical balance between lifestyle positioning and player-centered storytelling. The recommendations below build on each tournament's existing identity while giving audiences more ways to connect with the players and the event.

Recommendations

Shanghai: Create a Fan Ritual That Connects Status to Belonging

The Shanghai Masters should build on its premium lifestyle image by creating fan rituals that encourage casual visitors to return. Its emphasis on VIP hospitality and ATP Experiences provides a starting point. A collectible "Shanghai Tennis Passport" could reward participation in matches and tournament activities while giving fans a way to express their interest in tennis. Fans could collect stamps for attending a match, visiting a practice court, or returning in a later year. The passport should feel like a record of their experiences, with designs inspired by Shanghai, rather than a purely decorative souvenir.

This recommendation applies to the Psychological Continuum Model, which describes awareness, attraction, attachment, and allegiance as distinct levels of connection to sport (Funk and James 121-22). The aim is to help initial interest develop into stronger attachment through repeated participation. Over a two-year pilot, the program could target a 10 percent increase in engagement with passport-related content and a 5 percent increase in repeat ticket sales. These would be proposed evaluation goals. Free digital activities and affordable participation options could make the program accessible to younger and lower-income fans. Keeping the passport optional and focused on personal memories could also appeal to visitors who prefer a relaxed experience over a commercial promotion.

Indian Wells: Build a First-Time Fan Pathway

Indian Wells should make its traditional "Tennis Paradise" identity more welcoming to first-time and younger visitors. The tournament already promotes player access, practice courts, and the wider tennis experience. A "First-Time Fan Pathway" could bring together short player introductions, an interactive court map, beginner-friendly tennis explanations, and social-media features about tournament traditions. A short account of how the event developed could help newcomers understand its history. These resources would give first-time visitors a clear starting point while allowing experienced fans to continue using more detailed match coverage.

The player-introduction element draws on research linking athlete identification with emotional attachment (Carlson and Donavan 193, 202). Applied here, it could help newcomers find players to follow and make the tournament feel more familiar. Proposed targets could include a 10 percent improvement in first-time visitor satisfaction, a 10 percent increase in beginner-focused content output, and a 5 percent increase in ticket sales to new customers, each assessed against a baseline established before the pilot. Serious fans, families, and casual visitors may want different experiences. A separate beginner section, alongside traditional coverage and established events, would let visitors choose the level of guidance they need.

Miami: Connect Lifestyle Content More Closely to Tennis

The Miami Open should strengthen its position as a bridge between tennis and lifestyle culture through a year-round "Road to Miami" campaign. Its existing Opening Day programming combines tennis with culinary experiences and campus activities. Monthly videos could connect players with Miami's restaurants, fashion, music, and multicultural identity. Behind-the-scenes features would give fans a reason to follow the event between tournament editions and build anticipation for both the tennis and the city. During the tournament, the same themes could extend to on-site experiences and family activities.

The campaign applies the experience-economy principle that memorable experiences can add value around a core offering (Pine and Gilmore 98-99). A pilot could aim for a 15 percent increase in social-media engagement and a 10 percent increase in campaign-related ticket sales, measured against a prior-year baseline. Miami's varied audience means that the campaign should offer more than one route into the event. Each lifestyle feature should connect to a player or a tennis activity so that food, music, and fashion strengthen the tournament experience without overshadowing the competition.

Cincinnati: Turn Player Access into Short-Form Storytelling

The Cincinnati Open should modernize its "Heart of Tennis" identity by making player-centered storytelling more accessible to younger and first-time audiences. The event's brand already emphasizes connections among the place, the people, and the players. Short-form "Meet the Player in Cincinnati" stories could introduce players' personalities, careers, rivalries, and connections to the tournament before and during the event. Fan questions, autograph opportunities, and practice-court access could supply material for these stories, helping players feel more approachable while retaining the event's welcoming atmosphere.

Research linking athlete identification with emotional attachment supports the focus on personal connection (Carlson and Donavan 202). Proposed goals could include a 10 percent increase in social-media engagement among younger audiences and a 10 percent increase in interactions with player-focused content, compared with a baseline established before the campaign. Shares, substantive comments, and participation in player-access events could provide additional measures. Cincinnati still needs to serve dedicated tennis fans as well as newer

audiences. Keeping the stories grounded in players' own voices and the tournament's history would help preserve authenticity as its digital content develops.

Conclusion

This study examined how the Shanghai Masters and three U.S. Masters 1000 events use marketing to engage audiences. Shanghai emphasized lifestyle, luxury, celebrities, and city branding. Indian Wells and Cincinnati relied more strongly on player stories, tennis traditions, and fan experiences, while Miami combined tennis with a broader lifestyle offering. Across all four samples, posts featuring prominent players or celebrities received more likes on average than the full sample for the same account.

These findings can be interpreted through glocalization, the experience economy, and identification research: tournaments adapt a shared sport to their local setting while creating experiences and personal connections around it. Shanghai's lifestyle positioning offers a way to introduce tennis to new audiences, while the U.S. events draw on established identities to sustain interest. The recommendations build on these strengths through fan rituals, beginner guidance, year-round lifestyle content, and accessible player stories. The broader implication is that international sports marketing needs to respond to the culture and audience of each event rather than rely on a single approach.

Works Cited

"ATP Masters 1000 Shanghai | Overview." ATP Tour, www.atptour.com/en/tournaments/shanghai/5014/overview. Accessed 7 Sept. 2026.

BNP Paribas Open. bnpparibasopen.com. Accessed 7 Sept. 2026.

Carlson, Brad D., and D. Todd Donavan. "Human Brands in Sport: Athlete Brand Personality and Identification." Journal of Sport Management, vol. 27, no. 3, 2013, pp. 193-206. https://doi.org/10.1123/jsm.27.3.193.

Funk, Daniel C., and Jeff James. "The Psychological Continuum Model: A Conceptual Framework for Understanding an Individual's Psychological Connection to Sport." Sport Management Review, vol. 4, no. 2, 2001, pp. 119-50. https://doi.org/10.1016/S1441-3523(01)70072-1.

Pine, B. Joseph, II, and James H. Gilmore. "Welcome to the Experience Economy." Harvard Business Review, vol. 76, no. 4, July-Aug. 1998, pp. 97-105. hbr.org/1998/07/welcome-to-the-experience-economy.

Robertson, Roland. "Glocalization: Time-Space and Homogeneity-Heterogeneity." Global Modernities, edited by Mike Featherstone et al., SAGE Publications, 1995, pp. 25-44. https://doi.org/10.4135/9781446250563.n2.

"Home." Rolex Shanghai Masters, en.rolexshanghaimasters.com/en. Accessed 7 Sept. 2026.

"Our Brand." Cincinnati Open, cincinnatiopen.com/our-brand/. Accessed 7 Sept. 2026.

"2026 Miami Open Presented by Itaú Announces Free Opening Day on March 15." Miami Open, 26 Feb. 2026, www.miamiopen.com/latest-news/miami-open-presented-by-itau-announces-free-opening-day/. Accessed 7 Sept. 2026.

Wann, Daniel L., and Nyla R. Branscombe. "Sports Fans: Measuring Degree of Identification with Their Team." International Journal of Sport Psychology, vol. 24, no. 1, 1993, pp. 1-17.

The Architecture of Time: Horizons, Constraints, and the Quality of Business Decision-Making By Riley Popovich

Abstract

This paper examines how time horizons and time constraints shape the quality of business decision-making across investing, negotiations, and organizational performance. While existing research explores short-termism, bounded rationality, technological acceleration, and behavioral bias independently, these topics are rarely unified within a single temporal framework. To address this gap, the paper introduces the Dual-Temporal Model of Decision Quality, which conceptualizes time along two dimensions: horizon, the span over which decision-makers evaluate outcomes, and constraint, the limits within which decisions must be made. The model argues that decision quality depends on the alignment between these dimensions. Long horizons paired with flexible constraints support reflection, innovation, learning, and strategic resilience, while short horizons combined with tight constraints encourage reactivity, bias, and short-termism. The paper applies this framework across investment behavior, negotiations and deal-making, and organizational operations, demonstrating that temporal structure shapes both outcomes and the cognitive processes underlying judgment. Evidence from behavioral finance, psychology, and management research shows that compressed horizons and excessive time pressure reduce analytical depth, increase emotional decision-making, and weaken long-term strategic performance. The analysis also explores how digital transformation, automation, and artificial intelligence intensify these pressures by accelerating information flows and compressing decision cycles. Although technological systems improve efficiency and responsiveness, they often reduce the temporal space available for reflection and ethical reasoning. The paper concludes that organizations should treat time as a strategic and ethical design variable by balancing efficiency with deliberation to improve decision quality, adaptability, and long-term sustainability in modern business.

Introduction

Time is one of the most underexamined forces shaping the quality of business decisions. Whether an investor responds to volatility, a negotiator evaluates a deal, or a manager directs daily operations, the depth and accuracy of their judgment depend on two temporal factors: how far ahead they are able to plan (time horizon) and how quickly they are required to act (constraints). For instance, imagine a firm that has invested heavily in developing a new product, and in its first year the project reports a sharp decline in quarterly earnings. In a company operating under high temporal constraints, where executives are judged on immediate performance, cash reserves are thin, and investors demand visible short-term returns, the decline triggers alarm. Board members call emergency meetings, analysts downgrade forecasts, and managers are pressed to “fix” the numbers before the next earnings call. The project is shut down, employees are reassigned or laid off, and remaining capital is redirected toward safer, faster-yielding initiatives. The firm stabilizes its quarterly metrics, but it also abandons the

possibility of future gains, discourages experimentation, and signals internally that volatility will not be tolerated. By contrast, in a firm supported by patient capital and greater liquidity, the same earnings dip is absorbed as part of a longer developmental arc. Leadership communicates continuity, funding remains in place, and teams are encouraged to diagnose and refine rather than retreat. In this environment, the temporary loss becomes an investment in learning, iteration, and long-term positioning. Over time, the first firm becomes increasingly optimized for short-term survival and external signaling, while the second builds innovation capacity and accumulates long-run value. The difference is not merely in decision preference but in structure: when time is compressed, uncertainty is treated as failure; when time is expanded, uncertainty can be cultivated into advantage. However, despite its influence, time is rarely treated as a structural variable in management or finance. It is usually discussed as a passive backdrop rather than as a concrete feature of organizational decision-making, such as how much time decision-makers are given to evaluate information or how far into the future outcomes are expected to matter, that actively shapes how decisions are made.

This review addresses a central question: How do time limits and different time horizons affect the quality of business decisions, especially in investing, dealmaking, and day-to-day performance? Popular topics that have been explored in previous works of literature, such as short-termism, bounded rationality, technological acceleration, and risk behavior, are rarely unified under a single temporal framework yet all are defined by their relationship with time. The act of overlooking the role of time has become increasingly consequential as firms operate in environments where planning cycles shorten, information flows accelerate, and the pressure to decide quickly intensifies.

To examine this issue, the paper introduces the Dual-Temporal Model of Decision Quality, which conceptualizes time along two dimensions—horizon and constraint—and argues that decision quality depends on the alignment between them. The review then applies this model across three domains of modern business: investment behavior, negotiations and dealmaking, and organizational operations. By synthesizing evidence from psychology, finance, and management, the paper demonstrates why understanding the temporal structure of decisions is essential for improving judgment, reducing bias, and sustaining long-term strategic performance.

Time as a Horizon and Constraint: The Dual-Temporal Foundations of Decision Quality

Time defines how organizations plan, decide, and compete. It influences not only the speed of business activity but also the quality of judgment that underlies it. A company choosing between rapid expansion and steady growth and an investor weighing short-term returns against long-term returns are both operating within the boundaries that time creates. In management and finance, time can be understood through two interdependent dimensions: horizons describe the temporal span over which decision-makers project and evaluate outcomes, while constraints represent the limits that determine how much flexibility in time and deliberation are possible. Together, these dimensions form the temporal architecture of decision-making and reveal whether firms act strategically over the long term or reactively in the short term.

A long horizon provides the foundation for patience, innovation, and sustained growth. For example, consider a firm deciding whether to invest in a research and development project that is expected to generate financial losses for the first two or three years before producing returns in year four or five. The key question is not simply whether the project is losing money, but how much time the firm has formally decided to give it before judging its success. If leadership has established a five-year evaluation window and explicitly defined the first three years as a development phase, then early losses are interpreted as expected and necessary. Managers compare results against the original timeline and ask whether progress aligns with projected milestones rather than whether the project is currently profitable. In this case, losses signal movement along a planned path. By contrast, if executives are evaluated quarterly or annually and face pressure to improve short-term earnings within twelve months, those same losses appear alarming. After several quarters of negative results, stakeholders may conclude that the project is failing simply because it has not yet produced visible returns. What changes is not the project itself, but the time frame used to measure it. A two-year loss can represent disciplined long-term investment under a five-year horizon, or unacceptable underperformance under a one-year horizon. Whether the firm re-strategizes or continues funding depends on the temporal boundary that defines when judgment occurs. Firms that operate with extended planning spans invest more consistently, build stronger innovation capacity, and perform more steadily across market cycles (Akguc and Choi; Flammer and Bansal). Short horizons, by contrast, compress evaluation windows and emphasize immediacy. A manager whose performance is assessed primarily on quarterly results may respond to a temporary downturn by prioritizing actions that improve short-term metrics, such as cutting long-term investment or deferring necessary expenditures, even when these actions weaken the firm's future position. Under a longer evaluation horizon, the same information would be weighed differently, allowing the manager to tolerate short-term declines in service of longer-term value creation. Publicly listed firms, constrained by quarterly reporting and shareholder expectations, often respond to short-term fluctuations at the expense of long-term experimentation (Asker et al.). This behavior illustrates corporate short-termism, a systemic tendency to privilege rapid outcomes over durable value creation (Wiersema et al.). Executive compensation and incentive structures reinforce this pattern: when stock-based pay vests over short time periods and performance is evaluated on near-term results, managers are incentivized to prioritize immediate financial metrics rather than long-term strategic endurance (Edmans et al.).

Horizons also condition how organizations perceive and tolerate risk. When time is perceived as abundant, managers treat volatility as an opportunity for information gathering and learning. In this sense, time functions as a strategic resource, because greater temporal slack allows organizations to explore, experiment, and generate the kinds of insights that support long-term innovation. When time feels scarce, volatility appears as a threat, producing defensive and conservative choices. Investors evaluating performance over short intervals exhibit stronger reactions to transient signals and allocate capital less efficiently, while those operating over longer horizons demonstrate greater patience and analytical depth (Dessaint et al.). In this way,

horizons shape not only strategic direction but also a frame of reference through which managers interpret information, evaluate risk, and decide how to respond. In practice, under conditions of temporal abundance, a firm with multi-year performance targets may respond to short-term earnings volatility by increasing experimentation or gathering additional information, whereas under conditions of temporal scarcity, a firm facing imminent reporting or funding deadlines may interpret the same volatility as a signal to reduce exposure or halt investment altogether. Constraints represent the second temporal force. They include deadlines, production schedules, and external demands that limit how long decision-makers can analyze information before acting. When constraints are moderate, decision-makers have sufficient time to collect and verify data, consult relevant stakeholders, compare alternatives, and revisit assumptions, enhancing focus and discipline. When constraints become excessive, these activities are curtailed, reducing analytical depth, increasing reliance on heuristics, and producing less consistent decisions. Under high time pressure, decision-makers become less consistent in risky choices and exhibit changes in risk preference and outcome evaluation (Olschewski and Rieskamp; Lin and Jia). In negotiation settings, strict deadlines substantially increase disagreement and can alter the types of agreements negotiators reach, demonstrating how temporal scarcity affects bargaining outcomes (Karagözoğlu and Kocher). Constraints, therefore, determine not only how quickly decisions must be made but also the way those decisions are formed, shaping whether managers rely on careful analysis or on quick, heuristic thinking.

After synthesizing prior research and analyzing a range of outcomes and theories, we developed the Dual-Temporal Model of Decision Quality. This model integrates the two dimensions developed in this research paper to explain how their alignment governs strategic performance. The model posits that decision quality depends less on the total amount of time available than on the coherence between horizon and constraint. When firms or people in a firm have a long time horizon with no particular constraints, decision-making becomes reflective, adaptive, and learning-oriented and not driven by external time pressure. On the other hand, when working on a shorter window of time, and having immediate constraints and short horizons combined with tight constraints, decision-making becomes reactive, myopic, and risk-averse. Accordingly, a manager facing an imminent deadline may respond to unexpected market volatility by delaying investment, cutting discretionary spending, or avoiding strategic commitments altogether, even when these actions limit longer-term opportunities. However, short horizons don't always align with short-term constraints and long horizons don't always align with long term constraints. This causes a misalignment – when long horizons are paired with tight short term constraints or when short horizons lack sufficient structure and are paired with loose, long term constraints – and produces predictable inefficiencies. A firm that articulates a long-term strategic plan but evaluates managers primarily on rigid quarterly performance targets creates misaligned incentives. Managers are encouraged to endorse long-term goals in principle while acting to satisfy short-term evaluation criteria in practice, producing cognitive dissonance and inconsistent decision-making that ultimately undermines performance. Conversely, a firm operating with short planning horizons and minimal temporal boundaries may

retain flexibility in the near term but lack sustained strategic direction, leading to reactive, tactically driven decisions and gradual organizational drift (Flammer and Bansal; DesJardine et al.).

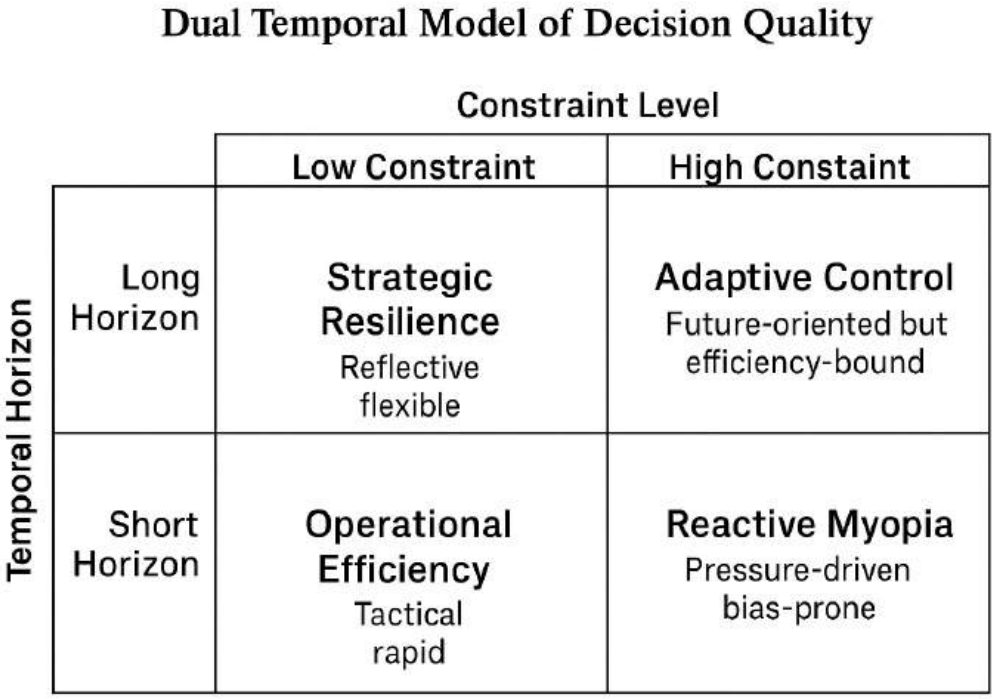


Fig 1: Dual Temporal Model of Decision Quality

This relationship can be visualized across four configurations of decision behavior (Figure 1). Below, I define each of the configurations:

1. Strategic Resilience (Long Horizon + Low Constraint) – I define strategic resilience as a combination of long time horizons paired with low constraints. With this combination, firms balance patience and flexibility, investing steadily in innovation and learning (Flammer and Bansal).

2. Adaptive Control (Long Horizon + High Constraint) – When configured towards adaptive control, long horizons and high constraints are used. Organizations maintain future orientation but operate under institutional or regulatory pressure, sustaining efficiency at the cost of experimentation (DesJardine et al.).

3. Operational Efficiency (Short Horizon + Low Constraint) – When short horizons and low constraints are paired, operational efficiency becomes prevalent. Firms act rapidly and

tactically, optimizing for output but risking volatility and drift (Wang and Liu).

4. Reactive Myopia (Short Horizon + High Constraint) – When a strict deadline (short horizon) and high constraint are paired, short-sightedness and poor, skewed decisions may be made (reactive myopia). This is the most fragile configuration, where pressure and immediacy reinforce bias and error (Dessaint et al.; Olschewski and Rieskamp).

Overall, horizons define foresight and patience; constraints determine focus and execution. Their interaction forms an organization's temporal alignment—the balance between anticipation and urgency that governs strategic behavior. Treating time as a design variable rather than a fixed background condition reveals why firms facing similar information and incentives can diverge dramatically in performance and resilience (Akguc and Choi; Asker et al.; Flammer and Bansal; Dessaint et al.; Olschewski and Rieskamp; Wiersema et al.; DesJardine et al.; Wang and Liu).

Applications Across Business Domains

The influence of time on decision-making becomes most visible in applied business settings, where trade-offs between speed and depth directly shape financial outcomes, relationships, and performance. A fund manager choosing whether to sell during a market dip, a negotiator deciding how long to wait for better terms, and a firm determining how quickly to scale production each encounter distinct combinations of horizon and constraint. As shown in Figure 1, the Dual Temporal Model explains why these situations elicit different cognitive and strategic responses. Long horizons paired with flexible constraints create the conditions for strategic resilience, where judgment becomes reflective, exploratory, and learning oriented. Short horizons paired with tight constraints compress deliberation and shift behavior toward the reactive myopia quadrant. Applied business settings therefore serve as natural experiments in how temporal structure shapes decision quality. Across domains, observable behaviors often align with the model's predicted configurations, demonstrating that temporal framing influences not only outcomes but also the underlying processes of reasoning.

Investing and Financial Decisions

Investment behavior offers one of the clearest illustrations of how horizon and constraint interact. Changes in investment horizon correspond directly to the vertical axis of Figure 1, shifting behavior between reflective strategic reasoning and reactive short-term evaluation. Short horizons heighten sensitivity to market fluctuations, discourage patience, and activate the cognitive tendencies associated with narrow temporal windows in the model. Extended horizons foster diversification, deliberate risk management, and long-run learning, reflecting the conditions of the strategic resilience quadrant. Evidence shows that investors operating within limited horizons react disproportionately to transitory information, producing volatile and inefficient capital allocation (Dessaint et al.). Experiments confirm that narrow horizons

exacerbate myopic loss aversion and emotional reactivity to volatility (Wesslen et al.), outcomes consistent with the reactive side of the model.

The constraint dimension operates in parallel. High reporting frequency, shareholder scrutiny, and regulatory oversight create the tight constraint conditions depicted on the right side of Figure 1, reinforcing reactive tendencies and shortening decision cycles. Firms with longer evaluation periods move toward the low-constraint quadrants, encouraging innovation, experimentation, and long-term positioning (Wang and Liu). Digital platforms increase both information velocity and temporal pressure, contracting horizons while simultaneously intensifying constraints (Zhang and Wang; Guo et al.). Investment behavior therefore demonstrates how horizon and constraint interact as a coupled temporal system, and how their alignment or misalignment produces distinct patterns of judgment consistent with the Dual Temporal Model.

Negotiations and Deal-Making

Negotiation and entrepreneurial deal-making environments foreground the constraint dimension even more directly. Deadlines, limited deliberation windows, and accelerated deal cycles place negotiators squarely in the high-constraint region of Figure 1. Under these conditions, cognitive capacity narrows, information search becomes truncated, and behavior moves toward the reactive myopia quadrant. Empirical work shows that strict deadlines substantially increase disagreement and can alter the types of agreements negotiators reach (Karagözoğlu and Kocher). These outcomes mirror the model's prediction that tight constraints compress deliberation and shift emphasis from reflective analysis to rapid closure.

Entrepreneurial finance exhibits similar dynamics. Venture capitalists evaluating many proposals with limited time operate within both short horizons and high constraints, a combination associated with the most fragile quadrant in the model. This configuration increases rejection of innovative but ambiguous ventures, favoring familiar patterns and easily measurable performance indicators (Vazirani and Bhattacharjee). Emotional responses also strengthen under dual compression of horizon and constraint, reducing the ability to process risk or uncertainty in an open-ended manner. When the temporal structure is loosened through staged evaluations or extended due diligence, negotiators and investors display behaviors more consistent with the adaptive control and strategic resilience quadrants in Figure 1. These patterns illustrate the theoretical claim that temporal design governs not only outcomes but also the cognitive pathways through which agreements are formed.

Organizational Operations and Performance

Organizational environments often demonstrate the full interaction of both temporal dimensions at once. Many modern firms operate within compressed horizons and heightened constraints, situating them in the short-horizon, high-constraint quadrant of Figure 1. Real-time information systems in finance exemplify this configuration. High-frequency trading reduces horizons to millisecond cycles while imposing rigid performance constraints, creating an

environment that rewards rapid output yet reduces the opportunity for reflective reasoning (MacKenzie). The Dual Temporal Model predicts that such conditions increase volatility, reduce learning, and elevate reliance on procedural heuristics, patterns observed empirically in these domains.

Legal and governance structures reinforce temporal compression by shortening planning horizons and constraining strategic experimentation (DesJardine et al.). These forces anchor firms to the reactive quadrants unless counterbalanced by cultural or structural mechanisms that extend horizons and relax constraints. Leaders with broader experiential backgrounds or cross-cultural exposure often adopt longer planning cycles and broader contextual awareness, illustrating movement toward the adaptive control quadrant (Ge et al.). Organizational cultures that emphasize long-term value correlate with innovation and resilience, traits associated with the upper-left quadrant of Figure 1 (Bukowski and Rudnicki; Muskat et al.). National cultural differences, including differences in long-term orientation, are associated with variation in corporate sustainability reporting practices (Hassanein et al.), consistent with the vulnerabilities of the short-horizon, high-constraint configuration.

These dynamics show that organizational time is socially constructed rather than externally imposed. When firms intentionally design temporal structures through flexible reporting schedules, staged project reviews, and incentive systems that support long-horizon reasoning, they move into the model's resilient and adaptive regions. Such alignment between horizon and constraint increases both performance and innovation, demonstrating how the Dual Temporal Model functions as a framework not only for analysis but also for temporal design. Across investing, negotiations, and operations, empirical findings consistently reflect the theoretical architecture of the Dual Temporal Model. Short horizons combined with high constraints generate reactive and fragile decision environments, while extended horizons paired with flexible constraints support learning, strategic coherence, and resilience. The model therefore provides a unified explanation for why temporal framing shapes behavior across diverse business settings and clarifies how the intentional structuring of time can improve decision quality, innovation, and organizational stability.

Technological Acceleration and the Compression of Time

Technological advancement has fundamentally redefined the temporal structure of business decision-making. The rise of instant communication, algorithmic trading, and automated supply chains has compressed both the horizons over which firms plan and the constraints under which they operate. Time is no longer a neutral background variable but an active product of technology itself. Digital transformation and artificial intelligence have made decision-making faster, more data-driven, and often less reflective. A manager can now analyze global market data in seconds, while an AI system executes thousands of investment orders before human judgment can intervene. These developments have reshaped how managers, investors, and organizations experience time, turning it from a managerial resource into a competitive constraint that continuously demands acceleration.

High-frequency and algorithmic trading illustrate this transformation most clearly. Automated systems execute financial decisions in milliseconds, far beyond human perceptual capacity, shifting the temporal boundary of market competition from human deliberation to machine processing. As MacKenzie observes, the logic of financial markets has evolved toward what he describes as "trading at the speed of light," in which technological capability rather than strategic foresight determines advantage. This acceleration has shortened effective investment horizons, prioritizing microsecond efficiency over long-term analysis. Similarly, corporate managers increasingly rely on real-time dashboards and predictive analytics that update continuously, compressing evaluation cycles and reducing opportunities for deep reflection (Guo et al.). When decision systems operate faster than human interpretation, the constraint is no longer informational scarcity but cognitive bandwidth.

AI amplifies these effects by altering both the content and pace of decision-making. Machine learning systems optimize for immediate performance metrics, adjusting continuously in response to new data. While this enhances responsiveness and efficiency, it also reinforces short-term orientation by rewarding instantaneous adaptation over strategic patience. Studies of digital transformation confirm that automation restructures corporate finance toward speed and liquidity, often increasing short-term asset allocation at the expense of durable investment (Zhang and Wang). These patterns parallel the broader phenomenon of corporate short-termism described in organizational research, where technological acceleration interacts with institutional incentives to narrow managerial horizons (DesJardine et al.). As AI systems increasingly participate in capital allocation and risk assessment, they embed algorithmic assumptions about time that may not align with human strategic goals. The result is an organizational environment in which temporal pressure is internalized through automated processes rather than imposed externally.

The compression of time through technology also affects how constraints are experienced within firms. Digital monitoring and data analytics have enabled continuous evaluation of employee and organizational performance. This constant visibility introduces a form of temporal discipline that blurs the boundary between productivity and surveillance. Managers and employees operate under implicit deadlines created by real-time feedback systems, which encourage constant responsiveness but limit sustained concentration. Research shows that such technological systems increase the speed of decision-making without necessarily improving quality, as human actors adapt behavior to algorithmic rhythms rather than strategic objectives (Guo et al.). Over time, these mechanisms normalize haste, making delay or reflection appear inefficient even when it might improve judgment.

AI-driven environments also create new asymmetries between human and machine cognition. Algorithms process vast quantities of data instantly but lack the contextual reasoning that supports long-term vision. Human decision-makers, in turn, are forced to adapt to machine time—accelerating their own thinking to remain competitive within digitally mediated systems. This temporal mismatch has implications for both strategy and ethics. As decision cycles shorten, the window for moral reflection narrows, increasing the risk of reactive or opaque

outcomes. Organizational research suggests that the speed of technological systems can outpace the development of governance frameworks capable of managing them responsibly (DesJardine et al.). Thus, the ethical dimension of time compression lies not only in efficiency but also in the loss of deliberative space.

At the same time, technology offers tools that can extend horizons if deployed intentionally. AI systems can simulate long-term outcomes, optimize planning under uncertainty, and model delayed feedback effects. When used to enhance foresight rather than accelerate reaction, technology can help restore balance between horizons and constraints. For example, digital forecasting tools can reduce uncertainty in multi-year investments, allowing firms to plan more confidently across longer horizons (Guo et al.). The challenge lies in designing algorithms and organizational systems that align technological speed with human strategic intent. The Dual-Temporal Model suggests that sustainable decision-making requires integrating fast, data-driven processes within structures that preserve deliberation and long-term orientation. Technological acceleration has made time both a scarce resource and a competitive weapon. Automation, AI, and digital analytics have collapsed traditional horizons and intensified constraints, redefining what it means to decide, plan, and act in business contexts. While these tools expand the boundaries of what can be analyzed, they also compress the temporal space available for reflection. The future of decision quality depends on whether organizations can reconcile machine speed with human foresight—using technology not merely to accelerate choice but to design temporal systems that balance efficiency, understanding, and sustainability.

Implications and Future Directions

The study of time in business decision-making reveals that temporal structure is not an external condition but an internal variable that can be designed, measured, and managed. The Dual-Temporal Model developed in this paper contributes to behavioral finance and organizational theory by explaining how decision quality depends on the alignment between time horizons and time constraints. This framework reframes short-termism, impulsivity, and overreaction not as isolated behavioral errors but as predictable outcomes of temporal misalignment. By integrating insights from finance, psychology, and management, the model establishes a basis for understanding how firms can intentionally structure time to improve both strategic and ethical outcomes.

Implications of the Framework

In theory, the Dual-Temporal Model clarifies how differences in time design produce systematically different decision behaviors. For instance, two firms may offer identical incentives and operate in the same market, yet reach different strategic outcomes because one evaluates performance over multi-year horizons while the other imposes tight quarterly deadlines. In the first case, managers are more likely to tolerate early volatility, invest in learning, and adjust strategies over time; in the second, they are more likely to prioritize immediate stability and avoid actions with delayed payoffs. By isolating horizons and constraints

as design choices rather than background conditions, the model explains why similar organizations respond differently to the same information and why decision quality varies even when incentives and resources are held constant. It explains why firms facing identical information and incentives may diverge in performance: they operate under different temporal architectures that shape cognition and execution. The model also bridges individual and institutional levels of analysis, showing that cognitive time perception (Olschewski and Rieskamp; Lin and Jia) scales into collective temporal systems such as reporting cycles, investment horizons, and technological infrastructures (MacKenzie; DesJardine et al.). Future theoretical work can formalize these relationships by modeling time alignment as a moderating variable that predicts variance in risk behavior, learning capacity, and innovation outcomes across organizations.

Implications for Practice

For practitioners, the framework suggests that time can be intentionally designed within firms and markets. Managers can shape decision quality by calibrating how time is allocated, measured, and rewarded. Expanding planning horizons, lengthening evaluation cycles, and balancing real-time analytics with periodic reflection can help organizations achieve temporal alignment. Investment and governance policies can be structured to prevent the collapse of horizons under technological or institutional pressure. For example, lengthening executive vesting schedules and de-emphasizing quarterly earnings can counteract the short-term focus produced by accelerated information systems (Edmans et al.; Wiersema et al.). Firms adopting digital transformation and AI should design feedback systems that encourage anticipation and foresight rather than constant reaction, such as by emphasizing trend-level indicators, scenario-based projections, and longer evaluation intervals that allow decision-makers to assess patterns over time rather than respond to isolated signals (Guo et al.; Zhang and Wang). In this view, temporal design becomes a dimension of strategic control comparable to capital structure or organizational culture.

Future Research Directions

Future research can extend this work in several directions. First, empirical studies can test the Dual-Temporal Model by measuring temporal alignment and its effects on decision accuracy, investment persistence, and organizational performance. Longitudinal and experimental designs could examine whether balancing horizon length and constraint intensity produces measurable improvements in innovation and risk calibration. Second, cross-cultural research can explore how national or institutional time orientations shape the balance between long-term vision and short-term constraint (Bukowski and Rudnicki; Muskat et al.; Hassanein et al.). Third, technology-focused studies can investigate how AI-mediated decision environments modify human temporal perception, introducing new ethical and governance challenges (MacKenzie; DesJardine et al.). Finally, financial modeling could incorporate temporal alignment variables

into portfolio and valuation theory, linking organizational decision behavior to macroeconomic performance over time.

Conclusion

Time is not a neutral backdrop to business decisions but the most fundamental variable shaping their quality, ethics, and outcomes. This paper has argued that understanding time as both a horizon and a constraint reveals the core of decision-making. The Dual-Temporal Model of Decision Quality shows that success in investing, negotiations, and organizational strategy depends not solely on access to information or capital, but on how effectively planning horizons are aligned with temporal limits of execution. When time is misaligned—when horizons are shortened or constraints tighten beyond cognition—decision-making deteriorates, and organizations lose their capacity for foresight.

Across the literature and domains analyzed, the same pattern emerged: temporal design defines judgment. Short horizons and excessive time pressure led to reaction, bias, and myopia, while extended horizons paired with deliberate pacing enabled creativity, learning, and sustainability. Yet the modern economy, accelerated by automation and AI, increasingly compresses both dimensions. Decisions are now made faster than they can be understood. The challenge for contemporary management is therefore not only to act efficiently but to reclaim time as a space for reflection. If technological speed is inevitable, intentional temporal design becomes the last defense of human judgment.

This paper argues that the next frontier of management and finance research lies in treating time as a strategic resource and an ethical responsibility. Every decision system—whether human, institutional, or algorithmic—embeds assumptions about how much time matters. Designing those systems consciously, with awareness of how time shapes cognition and value, is essential to the integrity of modern business. The capacity to think well under changing temporal conditions will define not only competitive advantage, but also the moral and intellectual quality of economic life itself.

Works Cited

Akguc, Serkan, and Jongmoo Jay Choi. “Time Horizon and Corporate Investment: Evidence from Private and Public Firms around the World.” Research in International Business and Finance, vol. 73, 2025, article 102577, doi:10.1016/j.ribaf.2024.102577.

Asker, John, Joan Farre-Mensa, and Alexander Ljungqvist. “Corporate Investment and Stock Market Listing: A Puzzle?” The Review of Financial Studies, vol. 28, no. 2, 2015, pp. 342–390, doi:10.1093/rfs/hhu077.

Bukowski, Andrzej, and Seweryn Rudnicki. “Not Only Individualism: The Effects of Long-Term Orientation and Other Cultural Variables on National Innovation Success.” Cross-Cultural Research, vol. 53, no. 2, 2019, pp. 119–162, doi:10.1177/1069397118785546.

DesJardine, Mark R., Wei Shi, and Yin Cheng. “Why Are Corporate Investment Horizons Shrinking? Uncovering the Spillover Effects of Shareholder Litigation.” Strategy Science, vol. 9, no. 3, 2024, pp. 229–247, doi:10.1287/stsc.2022.0111.

Dessaint, Olivier, Thierry Foucault, and Laurent Frésard. “The Horizon of Investors’ Information and Corporate Investment.” HEC Paris Research Paper, no. FIN-2022-1462, HEC Paris, 2022, doi:10.2139/ssrn.4276832.

Edmans, Alex, Vivian W. Fang, and Katharina A. Lewellen. “Equity Vesting and Investment.” The Review of Financial Studies, vol. 30, no. 7, 2017, pp. 2229–2271, doi:10.1093/rfs/hhx018.

Flammer, Caroline, and Pratima Bansal. “Does a Long-Term Orientation Create Value? Evidence from a Regression Discontinuity.” Strategic Management Journal, vol. 38, no. 9, 2017, pp. 1827–1847, doi:10.1002/smj.2629.

Ge, L. G., A. Muller, T. Gong, and C. Qian. “CEO International Experience in Advanced Market Economies and Firm Investment Horizon in a Transitioning Economy.” Management and Organization Review, vol. 20, no. 1, 2024, pp. 111–138, doi:10.1017/mor.2023.41.

Guo, Yundan, Han Liang, and Li Shen. “Digital Transformation and Corporate Financial Asset Allocation: Evidence from China.” arXiv, 2025, arXiv:2509.09095.

Hassanein, Ahmed, Ahmed Bani-Mustafa, and Khalil Nimer. “A Country’s Culture and Reporting of Sustainability Practices in Energy Industries: Does a Corporate Sustainability Committee Matter?” Humanities and Social Sciences Communications, vol. 11, 2024, article 1140, doi:10.1057/s41599-024-03536-x.

Karagözoğlu, Emin, and Martin G. Kocher. “Bargaining under Time Pressure from Deadlines.” Experimental Economics, vol. 22, 2019, pp. 419–440, doi:10.1007/s10683-018-9579-y.

Lin, Chiuhsiang Joe, and Huiqiao Jia. “Time Pressure Affects the Risk Preference and Outcome Evaluation.” International Journal of Environmental Research and Public Health, vol. 20, no. 4, 2023, article 3205, doi:10.3390/ijerph20043205.

MacKenzie, Donald. Trading at the Speed of Light: How Ultrafast Algorithms Are Transforming Financial Markets. Princeton University Press, 2021.

Muskat, Birgit, Tanja Hörtnagl, Mike Peters, and Anita Zehrer. “Innovation Capability and Culture: How Time-Orientation Shapes Owner-Managers’ Perceptions.” Journal of Hospitality and Tourism Management, vol. 47, 2021, pp. 217–227, doi:10.1016/j.jhtm.2021.03.018.

Olschewski, Sebastian, and Jörg Rieskamp. “Distinguishing Three Effects of Time Pressure on Risk Taking: Choice Consistency, Risk Preference, and Strategy Selection.” Journal of Behavioral Decision Making, vol. 34, no. 4, 2021, pp. 541–554, doi:10.1002/bdm.2228.

Vazirani, Ashish, and Titas Bhattacharjee. “The Impact of Time Constraints on New Venture Investment Decisions: An Experimental Study.” Venture Capital, vol. 28, no. 1, 2026, pp. 93–117, doi:10.1080/13691066.2024.2360914.

Wang, Yan, and Shen Liu. “Managerial Short-Termism, ESG Performance, and Corporate Financial Fraud: A Heterogeneity Analysis Based on Corporate Ownership.” International Review of Economics & Finance, vol. 102, 2025, article 104385, doi:10.1016/j.iref.2025.104385.

Wesslen, Ryan, Alireza Karduni, Douglas Markant, and Wenwen Dou. “Effect of Uncertainty Visualizations on Myopic Loss Aversion and the Equity Premium Puzzle in Retirement Investment Decisions.” IEEE Transactions on Visualization and Computer Graphics, vol. 28, no. 1, 2022, pp. 454–464, doi:10.1109/TVCG.2021.3114692.

Wiersema, Margarethe, Haeyoung Koo, Weiru Chen, and Yu Zhang. “Corporate Short-Termism: A Review and Research Agenda.” Journal of Management, vol. 51, no. 6, 2025, pp. 2389–2418, doi:10.1177/01492063241303392.

Zhang, Ping, and Yiru Wang. “Digital Transformation: A Systematic Review and Bibliometric Analysis from the Corporate Finance Perspective.” arXiv, 2024, arXiv:2412.19817.

Beyond the Affordable Housing Label: Rent-Based Classification, Cyclicality, Growth, and Valuation in U.S. Multifamily REITs By Kavin Krishnarajah

Abstract

This study develops a geographically weighted framework to classify publicly traded U.S. multifamily residential real estate investment trusts (REITs) by the average monthly rents these firms charge relative to the U.S. Department of Housing and Urban Development (HUD) fair market rents (FMR). For each REIT, we calculate a unit-weighted ratio of average monthly portfolio rent to the HUD's two-bedroom FMR, using FY2025 rents from 10-K reports and the HUD's published county-level FMR files from FY2025-original. The multifamily residential REIT universe in our study includes eleven firms: six at or below the geographically weighted FMR benchmark and five as controls. We collected approximately 10 years of firm-level financial data from SEC EDGAR 10-K reports from FY2016- FY2025 and used it to test whether lower-rent REITs display lower volatility (H1), faster growth in funds from operations (H2), or a valuation discount (H3). Because the sample has only 11 firms, we evaluate statistical significance over all 462 possible assignments of the six-firm lower-rent group, with a Romano-Wolf stepdown correction across hypotheses and a Monte Carlo simulation harness to determine the size of each test. Our results do not support cyclical decoupling or a valuation discount. However, it does show a statistically significant, economically considerable difference in FFO-margin growth path. This finding is still positive but loses statistical significance in a re-run with near-threshold exclusion and in a continuous-ratio specification. The main contribution is a reproducible classification system that differentiates relatively lower-rent multifamily residential REITs from officially subsidized affordable-housing firms.

Keywords multifamily REITs; housing affordability; Fair Market Rent; REIT valuation; cyclical risk; randomization inference; small-sample econometrics

1. Introduction

Housing is in an unusual position compared to other financial assets because it is also a basic need. This duality leads to a common intuition among financiers, housing researchers, and decision-makers: apartment portfolios that charge lower rents should be more defensive because demand for essential, lower-rent housing is presumed to be more resilient in economic slumps. However, the counterargument is equally plausible: tenants in lower-rent apartments are more exposed to unemployment, inflation, and shocks, so lower rent doesn't necessarily mean lower risk. Which argument is more dominant is an empirical question and hard to answer, because listed multifamily REITs aren't labeled in a way that allows a straightforward comparison. Existing REIT research classifies by sector, geography, size, leverage, or even strategy, but little evidence on position within local rent distributions. Additionally, raw dollar rents are misleading; for example, $1500 is cheap in San Francisco, but expensive in Memphis. Therefore, rent classification must be relative to each REIT's local market where its units are located.

This paper asks whether REITs with lower rents relative to local benchmarks differ in cyclicality, operating growth, or valuation. This question is examined in three ways. First, are relatively lower-rent REITs more decoupled from the market and less volatile compared to the remaining multifamily REITs (H1)? Second, do relatively lower-rent REITs have faster FFO-margin growth (H2)? Third, do relatively lower-rent REITs trade at a valuation discount after observable fundamentals are held constant (H3)?

Our paper develops a classification system that generates a classification ratio for every multifamily REIT in our study. The ratio is the REIT's unit-weighted average portfolio rent divided by the unit-weighted HUD two-bedroom FMR across the REIT's disclosed markets. Every FMR value comes from HUD's published country files in the first FY2025 report. Once a ratio has been determined for every REIT in the study, the classification system then sorts the REITs based on a threshold of 1.0, meaning that if the average rent is at or below the benchmark that federal voucher payments are based on, then that REIT would have a ratio at or below 1.0 and would be placed into the lower-rent bucket. REITs with a ratio above the threshold of 1.0 are placed in the control group. Our classification framework yielded a lower-rent group of six treated firms (NXRT, CLPR, BRT, ESS, MAA, IRT) and a control group of five firms (CSR, CPT, UDR, AVB, EQR). The ratios range from 0.77 to 1.21. Throughout the paper, we refer to treated firms as HUD-relative lower-rent and never affordable housing REITs because our classification cannot establish that a REIT has subsidy participation, voucher acceptance, or a legal affordability designation.

The study uses eleven firms, the entire population of listed U.S. multifamily equity REITs with data from most experimental years; therefore, asymptotic inference is unreliable. Instead, the study tests every test statistic against its exact permutation distribution over all $C(11,6) = 462$ possible assignments of six firms to the treated group, applies a Romano–Wolf stepdown correction across the hypothesis family (Romano and Wolf), and calibrates the entire pipeline with simulated positive and null controls that verify size, power, and parameter recovery before any real-data conclusion is drawn.

The findings are candid and mixed. The cyclical-beta contrast carried the predicted sign but was far from significant, and the calibration shows this test now has substantial power, so large decoupling effects can genuinely be ruled out; the residual-volatility contrast ran opposite to prediction but comes from a component with very low calibrated power. The 2020 COVID-19 contrast favored lower-rent firms against clean pre-trends but is a low-powered event design. The only statistically significant result based on real data relates to the trend in FFO margins: firms with lower rents saw their margins increase meaningfully faster (adjusted $p = 0.024$), a conclusion which maintains the same sign regardless of the alternative classification rules examined but only achieves conventional significance when the significance threshold is applied.

The conditional valuation coefficient was only slightly positive; its sign was unstable across specifications, and it was statistically indistinguishable from zero in all cases. The paper thus makes four contributions: it provides a clear and explicitly vintage-based measurement approach; it offers immediate empirical support for three propositions which have generally been

assumed; it shows an example of small-sample inference in a situation where cluster-based methods fail; and it includes a completely auditable research pipeline that has been validated by simulation. The rest of the paper is arranged as follows: Section 2 sets out the conceptual framework, Section 3 outlines the classification methodology, Section 4 describes the financial data, Section 5 gives the details of the econometric and inference methods, Section 6 reports the results, Section 7 presents the robustness and calibration evidence, and Section 8 discusses the implications and limitations before Section 9 concludes.

2. Institutional Background and Conceptual Framework

2.1 The Multifamily REIT Operating Model

Multifamily equity REITs earn revenue primarily from residential rents; therefore, their cash flow depends on occupancy, achieved rents, concessions, turnover, maintenance, and financing costs. FFO (net income + real estate depreciation/amortization added back) is a financial data point used to measure cash flow for standard REITs. It adds back depreciation and amortization because accounting depreciation understates the economic performance of appreciating property (NAREIT). Therefore, we use FFO, rather than net income, as the outcome measure throughout the study.

2.2 What "Lower-Rent" Means in This Paper

Fair Market Rent is HUD's annual estimate of gross rent, typically near the 40th percentile of local market rents, and it anchors the payment standards for the Housing Choice Voucher program (U.S. Department of Housing and Urban Development). If a REIT's average rent is at or below its local FMRs, then that portfolio's rent is at or below the government's own assistance benchmark for qualifying low-income families in those markets. The classification ratio captures a REIT's average rent relative to the local benchmark. It does not establish that the REIT participates in a subsidy program, accepts vouchers at every property, satisfies income-restricted affordability requirements, or operates legally designated affordable housing.

2.3 Competing Theoretical Mechanisms

Every hypothesis has a mechanism predicting it and a countering mechanism predicting the opposite. For H1 (stability), essential housing with very high demand predicts cash flows that resist recession, leading to lower cyclical beta and lower residual variance. In contrast, tenants more exposed to job loss and financial stress predict the opposite. For H2 (margin trajectory), full buildings, no costly amenities, and flat costs with rent flows predict margins compound faster over time, while older assets, heavy maintenance, and tenants that can't absorb rent hikes predict the opposite. For H3 (valuation), investor stigma toward a neglected segment predicts a negative conditional coefficient on the lower-rent indicator.

In contrast, rational pricing of leverage, quality, geography, and growth predicts no conditional difference. Because the mechanisms have competing outcomes, the hypotheses are

directional, but none of them are presumed true in advance. Note that H2 tests the margin outcome, not the turnover or maintenance mechanism because EDGAR doesn't consistently disclose recurring maintenance capital expenditures.

3. Portfolio Classification Methodology

3.1 REIT Universe

Eligibility required the following criteria: publicly traded, U.S.-listed, equity REIT, and multifamily-focused. We needed the REITs to be publicly traded because our H3 requires a market valuation (P/FFO needs a price), and publicly traded firms must file with the SEC. Our classification process is built on HUD FMRs, which exist only for U.S. metros, so the REITs must be based in the U.S. Since we are comparing a two-bedroom apartment FMR, the REIT must be multifamily, and restricting to one property type ensures outcome differences are not just property-type differences. The REITs also need to be equity REITs, as mortgage REITs hold loans, not properties; therefore, they have no portfolio rents to classify.

The sample is the population of currently listed U.S. multifamily equity REITs as of FY2025, not a convenience subset. Firms acquired or delisted mid-window are absent from the sample: Bluerock Residential Growth REIT and Preferred Apartment Communities, both acquired by Blackstone affiliates in 2022, and AIR Communities, taken private by Blackstone in 2024. Each currently listed residential REIT not in the universe fails a stated criterion. Elme is in liquidation, having sold its multifamily portfolio in 2025. Aimco is a development-focused platform in liquidation. Veris Residential is multifamily-dominant only from the early 2020s. The panel draws on the 10-K filings from the 11 REITs in our universe from 2016-2025 via EDGAR XBRL, while the classification inputs come specifically from the FY2025 10-Ks benchmarked against HUD FY2025 FMRs as originally published.

A final screening issue concerns mixed-use exposure. We classify firms as multifamily when multifamily is the dominant reported segment, and we construct the classification ratio from residential rents and units only. Two firms warrant disclosure. CLPR operated in two segments: residential and commercial, with the residential segment contributing approximately 78% of FY2025 revenue (Clipper Realty 10-K 2025, p. 6). CSR was a diversified REIT until it divested its healthcare, industrial, and commercial portfolios between 2016 and 2018, after which it became multifamily-focused. Although the classification uses residential rents only, the outcome variables (FFO, margins, and valuation) are firm-level. Therefore, they include the non-residential income, a limitation discussed in Section 8.

3.2 Classification Data and Construction

We collected each firm's average monthly portfolio rent from its FY2025 10-K, primarily from the Item 2 Properties Tables. However, conventions vary: MAA and EQR report same-store figures; BRT reports unit-weighted rent including joint venture properties; UDR's rent is derived as revenue per occupied home. Metro or regional-level unit counts were taken from the Item 2

Properties Table, with the following exceptions: same-store portfolio disclosures for MAA and EQR; Schedule III for CSR.

Benchmark rents come from the FY2025 fair market rents originally published by HUD's Office of Policy Development and Research, which sets FMRs at the metropolitan-area and county level. We used the FY2025 two-bedroom FMR because it is HUD's standard reference unit. The original schedule that we use took effect October 1, 2024, making it essentially contemporaneous with FY2025 year-end rents. Where filings report broad regions spanning multiple FMR areas (e.g., "Southern California", "New England"), each region was assigned the two-bedroom FMR of its dominant metropolitan area. Several mappings required judgment beyond this rule. These include the residual buckets, "Other Expansion" and "Other Florida", which aggregate several smaller markets; and the combined markets, CPT's paired metros such as Dallas/Fort Worth, which are assigned blended values. Classification is insensitive to these judgment values: varying each across the full range of plausible constituent-market FMRs — individually and jointly — changes no firm's assignment, because the affected firms all lie well above the threshold. We identified one mapping error: Brooklyn had been assigned a non-New-York value. We corrected this, moving CLPR further below the threshold. Calculations were verified three ways. First, we independently recomputed the pipeline ratio for every firm. Second, we summed metro unit counts and checked them against the portfolio totals firms themselves report. Third, we re-verified every FMR value against HUD's published FY2025 original schedule.

$$w_{im} = \text{Units}_{im} / \Sigma_{m'} \text{Units}_{im'} \quad (1)$$

$$\text{FMR}_i^W = \Sigma_m \, w_{im} \cdot \text{FMR}_m \quad (2)$$

$$\text{RentRatio}_i = \text{PortfolioRent}_i / \text{FMR}_i^W \quad (3)$$

In the expressions above, i indexes the eleven sample firms and m indexes the geographic markets a firm reports in its Item 2 properties disclosure. The weight w_{im} is the share of firm i's total units that are located in market m. FMR_i^W therefore represents the rent the firm's portfolio would command if every unit were priced at HUD's two-bedroom benchmark in its market. At the same time, PortfolioRent_i is the firm's average monthly rent per unit. Because both the numerator and the denominator are measured in dollars per month, the ratio is dimensionless. This lets firms in different metropolitan areas be compared on a single scale, since geography enters through the denominator rather than the comparison itself.

Revenue weights were rejected because revenue is rent multiplied by occupied units. Such weights would automatically give greater weight to high-rent markets and therefore raise the weighted benchmark most for firms with the highest rents. In practice, firms disclose unit counts by market consistently but rarely disclose revenue at the market level.

3.3 The Threshold

$$D_i = 1\{\text{RentRatio}_i \leq 1.00\} \quad (4)$$

If $RentRatio_i$ is less than or equal to 1.00, D_i is classified as having a portfolio average monthly rent at or below the HUD's FMR, therefore placing it in our treatment group.

HUD's fair market rent forms the basis for setting voucher payment amounts. Hence, a ratio at or below 1.00 indicates the portfolio's average rent is at or below the benchmark the federal government uses to determine the level of assistance for that area. This line is set externally and is relevant to policy rather than a statistical result or a threshold chosen to fit the data. Nevertheless, the threshold is still a classification convention, not a legal test; a rent at or below the FMR does not mean a voucher applicant is eligible because payment standards differ from one housing authority to another. Actual eligibility is determined by program rules and unit-specific conditions such as inspections and the tenant's income.

The threshold is treated as a matter of priority rather than derived from the data, and the sample distribution does not support this; this constitutes a finding regarding the measure, not a design flaw. When ranked by the size of the gaps, the one at the statutory cut is the fourth out of ten, amounting to 0.035; this is the gap between IRT (0.966) and CSR (1.001), while the three largest gaps are found elsewhere in the distribution: CLPR|BRT at 0.112, CPT|UDR at 0.0103, and UDR|AVB at 0.056.

Table 1. HUD-Relative Portfolio Classification, FY2025

NXRT	1,492	1,930	0.773	13,305	Lower-rent
CLPR	2,050	2,580	0.795	4,100	Lower-rent
BRT	1,413	1,558	0.907	8,311	Lower-rent
ESS	2,699	2,900	0.931	63,077	Lower-rent
MAA	1,690	1,787	0.946	96,568	Lower-rent
IRT	1,593	1,649	0.966	33,758	Lower-rent
CSR	1,558	1,557	1.001	12,262	Control
CPT	2,006	1,975	1.016	58,759	Control
UDR	2,605	2,329	1.119	55,240	Control
AVB	3,062	2,607	1.175	85,107	Control
EQR	3,203	2,650	1.209	73,465	Control

Note. Ratio = unit-weighted average portfolio rent (FY2025 Form 10-K) divided by unit-weighted HUD FY2025 two-bedroom Fair Market Rent (original schedule) across each firm's disclosed markets. Lower-rent: ratio ≤ 1.00.

3.4 Classification Uncertainty and Benchmark Vintage

Several sources of error affect the numerator. Each firm's figure is a single annual figure for a portfolio that changes during the year through acquisitions, dispositions, and lease-up, so it approximates rather than exactly measures the rent earned across the units included in the unit counts. The disclosed rent averages across all unit types in the portfolio from studios to three- and four-bedroom apartments, whereas the benchmark is HUD's two-bedroom FMR specifically, so the comparison is not strictly like-for-like at the unit level, and a firm whose mix skews towards smaller units will show a lower ratio for reasons unrelated to its rent position. A further distinction separates the two sides of the ratio: HUD's benchmark is a gross rent, whereas the rents firms disclose are generally contract or base rents, which exclude utilities. The gross-versus-contract distinction is different in kind: because it applies to all eleven firms, it shifts the entire distribution of ratios downward rather than reordering firms relative to one another, with the consequence that 1.00 is not the neutral dividing line it appears to be.

To quantify this, we applied a Monte Carlo perturbation to each firm's ratio, drawing multiplicative measurement error with a 3% coefficient of variation across 20,000 replications, and recorded the share of draws in which the firm's assignment changed. CSR's probability of treatment assignment is 0.49 (effectively a coin flip). CPT is assigned to the control group in 70% of draws and IRT to the treatment group in 88%. Three firms lie within roughly three and a half percentage points of the boundary: CSR at 1.001, below the cut by less than one-tenth of one percent; CPT at 1.016; and IRT at 0.966. The eight remaining firms lie at least 5 percentage points from the boundary, with the nearest (MAA at 0.946) roughly sixty times CSR's distance. The reference unit itself is a researcher's choice rather than a property of any firm's portfolio. Rebuilding the weighted benchmark from the one-bedroom schedule leaves two firms (CLPR and NXRT) at or below the benchmark, while the three-bedroom schedule places all eleven firms in the treated group. Because no firm's rents or unit counts change across these three benchmarks, the treated count moves from two to eleven on a choice external to the firms being classified.

4. Financial Data and Variable Construction

We collected our data from firm-year financials for 2016-2025 programmatically from SEC EDGAR’s XBRL company-facts endpoint, along with year-end stock prices. Data points that couldn’t be extracted from the program were manually found in the firm’s 10-K report filed on SEC EDGAR. The program followed an accurate-or-blank policy, meaning it imputed or fabricated no values. Every data point has a source tag. For every blank cell, the pipeline creates a manual-check report instead of silently filling in data. We calculate FFO using the Nareit definition: net income available to common shareholders plus real-estate depreciation and amortization, minus gains on sales of depreciable property, plus impairments. The gain and impairment adjustments are used where XBRL tags exist and are taken as zero otherwise, so this is a derived FFO from the filings, not a company-reported FFO. Reconciliation against company-reported figures is a future refinement. The main calculated variables are: FFO yield (derived FFO over lagged total assets, used for H1); cyclical beta (the firm-level slope of FFO

yield on a common cycle factor, H1a); residual log variance after the cycle regression (H1b); FFO margin (derived FFO over revenue) and its firm-level log slope over time (H2); the P/FFO multiple (year-end price over derived FFO per share, H3); and controls for leverage (total debt over total assets), size (log total assets), and growth. P/FFO values that are outside an admissibility window of 3 to 50 are treated as invalid. This screen determines which firm-years enter H3.

5. Empirical Methodology

5.1 Small-Sample Design and Exact Permutation Inference

The panel contains 110 firm-years, but there are only 11 firms. Since observations within a firm are dependent, the firm is the real inference unit for the tests. The standard fix is to use cluster-robust standard errors, but, with only 11 clusters, they would understate uncertainty and over-reject. Therefore, conventional asymptotic inference is unreliable in this design. The study therefore uses exact firm-level permutation inference under the null and an exchangeability assumption. The treatment group is 6 out of the 11 firms, so $C(11,6) = 462$ possible assignments, small enough to enumerate exhaustively, so the distribution is exact rather than sampled. The test statistic we calculate is the treated mean minus the equal-weighted control mean, recomputed the same way under every relabeling. The p-value is the proportion of the relabelings at least as extreme as the observed assignment included. The floor for this p-value is $1/462 = 0.002$. The labels were not literally randomized; this is a permutation inference under a maintained assumption. All hypotheses share the same 462 relabelings, and a Romano-Wolf stepdown correction controls the family-wise error rate across the hypothesis family (Romano and Wolf).

5.2 H1: Cyclicality and Stability

We extract a common operating-cycle factor from the control firms' FFO yields as the first principal component, combined with the cross-firm median and oriented so stronger years are positive. Full panel extraction would let the treated firms help define the cycle they are being tested against, therefore shrinking any decoupling that could exist. We re-extract the factor within each of the 462 relabelings because the control firms change with the relabelings. Each firm’s FFO yield is regressed on the factor to get a firm-specific cyclical beta and a residual variance. H1a tests whether the mean cyclical beta of the lower-rent group is below the mean cyclical beta of the control group. H1b tests whether the mean log residual variance of the lower-rent group is lower than the mean log residual variance of the control group. A complementary COVID-19 event analysis compares the 2020 change in FFO yield across the groups with a pre-2020 placebo year as a parallel-trends check. The event contrast serves as a design anchor rather than a conventional causal difference-in-differences estimate, because the pandemic affected firms through many simultaneous channels.

5.3 H2: FFO-Margin Trajectory

For each firm, we regress log FFO margin on time, giving each firm its own projection with an intercept and slope. H2 tests whether the mean slope of the lower-rent group is greater than the mean slope of the control group. The logarithm is used so the slopes approximate proportional growth rates. The outcome is described as a margin trajectory rather than compounding, since the exponential interpretation is not separately established.

5.4 H3: Conditional Valuation

The valuation model is a Mundlak correlated-random-effects panel. Log P/FFO is regressed on the lower-rent indicator, the fundamentals that vary over time (leverage, size, growth), each firm's time-averages of those fundamentals, and year effects (Mundlak). Firm fixed effects would absorb the time-invariant classification, while Mundlak keeps it estimable and handles firm-effect and regressor correlation, preserving the coefficient of interest.
We evaluate the coefficient's robustness to omitted variables within the Oster coefficient-stability framework (Oster). The leave-one-firm-out reestimation checks whether any single firm controls the result.

6. Results

6.1 Descriptive Distribution of the Classification Ratio

The calculated RentRatios exhibit a spectrum from 0.773 for NXRT to 1.209 for EQR, with a total of six entities situated at or below the 1.00 parity line: NXRT (0.773), CLPR (0.795), BRT (0.907), ESS (0.931), MAA (0.946), and IRT (0.966). This lower-rent cohort is characterized by significant economic heterogeneity, encompassing niche Sunbelt value-add specialists like NXRT and BRT, a geographically concentrated rent-stabilized platform in CLPR, and industry giants such as MAA and ESS. Including the latter two firms shows that a lower-rent designation is not a proxy for small-cap status. A critical boundary observation is CSR at 1.001; positioned less than ten basis points above the threshold, its classification proves sensitive to HUD benchmark vintage, as explored through the reliability checks in Section 7 alongside other near-threshold firms CPT and IRT.

6.2 H1: Cyclicality and Stability

The observed cyclical-beta contrast was −0.0064, indicating that the lower-rent cohort exhibited slightly less sensitivity to the synthetic operating cycle than the control group. While this direction aligns with the decoupling hypothesis (H1a), the Romano–Wolf adjusted p-value of 0.464 renders the result statistically insignificant. Because our simulation harness confirms 75% power for this test against the calibrated effect size, this null result suggests that large-scale decoupling effects are likely absent from the population. The residual log-variance contrast (+0.096) moved contrary to the stability prediction with an adjusted p-value of 0.464; however, the 9% power rating for this component suggests that H1b remains uninformative. Although the 2020 COVID-19 event contrast (+0.010) favored the lower-rent group amid clean placebo

pre-trends, the event-study design lacks sufficient power for a confirmatory conclusion. Consequently, H1 is not supported: beta estimates rule out significant decoupling, volatility measures are indeterminate, and the pandemic-era resilience remains merely suggestive.

6.3 H2: FFO-Margin Trajectory

We identify a meaningful divergence in operating trajectories, with the lower-rent group's annual log FFO-margin slopes exceeding the controls by +0.034. This implies that lower-rent firms achieved margin growth approximately 3.4 percent faster per year in proportional terms, translating to a relative improvement of roughly 40 percent over the decade. This finding is statistically significant (adjusted $p = 0.024$), and the pipeline is well calibrated, with simulation results showing 98.7% power for this test and a low 1.3% false rejection rate. While H2 is supported regarding the margin outcome, two caveats are necessary. First, the test evaluates the final margin rather than the specific maintenance-and-turnover mechanisms, which EDGAR filings do not consistently disclose. Second, as Section 7 details, statistical strength depends on the categorical threshold: significance fluctuates when near-threshold firms are removed ($p = 0.053$) or when using a continuous ratio ($p = 0.059$). We therefore characterize this as a suggestive pattern whose robustness depends on the group-based design.

6.4 H3: Valuation Discount Analysis

The Mundlak model yields a conditional lower-rent coefficient of +0.124 on log P/FFO, suggesting a slight premium rather than a discount, though it remains statistically indistinguishable from zero (adjusted $p = 0.631$). No evidence supports a valuation discount once fundamentals—leverage, size, growth, and time effects—are held constant. The coefficient is unstable across alternative specifications; for instance, excluding NXRT or near-threshold firms shifts the coefficient into negative territory, though it never nears conventional significance levels. Given that the simulation harness successfully detects injected discounts 98.7% of the time and correctly identifies risk-justified discounts in the null control, this null result is highly informative. It suggests that if a residual valuation discount of the calibrated magnitude existed, the pipeline would have recovered it. Table 2 consolidates the real-data estimates and their corresponding simulated controls.

Table 2. Hypothesis Test Results (Real Data and Simulated Controls)

H1a cyclical-beta contrast	−0.0064	−0.0041	+0.0013
H1a adjusted p-value	0.464	0.004	0.998
H1b residual log-variance contrast	+0.096	−0.039	+1.556
H1b adjusted p-value	0.464	0.149	0.998
COVID 2020 contrast	+0.010	+0.014	−0.005

Pre-period placebo contrast	−0.013	−0.001	−0.001
H2 margin-slope contrast	+0.034	+0.018	−0.002
H2 adjusted p-value	0.024	0.004	0.607
H3 valuation coefficient (log P/FFO)	+0.124	−0.347	−0.514
H3 adjusted p-value	0.631	0.004	0.004
Verdict: H1 (stability)	Not supported	Partial (decoupling)	Not supported
Verdict: H2 (margin trajectory)	Supported	Supported	Not supported
Verdict: H3 (valuation discount)	Not supported	Supported	Not supported

Note. p-values are Romano–Wolf adjusted exact permutation p-values over all 462 firm assignments; the smallest attainable value is 1/462 ≈ 0.002. The null-control H3 coefficient is significantly negative in the raw specification but collapses once leverage is controlled, correctly identifying the injected discount as risk-justified rather than residual.

7. Simulation Calibration and Robustness

7.1 Monte Carlo Calibration of the Pipeline

Before analyzing the real data, we ran the full pipeline on two simulated data sets with known ground truth. These simulated datasets matched the real design: six treated firms and 5 control firms. Of the two simulated datasets, one is a positive control, and one is a null control. The positive control assumes lower cyclicality, lower idiosyncratic volatility, a steeper lower-rent margin slope, and a valuation discount to see whether the pipeline would detect these positive signals. The null control removes or reverses these effects to ensure that the pipeline doesn't falsely report positives. A 150-replication harness then measured the empirical size and power. The size results showed no test over-rejection, as false rejections were 0% for H1a, H1b, and H3, with H2 at 1.3%. The power results were 75.3% for H1a, 98.7% for H2, 98.7% for H3, but only 8.7% for H1b. The calibration dictates how each real result must be interpreted. H1a, H2, and H3 all rest on powered tests, but H1b is too weak for its unfavorable sign to count as evidence against stability. The weak H1b power likely reflects the instability of residual-variance estimated from only ten annual observations per firm. This component is the design's most important refinement.

7.2 Classification and Specification Robustness

We re-derived the label at cutoffs from 0.90 to 1.10 and under median and bottom-tercile splits; the margin-slope contrast keeps its sign under all nine rules and stays significant under six of them, including the median split, which produces the same partition as the statutory cut.

To test whether the findings depend on the firms whose assignment is least secure, we excluded the three firms within 5 percentage points of the threshold—CPT (1.016), CSR (1.001), and IRT (0.966)—and re-estimated the tests on the remaining eight firms. The exclusion rule is stated in terms of distance from the cut and applied before examining the results, not as a selection based on which exclusion changes the outcome. On the reduced sample, the margin-slope contrast retains its sign and magnitude (+0.031, against +0.034 in the full sample) but no longer reaches conventional significance ($p = 0.053$). At the same time, the valuation coefficient becomes negative (-0.433) yet remains not statistically different from zero ($p = 0.228$). Because the reduced sample admits only $C(8,5) = 56$ distinct relabelings, the smallest attainable p-value is 0.018.

The continuous rent ratio can be used directly, which removes the threshold from the design entirely. Inference permutes the ratio vector across firms, drawing 5,000 random orderings, which makes these p-values approximate, with a Monte Carlo standard error of roughly 0.007 rather than exact. Under this specification, no relationship is significant at the 0.05 level, with the margin-slope relationship ($p = 0.059$) closest to conventional significance and carrying the expected sign—a lower ratio associated with faster margin growth.
The finding survives moving the boundary but not removing it. The effect appears as a difference between groups rather than as a monotone relationship across the range of the ratio. Taken together with the exclusion analysis, the one positive finding in the categorical analysis is suggestive rather than robust: it holds its sign across all three specifications. Still, it reaches significance only when the threshold is imposed.

8. Discussion and Limitations

8.1 Interpreting the Findings

The results replace two common narrative accounts with a more precise picture. Lower-rent portfolios were not significantly or decoupled from the overall REIT sector's operating cycle: the beta contrast carried the predicted sign but was small, and a well-powered test could not distinguish it from zero, so modest defensive effects remain compatible with the data. Additionally, there was no conditional valuation discount. The coefficient was slightly positive, sign-unstable across specifications, and never approached significance under a test that detects real discounts almost every time, so the data do not support the stigma narrative. The one positive result is that lower-rent firms' FFO margins improved substantially faster over 2016-2025, and the test that found this result is well calibrated. However, it's an outcome difference, not a mechanism, so the data can't distinguish operating efficiencies or occupancy stability from composition effects, asset age, or market-level rent dynamics.

Additionally, its statistical strength is categorical: a difference between groups, not a relationship in the continuous ratio, which is why the paper calls the result suggestive. Failing to reject never proves that segments are identical, but calibrated power fixes what each null means.

The H1a and H3 nulls speak against effects of the injected magnitudes with reasonable confidence, while on the other hand, the H1b null is uninformative.

8.2 Implications

For investors, the results caution against assuming that lower-rent portfolios are automatically defensive or systematically mispriced. Once core fundamentals are controlled for, there is no valuation discount, and cyclical exposure looks broadly similar to the control group. The margin trajectory implies that rent positioning may carry operating information beyond just leverage and size, but it does not support an investment strategy by itself. For housing research, the study shows that "affordability" in the public multifamily sector must be measured locally. Six of the eleven listed firms, including two of the largest in the REIT sector, charge average rents at or below the federal voucher benchmark in their markets, a fact that is hidden by the REIT sector's high-end image. Methodologically, the study illustrates a reusable workflow for small public-company populations. Exact enumeration of the permutation distribution, family-wise error regulation, an explicit auditable vintage choice, and simulation controls that certify every test before real data is touched.

8.3 Limitations

The limitations are stated candidly: The study only has eleven firms. The classification is relative, not statutory, as it doesn't establish subsidy participation, tenant eligibility, or voucher acceptance. The FY2025 classification may not reflect each firm's positioning throughout 2016-2025, and firm-level averages don't fully capture variation across properties, units, and tenants. Firms also disclose rents differently (effective versus average versus same-store). HUD's benchmark includes utilities, but company rents mostly don't, which pushes every ratio down. The bedroom benchmark also matters significantly for the classification: 1BR rents label only 2 firms as lower-rent, while 3BR rents label all 11 firms. CSR's assignment flips under HUD's revised FY2025 benchmark. FFO is calculated from the pipeline rather than pulled from an official report, so it may not exactly reproduce the company-reported FFO. Firms choose their markets, leverage, and acquisition strategies, so no coefficient here is a randomized causal effect. The panel-internal cycle factor is estimated from only five control firms and may therefore be unstable; the pandemic affects firms through many channels; and the H1b calibration indicates very weak power. Finally, defining the universe as of FY2025 excludes firms acquired mid-window, creating a potential source of survivorship bias.

9. Conclusion

Lower-rent portfolios are frequently assumed to be defensive or mispriced. Yet these assumptions have been difficult to test because publicly traded multifamily REITs are not classified by rent relative to the local market in which they operate. This study developed a classification grounded in HUD's published Fair Market Rent schedules, each firm's disclosed geographic unit distribution, and an auditable benchmark vintage. The study also evaluated three

associated propositions using exact permutation inference and a simulation-calibrated pipeline. The evidence from the tests does not support cyclical decoupling or a valuation discount, as the nulls in both are well powered. The study does reveal a faster FFO-margin trajectory for the lower-rent firms. This economically meaningful pattern holds its direction across every specification examined but is statistically significant only under the categorical design. Therefore, this finding is suggestive evidence warranting replication, not an established regularity. Readers should not conclude that rent segments are identical, nor that lower-rent portfolios consistently outperform. They should conclude that the informative tests found no defensive or mispricing advantage for the lower-rent group, while the margin results are the most encouraging direction for further work. Future researchers should extend the classification historically with property-level rent data, incorporate tenant-income measures, change the derived FFO to reported figures, and use a larger universe that incorporates private and nontraded owners. The evidence from the study therefore doesn't establish a distinct lower-rent premium, discount, or defensive operating profile. However, it provides a framework to test those hypotheses more accurately as richer property-level data becomes available.

Works Cited

AvalonBay Communities, Inc. Form 10-K for Fiscal Year 2025. U.S. Securities and Exchange Commission, EDGAR, 2026.

BRT Apartments Corp. Form 10-K for Fiscal Year 2025. U.S. Securities and Exchange Commission, EDGAR, 2026.

Camden Property Trust. Form 10-K for Fiscal Year 2025. U.S. Securities and Exchange Commission, EDGAR, 2026.

Centerspace. Form 10-K for Fiscal Year 2025. U.S. Securities and Exchange Commission, EDGAR, 2026.

Clipper Realty Inc. Form 10-K for Fiscal Year 2025. U.S. Securities and Exchange Commission, EDGAR, 2026.

Equity Residential. Form 10-K for Fiscal Year 2025. U.S. Securities and Exchange Commission, EDGAR, 2026.

Essex Property Trust, Inc. Form 10-K for Fiscal Year 2025. U.S. Securities and Exchange Commission, EDGAR, 2026.

Independence Realty Trust, Inc. Form 10-K for Fiscal Year 2025. U.S. Securities and Exchange Commission, EDGAR, 2026.

Mid-America Apartment Communities, Inc. Form 10-K for Fiscal Year 2025. U.S. Securities and Exchange Commission, EDGAR, 2026.

Mundlak, Yair. “On the Pooling of Time Series and Cross-Section Data.” Econometrica, vol. 46, no. 1, 1978, pp. 69–85.

Nareit. Nareit Funds from Operations White Paper – 2018 Restatement. National Association of Real Estate Investment Trusts, 2018.

NexPoint Residential Trust, Inc. Form 10-K for Fiscal Year 2025. U.S. Securities and Exchange Commission, EDGAR, 2026.

Oster, Emily. “Unobservable Selection and Coefficient Stability: Theory and Evidence.” Journal of Business & Economic Statistics, vol. 37, no. 2, 2019, pp. 187–204.

Romano, Joseph P., and Michael Wolf. “Stepwise Multiple Testing as Formalized Data Snooping.” Econometrica, vol. 73, no. 4, 2005, pp. 1237–1282.

UDR, Inc. Form 10-K for Fiscal Year 2025. U.S. Securities and Exchange Commission, EDGAR, 2026.

U.S. Department of Housing and Urban Development, Office of Policy Development and Research. Fair Market Rents for the Housing Choice Voucher Program, Fiscal Year 2025. HUD, 2024, www.huduser.gov/portal/datasets/fmr.html.

U.S. Department of Housing and Urban Development, Office of Policy Development and Research. Fair Market Rents for the Housing Choice Voucher Program, Fiscal Year 2026. HUD, 2025, www.huduser.gov/portal/datasets/fmr.html.

U.S. Securities and Exchange Commission. EDGAR Company Facts API (XBRL Financial Data). SEC, 2026.

College Debt: Institutional Factors By Lalith Durbhakula

Abstract

This research paper examines how institutions are directly impacting student college debt in the United States while also investigating various student factors that influence the accumulation of student debt – this ranges from the type of college that the student is attending to the overall financial literacy of students. Throughout this paper, for purposes of conducting research, each section contains a comprehensive literature review while critiquing existing research. This allows for a cohesive understanding of each section using existing data while providing insight into the limitation of such sources and how to fill in any missing information that would prove valuable towards expanding this field of research.

Public education in the United States has successfully addressed gaps due to wealth by making it accessible for anyone residing in the 3.8 million square mile country. However, as students grow into the next phase of their life, disparities unconventionally arise among individuals, most of which are associated with college debt. As a result, it is imperative to find a way in which college debt has minimal impact on society.

To start, this paper reviews why college debt has detrimental effects on students themselves and on their future. By further acquainting the reader with the notion of college debt, this section serves as a solid foundation for the analysis presented in this paper. Previous research indicates that most effects are primarily financial, with others being related to personal matters such as family. In addition, students' mental health levels are compromised with some reporting incidents of experiencing chronic stress.

Next, this paper assesses the correlation between attending for-profit colleges and the average debt level of graduates in comparison to non-profit colleges in the United States. By understanding how college debt statistically varies between diverse types of colleges, one can make an educated financial decision on their future college. Prior research and data suggest that private nonprofit colleges have the highest debt amount among college students, but private for-profit colleges have the highest debt rate among college students.

Afterwards, this paper examines how the financial literacy of students entering college directly impacts their expected salary post-graduation. In most schools, a commonly overlooked course is fiscal management – undoubtedly, students must oversee their own wealth as they progress through life and obtain higher responsibilities. Anterior research depicts that students who were willing to cope with more debt were projected to have higher annual salaries post-graduation.

Finally, this paper analyzes the effects of receiving misleading information about loan terms on the repayment success rates of students from technical schools in the U.S. Preventing the spread of misguided information is crucial in ensuring that a student is fully aware of all the assistance that they could receive financially. Earlier research shows that from the late 1990s/early 2000s to 2015, small colleges which were previously owned by big incorporations

have been misleading students regarding career prospects, transferability of credits, and program duration – this led to the intervention of the Department of Education (ED).

An area that current research fails to include with regards to college debt is substantial datasets that account for the direct input of the human population. Although there exist a few datasets on college debt, most of them merely discuss the dispersion of student debt among undergraduates, graduates, and postgraduate students; while utilizing this data, this paper also aims to explain factors that correlate with student debt. Therefore, it would be beneficial to publish mass surveys that ask students about their opinion on factors such as financial literacy and how it can influence college debt.

Effects of College Debt

College debt has countlessly been proven to deteriorate mental health across students of all ages and is linked to altering their decision-making process after college. Despite COVID-19, the national student loan debt has increased from $52 billion in 2006 to $1.73 trillion in 2023, which is an average of $77.5 billion increase per year – the average monthly student loan payment is $390 as of 2026 (Hanson). Most students enter college with the prospect of long-term success; instead, they leave with an abundant stockpile of debt and an unsettling future.

During college, depending on the loan, the monthly payments can be exorbitant, and bills can quickly amass. This proves difficult especially for those students who have never managed money before – in addition to keeping up with monthly payments, the student must juggle coursework, extracurricular activities, and personal budgeting with the remainder of their money. Thus, many students seek part-time jobs in college which they are not necessarily fond of, leading to decreased attention towards academics. A survey conducted by the American Student Association concluded that half of graduates attributed college debt to their lack of freedom when choosing a place to work (Porter). Another study conducted by the National Association of REALTORS in 2021 shows that 58% of student loan debt holders had to work at jobs that they were not interested in to pay off student loans (Realtors). In addition to affecting students in college, student debt can significantly grow beyond college.

Despite graduates acquiring numerous degrees and certificates, they often have a negative net-worth because of ongoing payments towards their student loans. Some students in college aim to prioritize academics over financial stability: as they wind up and graduate, they are left with a massive dip in their net-worth compared to other students. This can lead to inconvenient hurdles if the student wants to take out any more loans or if they want to make large purchases. Therefore, it takes a considerable amount of time before student borrowers financially overtake their peers who have never encountered debt. Even if the student manages to get accepted for further loans, it is unlikely that they are willing to do so (Porter).

One application of loans is buying a home through taking out a mortgage. The prospect of buying a home, however, is particularly dreadful in the minds of student borrowers. Porter notions that due to the existing burdensome pile of college debt that student borrowers are handling, a negative correlation emerges between student loan debt and homeownership rates.

Unfortunately, most students are unaware that owning a home can contribute toward building and financing their personal wealth: this causes a hindrance in the ability for a student borrower to gain wealth. The study from REALTORS indicates that 29% of student borrowers associate debt as a reason for their inability to purchase a home. Without owning a home, it becomes even more challenging to start a family (Porter).

To raise the average child, it costs $1100 per month – thus, most students in debt steer away from raising a family until they are financially stable (Porter). Raising a child not only requires long-term financial stability; it also demands tremendous discipline and chaotic stress, especially during the infant years. Furthermore, REALTORS have determined that one-seventh of all student borrowers link college debt to delaying the process of raising a family. With many borrowers deterring from raising a family, it is evident that there lies an underlying factor of stress when making choices based on student debt.

For student borrowers, stress is primarily triggered because of unpaid student loans. However, this principle extends to other types of debt, such as credit or medical. An article written by *The Practice* at Harvard Law School suggests that external stress from day-to-day life can cause unsecured debt, and debt can cause internal stress within the body, therefore indicating an anatomical correlation between debt and debtholders. Moreover, the article explains that the mental repercussions are equally as overwhelming – for the majority of young borrowers, "the researchers determine that those with consistently high debt 'report 21 percent higher depressive symptoms than respondents who report no or low debt across the life course' " (Fong et al.). Prolonged stress can lead to considerable impact on one's financial life.

Student borrowers who have issues with repayment have a tremendous dip in credit score compared to other students who do not acquire debt. Furthermore, this can lead to aforementioned problems such as purchasing a house or starting a family. With regards to credit score, "The average person with student loans has a credit score of 656, but the average for the entire United States is 711" (Porter). This ubiquitous gap in credit score implies that it can take a significant amount of time before student borrowers regain their credit.

Many families are aware of these insurmountable problems beforehand; thus, they prevent their children from attending college due to the apprehension of taking on debt. However, this is suboptimal for the education sector – as Porter emphasized, "If college debt were less of an issue, more people might be willing to pursue higher education, increasing their lifetime income and building a more educated workforce."

The short-term and long-term effects of accruing student loan debt are debilitating for a young college student – therefore, it is necessary for incoming college first-year students to do extensive research on their selected college. Additionally, they should be financially well-equipped to combat debt.

For-profit vs. Non-profit Colleges

As students wrap up their final year of high school, choosing the right college to attend can have a substantial impact on their overall financial stability. A main criterion is

understanding the distinction between the operation mechanisms of for-profit and non-profit colleges. For-profit colleges have been historically known to aggressively recruit students, while non-profit colleges rely on student admission fees for funding. By comprehending both options, one can make an educated choice on their prospective college and look after their financial well-being.

For-Profit Colleges

To start, more than 80% of 2016 graduates from for-profit colleges had student loans, compared with 68% of those from private nonprofit colleges and 66% from public colleges. This is correlated with the recruitment tactics of for-profit colleges – prospective students are constantly coaxed by these colleges to sign up for a program. In most scenarios, the student is told that they have a higher likelihood of getting accepted for a job by completing a certain program. Afterwards, to finance the program, the college will 'assist' the student by encouraging them to take out student loans – in reality, the student is unaware of the complications of high-interest loans .

Moreover, colleges are fully aware of the educational awareness of various communities – thus, these for-profit institutions primarily target Black and Latino students as a means of generating revenue. These minorities are more likely to be unacquainted with economic management; this leads to the for-profit colleges making profits out of admittance while the student is tumbling in debt. Similarly, the GI Bill pays private colleges $25,000 per veteran enrollment, making veterans susceptible to aforementioned recruitment schemes.

Due to these conscious enlistment strategies incorporated by for-profit colleges, there must be a comparison between taking out loans at nonprofit colleges compared to for-profit colleges. Indeed, a study conducted in 2020 by The Institute for College Access & Success (Schak et al.) concluded that "Even among bachelor's degree recipients, 30 percent of those who started at for-profit colleges defaulted on their federal student loans within 12 years of entering college … six times the rate of those who started at nonprofit colleges (5%)." Defaulting on student loans can prevent graduates from getting a job primarily because their wages get garnished, and their credit score is extraordinarily depleted.

Non-Profit Colleges

In contrast, nonprofit colleges are not as demanding as for-profit colleges, yet they manage to accumulate more cumulative debt among students. Compared to public colleges, private nonprofit colleges typically have higher admission fees because they are externally funded through private grants and internally funded through student fees. As they are private, the government does not financially assist them; hence, a considerable amount of money is streamlined from student fees in order to maintain the college (Schak et al.).

To minimize debt acquisition, nonprofit colleges often implement various strategies when considering recruits. Students that attend private nonprofit colleges pay the net price; that is, they pay the sticker price minus any scholarships or discounts. This allows for two key areas of focus.

The first contributing factor is the individual wealth of the student. If the student is from a high-income family, they will be able to afford the sticker price without any scholarship – consequently, some nonprofit colleges primarily select students from wealthy families and the minority of the students are from families that have low to moderate incomes. Another way that nonprofit colleges compromise for students is by offering financial aid and scholarships – this way, they reduce the disproportion between students from upper-class families and lower-class families by enrolling more students in lower and middle-class families. In either case, graduate students tend to have lower debt levels (Schak et al.).

As there are a wide variety of techniques to enlist students, cumulative student debt levels fluctuate between individual colleges. For instance, the TICAS report discovered that " Two-hundred and twenty-nine nonprofit colleges (34%) reported that more than 75 percent of their 2020 graduates had debt, but 119 colleges (18%) reported less than 50 percent of their graduates with debt." Despite methods to alleviate debt, nonprofit schools manage to have the most debt across all types of colleges. Graduates from public institutions graduate with an average of $26,100 in debt, graduates from private universities graduate with an average of $29,000 in debt, and graduates from private nonprofit colleges graduate with an average of $35,000 in debt (Quintana).

Data Analysis

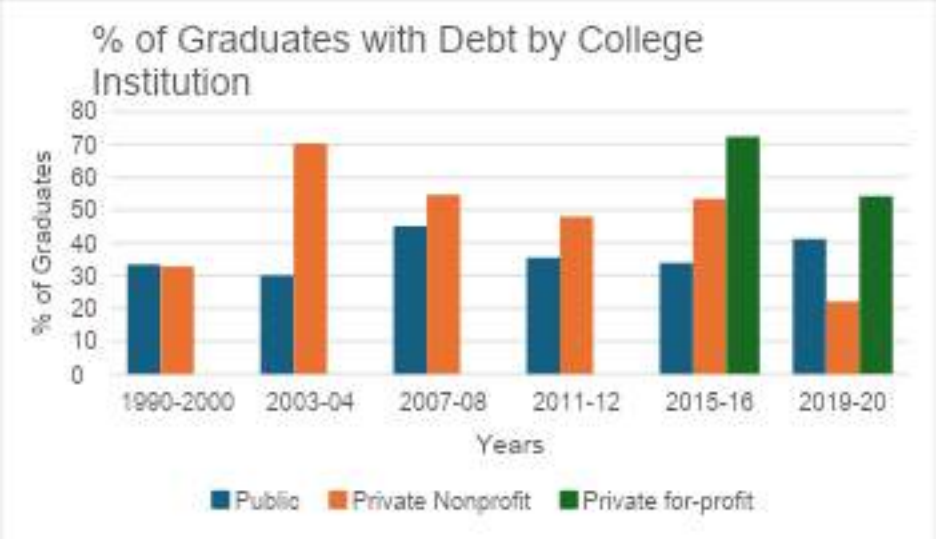


Figure 1.1: Percentage of College Graduates Completing Degrees with Student Loan Debt by Institutional Control, 1999–2020

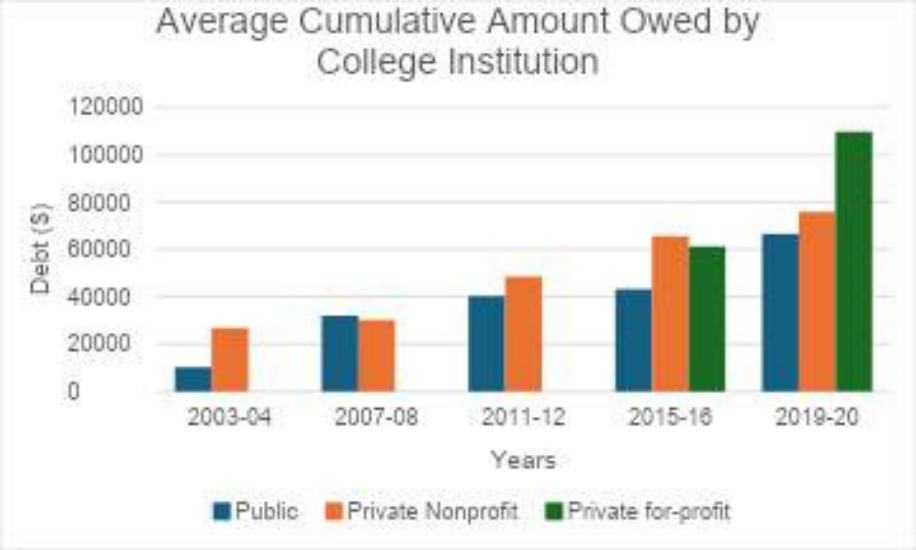


Figure 1.2: Average Cumulative Student Loan Debt Owed per Graduate by Institutional Control, 1999–2020

The two graphs were adapted from the National Center for Education Statistics (NCES). In figure 1.2, debt for graduates at a nonprofit college had an average increase of $3069.68 per year, which contributes to $12,278.72 per four-year degree program as depicted in the graph. Across all three institutions, the average loan amount owed increases as the graph approaches the present – reasons vary by institution. At public colleges, state funding has decreased by 38% from 1980 to 2020; Sean Ruddy, the project associate for the Bipartisan Policy Center explains that "Over the past few decades, however, state support for higher education has gradually declined as a result of tax cuts, competing priorities, and the impact of recession on state finances." This explains the gradual increase in graduate debt at public colleges. Moreover, graduates at private for-profit colleges had 79% higher loan amounts in 2019-20 compared to 2015-16. At the same time, for-profit colleges experienced a rise in enrollment during the pandemic, which can be linked to the surge in graduate loan amounts (Ott). Finally, for-profit companies opted to withhold statistics from 2003-2012 during this survey. Their participation would highlight unusually elevated levels of debt; since most for-profit colleges heavily rely on student admissions and investors, displaying such data would tend to damage their reputation and deter prospective investors.

In figure 1.1, the correlation between the time period and the percentage of graduates who have debt in public colleges appears to be sinusoidal as it closely maps the function $f(t) = 9 \sin sin\left(\frac{2\pi}{13}t + \frac{11}{2}\right) + 35$, where t is the year and $f(t)$ represents the percentage of graduates with debt at public colleges ($\sin sin\,(x)$ is evaluated in radians). However, more data needs to be given to accurately describe the trend. In addition, 20-40% of graduates consistently tend to have debt in public colleges. The relationship between the time period and the percentage of graduates who have debt in nonprofit colleges has a strong correlation to the quartic regression $f(t) = -0.003t^4 + 26.32t^3 - 79207.05t^3 + 105939059$ ($R^2 = 0.9896$), where

t is the year and $f(t)$ represents the percentage of graduates with debt at public colleges. There is insufficient data to describe a pattern regarding for-profit colleges and the percentage of graduates with debt.

Critiquing Research

The sources that are used for this subtopic thoroughly present surface-level findings when inspecting data. Most of the findings are presented in the form of a percentage of a quantity or a description of a percentage increase or decrease between two correlative points. For instance, the U.S. News mentioned that the net price of college for students from poorer families declined by 2% from 2003-04 to 2015-16 in nonprofit colleges, but it rose by 36% at for-profit colleges during this time. In addition, these sources cite TICAS or NCES for obtaining data.

The 2020 TICAS report and the Education Data website primarily refer to external sources rather than conducting their own research – like other sources in this paper, their findings are widely displayed in the form of percentages where the change between two data points are evaluated. An Education Data website will have an average of 17-25 sources, whereas a TICAS comprehensive report will have 90-100 sources in their works cited.

In all the websites used for this subtopic, some links that are directed to the website's sources within the site are faulty – the page is either removed or does not exist altogether. This indicates that the author/editor does not regularly check the status of the sources that they cite in the article.

The larger issue arises in the way that the research is analyzed. The sources' findings require basic arithmetic and can be visualized by most individuals. In most of the reports, whenever a graph is provided, the report will repeat statistics that can be found on the graph – otherwise, they will merely state the percentage change between two time periods or two relevant data points (e.g., a 1.7% decrease in national student loan debt from 2022 to 2023).

The sources fail to provide predictions and trends when analyzing their data. For instance, in the earlier analysis provided, one can clearly understand the trend between the time period and the percentage of graduates that have debt. In addition, based on the function that was given, one can make a relatively accurate estimate for future data. Such depictions are not seen in any report.

Moreover, all of the sources for this section have a root source – NCES. This means that there is only one governmental website that provides firsthand surveyed data on the correlation between attending for-profit colleges and the average debt level of graduates in comparison to nonprofit colleges in the United States. Therefore, further publicly available data on this topic would be highly beneficial for incoming college first-year students as deciding whether to attend a for-profit college or a nonprofit college is a crucial decision to make.

Conclusion

When selecting a potential college to attend, it is indispensable to understand the differences between the strategies incorporated by for-profit colleges compared to nonprofit

colleges, thereby learning the assorted reasons for which debt is accumulated depending on the type of institution.

For-profit colleges are established based on the motive to make profit from students – earnings from tuition rates are split between shareholders and investors at the college. Thus, for-profit colleges aim to enroll as many students as possible. In doing so, these for-profit colleges target minorities that are unaware of the complications of high-interest loans; the colleges inform these minorities that they can acquire a job if they complete a certain program in 3-6 months. Consequently, the college profits from these abundant enrollment fees while the individual student accumulates debt. A greater percentage of for-profit college graduates have debt compared to nonprofit college graduates.

Nonprofit colleges are created to redirect any earnings towards institutional needs. Rather than admitting as many students as possible, nonprofit colleges tend to raise tuition fees in order to keep up with rising maintenance costs. However, nonprofit colleges are more aware of the debt crisis and are implementing ways to mitigate student debt. One strategy that colleges incorporate is mainly recruiting students from high-income families; the student will be able to afford the sticker price and is unlikely to aggregate debt. Some nonprofit colleges manage to proportionately recruit students based on income level; they are able to compensate by providing financial aid to the lower and middle-class students. Non-profit colleges usually have more cumulative debt compared to for-profit colleges, but in 2019-20, graduates had a 30% higher average cumulative debt at for-profit colleges than nonprofit colleges.

Further research should include descriptions of trends that happen within the data – rather than analyzing the relationship between two data points, the report should assess the overall data and use algorithms to deduce a recurring pattern. Moreover, this allows for predictions to be made – given hundreds of thousands of lines of data and a general trend, an accurate prediction can be made for future reference. However, this becomes problematic because there isn't enough data specifically regarding debt at for-profit colleges – they prefer to not participate in public surveys that would damage their reputation.

Therefore, it is possible to conduct more in-depth research on nonprofit colleges. A breakdown of college debt per individual nonprofit college would be beneficial because one can make geographical connections between certain colleges that have high or low amounts of debt. Moreover, the display methods (such as Excel files) on governmental websites such as NCES should be updated to the latest versions to present the findings cleanly. Finally, different government agencies should conduct these surveys to enhance diversity in selections and findings.

Student Financial Literacy

Regardless of the type of college that one attends, "42% of students find themselves grappling with credit card debt." Knowing the right assets at one's disposal when trying to pay off a loan differentiates those who are stuck in debt crisis and those who can handle high loan levels and responsibly finance their money. By outlining the significance of being financially

literate with regards to money management, one can properly educate themselves and get to have a debt-free college experience.

Importance of Financial Literacy

Financial literacy is about having the basic understanding of requisite financial principles in order to be financially stable for the future. Being financially literate offers a plethora of benefits to one's economic profile; unfortunately, it is not effectively taught in the school curriculum.

Financial literacy empowers individuals to tackle financial challenges and be responsible with their financial decisions by providing them with essential fiscal management knowledge. These individuals are better equipped to make sophisticated decisions regarding any payments that require a sum of money to be repaid, such as student loans. The students are mindful of the different options, aid programs, and negotiation tactics when taking out loans.

Since these students are financially knowledgeable, they are more aware of poor economic management which can lead to considerable drawbacks such as student loan debt. Thus, these students know how to efficiently budget their money to gather all their payments on time. This ensures that they do not succumb to omnipresent challenges faced when handling high-interest loans in college (iGrad).

In the event that the student faces financial hardship such as being fired from a job, they are able to quickly recover and discover other methods to control their student loan debt. Moreover, this implies that the student is capable of making independent choices and has decreased their reliance on family altogether. This is crucial for college students because they will be faced with independent financial hurdles; oftentimes their parents will not be available to support them (iGrad).

A frequently undervalued benefit of being financially literate is "its direct connection to graduation rates" (iGrad). Stress is a copious problem that causes students to inadequately perform on academic tests; contrary to popular belief, most students obtain external stress from financial apprehensions. Therefore, if students can appropriately manage their money in college by acquiring suitable financial education, they are bound to perform comparatively better than their peers who lack economic management skills (iGrad).

Further Implications

When students learn about finance to prepare for potential challenges that could be faced in college, students repeatedly study common pitfalls that lead to student loan debt. However, "When it comes to student financial literacy, it's essential to recognize that it goes far beyond just understanding student loan debt" (iGrad). Aspects such as budgeting and retirement planning are equally important when being financially literate.

When students are budgeting in college, they must manage their own money; thus, it is fundamental to learn how to proportionately divide earnings, expenditures, and savings. Not only will this be beneficial during college, but it is vital for the rest of their lives. The inability

otherwise could lead to bad habits in college such as spontaneously plunging into student loan debt that will continue into the student's future, making it difficult for one to take out any loan (iGrad).

Financially literate students are conscious of revenue and income sources that can easily fund a loan payment. Since most students end up taking a loan, they should learn ways to stay on track with monthly payments. This includes earning scholarships and grants, or it can result in a long-term approach such as working in a job. Such education at a young age will foster better administration of one's money as it opens up opportunities to expand revenue streams in the future (iGrad).

College is the time when most students are introduced to credit cards as they must authorize their own payments. Hence, students who are taught how to make informed choices on "topics like credit scores, credit reports and responsible credit card usage" (iGrad) will lead to an abundance of openings financially and can avoid any damaging consequences to their economic status. By having stable credit reports and satisfactory credit scores in college, a student is prepared to take risks in the future if necessary.

Although retirement feels far-off from college, according to iGrad, "Only 40% of non-retirees thought their retirement savings were on track." Another significance of college students being financially well-educated is that it will give them an advantage when planning for retirement compared to their contemporaries who likely have no comprehension of fiscal management. Financially literate students will intelligently oversee their investment, net worth, and real estate, all of which are invaluable towards retiring early (iGrad).

Analysis

The financial status of a student has been proven to be directly correlated with retention and graduation rates in college. As reported by iGrad, "less than two-thirds of U.S. college students will finish their degrees and half of all college dropouts will do so due to financial concerns."

Returning to stress, it can quickly stockpile and interfere with a student's day-to-day life; managing student debt while maintaining a constant source of income leads to decreased attention to studying. Over time, this can lead to the incompleteness of a student's prospective degree and can cause them to change their field of study entirely (iGrad). A study conducted by Inside Higher Education and College Pulse concluded that 40% of students selected that balancing school with other obligations was one of the top three factors contributing to academic stress. The same article discusses that stress can affect a student's academic performance by pressuring them to do well on tests, group projects, or authoring papers.

Figure 1.3: Student Willingness to Accumulate Debt by Expected Post-Graduation Starting Salary Category

Willingness to Accumulate Debt by expected starting salary categories**

	Willingness to Accumulate Student Debt						
Expected Salary ($)	$1-$9999	$10000-$19999	$20000-$29999	$30000-$39999	$40000-$49999	$50000+	Don't know
0-29999	23.50%	18.80%	15.80%	9.50%	6.90%	9.50%	16.10%
30000-39999	18.30%	18.80%	16.70%	12.10%	8.50%	12.80%	13.00%
40000-59999	16.20%	18.20%	16.20%	12.70%	9.30%	15.40%	12.00%
60000-79999	14.20%	17.00%	15.20%	9.70%	10.80%	21.20%	12.00%
80000+	11.90%	11.80%	14.40%	9.50%	8.50%	28.00%	15.90%

This data was adapted from the National Student Financial Wellness Study (NSFWS) from the Ohio State University (OSU); 19000 students were surveyed, and the chart depicts the percentage of students who are willing to accumulate a certain amount of debt based on their projected income after college. Clearly, the trend is a positive correlation between expected salary and acceptable student debt – the yellow highlighted boxes represent the highest percentage in each row for expected salary. Specifically, 28% of students at OSU who were expected to earn $80,000+ annually after graduation were willing to accumulate $50,000+ to finish their degree. In contrast, 10% of students projected to earn less than $30,000 annually were willing to accumulate $50,000 to finish their degree. So, one can conclude that students who are likely to earn more are financially responsible and can oversee larger amounts of debt, whereas students who cannot manage loads of debt are unlikely to earn as much.

Solutions

The only way to be certain that incoming college first-year students have a sound understanding of finance to be successful is to incorporate it into the high school curriculum. By making students financially literate, they have a much higher probability of being successful in college and general life.

When considering how to implement financial courses into everyday school curriculum, it appears that finance and math go hand in hand. Common core high school mathematics courses are rigorous enough for an average high school student to complete basic financial education and including it in a math unit would be sufficient for comprehension. An alternative option would be to make financial classes separate from school classes and to make them mandatory for students in 11^{th} or 12^{th} grade. Therefore, it would be overwhelmingly advantageous for those students who are financially literate on an everyday basis.

After choosing how to add classes to the high school curriculum, schools must invest in faculty that can efficiently teach financial topics to other students. Classes could have two instructors: one teaches the main course, while the other teaches the financial curriculum and makes it engaging for students. If two instructors are not affordable, the teacher for that course must be financially literate themselves to ensure that they can deliver the content promptly and thoroughly.

If possible, schools should aim to teach student populations that are prone to facing financial hurdles in college. These include “older and other non-traditional students, low-income students, first generation students, and students of color” (FLEC). By utilizing institutional data on former students that experienced financial hardships and impacting their quality of education and international data on the financial well-being of foreign students, schools can make an educated choice on the types of students that need more outreach and assistance when it comes to being financially viable (FLEC).

Although schools have a crucial gap in financially educating students, colleges must take immediate action to ensure that students are aware of their financial requirements and necessities. Colleges should make the cost of attendance explicit to students; by giving them the

upfront cost of attending college, their respective families can decide whether it is affordable. If students decide to go, the college should state the distinctions between financial aid offers such as scholarships and job opportunities. Moreover, institutions should categorize critical details within the financial aid section. For instance, when explaining college grants, institutions should state "the term (such as one-year renewable) and conditions (such as minimum credits and/or grants)" (FLEC). Finally, the college should walk the student through the financial obligations that must be executed. Directions on enrollment offers and "information on how to contact the financial aid office and other sources for additional assistance and resources" (FLEC) will phenomenally support students when applying for financial aid offers and makes it simple to understand the overall financial expectations for an institution.

Critiquing Research

The sources used for this section primarily present their own interpretations and evaluations of other data sources with the exception being the survey conducted by OSU. IGrad is a blog that emphasizes the goal of making every student financially literate. In addition, it offers its own personalized financial courses for students to take. The U.S. Financial Literacy and Education Commission (FLEC) consists of many governmental departments and agencies that write reports and work to improve financial literacy and education for all Americans.

Additionally, most sources explain that other surveys mainly focus on the correlation between financial literacy and stress levels; expectedly, findings reveal that financial problems are one of the largest factors of stress and financially literate students tend to have less stress on a daily basis. However, there are limited studies on the financial literacy of individuals – this can be evaluated by asking how one budgets their money based on spendings, savings, and investments. Another way to assess one's individual financial literacy is by asking them about their borrowing decisions and their understanding of repayment at mid-sized public universities. Although it logically makes sense that financially literate students would have more expected earnings, it would be invaluable to conduct similar studies to OSU on a larger scale.

Conclusion

The reason behind the vast majority of students attending college is to enrich themselves in their field of study and to find a guaranteed way of securing a job. Therefore, it is substantially favorable for students to understand the significance of being financially educated with regards to earning a good salary.

Being financially literate from a young age seldom offers benefits towards acquiring opportunities later in one's career. Within the scope of college, students can efficiently make informed decisions regarding credit card debt or student loan debt. They are wisely observant of any aid programs or repayment options available when handling money. On the other end of the spectrum, they are aware of traps and bad habits that financially uneducated students tend to fall into, such as procrastinating monthly payments. Moreover, they are anticipated to earn more due to the direct correlation between graduation rates and financial literacy. Financially educated

students have superior academic performances because they are less subject to economic burdens and can focus their attention towards studying – this leaves them more likely to graduate and earn higher starting salaries.

College students primarily focus on student loan debt and credit card debt when learning financial management, but to ensure that college students are sufficiently and appropriately being financially educated, students must learn how to budget their money when they earn it. Not all the money can be put towards loans, but not all of it can be spent egregiously – it must be equitably divided between savings and spendings which can be filtered into further subcategories. Furthermore, it is essential to learn available part-time jobs that a college student can apply for during college – students do not educate themselves adequately in this aspect of finance, which is why they miss out on opportunities to attain occupations that pay higher salaries. By combining these procedures, one can successfully plan their retirement from as early as college.

Studies from OSU have concluded that there is a positive correlation between one's expected salary and their willingness to accumulate debt. This implies that the students who are projected to earn more are financially literate as they are inclined to accumulate more debt to complete their college degree.

To financially educate students, high schools must combine financial curriculum and common core curriculum. Schools can make finance a separate course, or it can be absorbed into the math coursework – they must aim to educate minorities as they are heavily underserved in the economic sector. Either way, it should be obligatory to learn finance in 11^{th} or 12^{th} grade – along with generating the content, schools should devote some time to acquiring financial experts that can effectively teach the curriculum to students. Simultaneously, colleges can make it easier for students to understand repayment options by giving the forthright cost of attendance; moreover, they can supplement their website with financial aid options or any other discounts that students can acquire.

Further research should be conducted similar to that of OSU but on a greater scale; by having national data on one's expected earnings and their debt levels, accurate trends can be made to assess and determine an individual's financial literacy rate. Non-statistical data can be surveyed through a series of questions that refer to borrowing methods and repayment options; another way to evaluate a student's financial education is by gauging one's understanding of these topics. Essentially, the more data that can be specifically towards financial education as opposed to external factors such as stress, additional research can be accurately collected.

Spread of Misinformation

Oftentimes, students look for information directly from colleges and universities when they are doing general research on their chosen institution. This is publicly available on the college's website, or it can be provided with the help of close affiliates. Problems arise when the institution gives false information regarding coursework, payments, or job opportunities that

would otherwise seem appealing to an incoming first-year college student. Therefore, it is vital that accurate and up-to-date information is presented by colleges.

ED Responses

On July 9, 2021, the Department of Education (ED) forgave $55.6 million towards student loan debt – this movement helped relinquish students who were financially misled under three schools: Marinello Schools of Beauty, the Court Reporting Institute, and Westwood College.

The majority of the claims appear to be from Westwood College. For thirteen years, Westwood College has been wrongfully informing students that they cannot transfer credits to other colleges or institutions, implying that they would have to restart their education if they did not choose to go to Westwood College. During the same time period (2004-15), Westwood College was also accused of falsely assuring students that completing their criminal justice program would make them eligible for a position with the Chicago Police Department. Contrary to these assumptions, the Chicago Police Department did not hire graduates from Westwood College. Moreover, these students would end up working at jobs that never required a degree, most of which pay minimum wage. The ED has provided exact numbers: "The Department has approved over 1600 claims, representing approximately $53 million in relief for former Westwood students." They also mentioned that "Westwood College was owned by Alta College, Inc. (Alta), which was located in Colorado."

The Marinello Schools of Beauty have significantly misguided students on the content that they would learn and the quality of teaching. Most students report a lack of curriculum engagement and depth. It has also been unearthed that instructors had been abandoning students because Marinello could not sufficiently provide for students. Therefore, most students were unable to pursue cosmetology as they could not acquire the necessary official documentation. "As of today, the Department has approved over 200 claims, representing approximately $2.2 million in relief for former Marinello students" (ED).

From 1998 to 2006, the majority of students at the Court Reporting Institute (CRI) were misled into thinking that ample time would be given to complete the court reporting program. Consequently, 94-98% of students could not complete the court reporting program. "As of today, the Department has approved 18 claims, representing approximately $340,000 in relief for former CRI students" (ED).

Costs

For many families, the cost of going to college can be burdensome; thus, it would be useful to clearly publish the cumulative cost of college. Along the same lines, in order to administer universities, Trump wanted them to publish data on their average expenditures and other related factors such as admissions. Most institutions complied and made the statistics publicly available; unfortunately, most colleges provide erroneous links and fail to check their authority. For instance, the link redirected by ED on the cost of attending the University of St.

Francis in Illinois prints the message "page can't be found." Moreover, the College Scorecard is the primary website used for individuals to evaluate the cost of college – in 2019, it was unearthed that this website uses falsified data from colleges that rarely solidify them. Concerningly, the University of Oklahoma admitted to inflating alumni giving figures for two decades, highlighting the broader issue of data misrepresentation by colleges. Misleading financial aid offers and placement rates, often based on incomplete or inaccurate data, further complicate students' understanding of their financial obligations and career outcomes. A lack of authenticity ultimately questions the security and officialdom of such colleges from the perspective of students, education specialists, and families (Marcus).

Critiquing Research

When analyzing misinformation reports from technical schools, there are insufficient sources that specifically address this area of research. Although there are always misunderstandings amongst students and colleges, few of them are rectified or publicized. The publicity of these reports would be significant for students and future researchers who want to inform people worldwide.

The sources used for this section were The Hechinger Report and an archived article from the Department of Education. The ED report presents why they forgave $55.6 million and gives a brief overview of the scams conducted by the three colleges. The Hechinger Report focuses on brightening education and aims to bring out any disparities among students and colleges – it also summarizes some of the more recent misrepresentations that have occurred.

It appears that most of the misrepresentations occurred in the Midwest; hence, further studies should focus on gathering details from Midwest colleges – the reports could have legal consequences, similar to that of Marinello, or they could be minor, such as false links on a website. Data should be collected preferably by conducting a survey on the opinions of students who have been affected by the false portrayals.

To scale it even further, cross-sectional studies should be performed in different geographic areas of the U.S.; this allows a wider audience to be the sample population and one can derive trends on the root cause of college misinformation. Longitudinal studies on a select population of students in the Midwest would also be beneficial in analyzing the frequency of misreports in that area.

Conclusion

Colleges are liable for spreading inaccurate information to their students, evidently confusing them and deterring future students from enrolling. Moreover, higher level institutions are generally prestigious and should have a level of excellence and profound trust when giving students crucial information such as affordability and expected coursework – that trust diminishes as colleges continue to inadvertently spread misinformation.

This section covered three colleges. Students who attended the Marinello Schools of Beauty misunderstood the quality of instruction that they would receive and the overall rigor of

coursework. Westwood College misinformed students that credits were not transferable between institutions, and the Court Reporting Institute did not give sufficient time for students to finish the course as it was misrepresented to students. Combined, the ED forgave students attending any of these colleges, resulting in a forgiveness of $55.6 million.

In general, there is insufficient publicized data available regarding the spread of misinformation amongst colleges. Throughout the U.S., cross-sectional studies should be conducted regularly on any misrepresentations that occur in various geographic regions. Within the Midwest, data should be surveyed by collecting the opinions of students and their firsthand experience with receiving misinformation from institutions. Longitudinal studies could also be incorporated into these methodologies – by continually surveying a selection of students, one can deduce the frequency of how much misinformation is spread over a period of time and how it varies year by year.

Conclusion

This research paper examined the deteriorating effects of college debt on one's health in college and their likelihood of being successful in their field of study, the distinction between techniques used by for-profit and nonprofit institutions in order to generate revenue, the effect of financial literacy on one's cumulative student loan debt levels, and the misrepresentations given by institutions, primarily in the Midwest.

College debt introduces a host of problems in college, ranging from mental depression, anxiety, and stress to unnecessary financial burdens that proves to be continuously concerning later in one's life, impacting family relationships and the prospect of marriage.

For-profit colleges aggressively recruit students in hopes that they will participate in one of their programs, thereby earning money from every admission that they receive. Here, dividends are split between investors and stakeholders of the college. Private nonprofit colleges raise tuition fees as opposed to focusing on the students themselves, but they provide financial support or primarily recruit upper class students who can afford the sticker price. This allows them to attenuate the prospective debt levels, whereas for-profit institutions do not engage in such tactics. As a result, a higher percentage of students at a for-profit institution have student debt compared to a private nonprofit institution. However, the average cumulative debt levels among students between for-profit colleges and private nonprofit colleges vary each year, but there has been a recent surge in 2019-20 in favor of for-profit colleges.

Student financial literacy rates have a negative correlation with debt levels and are crucial for one's long-term financial health. 28% of students who were projected to earn $80,000+ annually were willing to accumulate $50,000+ in debt to complete their degree, whereas less than 10% of students who were projected to earn $30000 annually were willing to accumulate $50,000+ in debt to complete their degree. This emphasizes that financial literacy and one's wealth are directly and positively correlated. Although finance is inadequately taught in common core high school classes, it can be embedded in math curricula, or it can be a mandatory elective

that students must take in 11th or 12th grade. Ensuring that students receive appropriate financial education will significantly impact their college experience and overall well-being.

Colleges have been spreading misinformation in the last two decades, most of which is unreported. These include misrepresentations of course completion, the transferability of credits, and the quality of learning that is provided at institutional facilities. These misunderstandings between students and colleges challenge the validity of higher education while confusing students in college altogether.

Research can be extended in numerous ways. Any study that is conducted must analyze continuous correlational data and should avoid merely stating a change in one aspect over a period of time. That is, a thorough analysis should be conducted using modern algorithms or artificial intelligence – this will deduce trends over thousands of data points, allowing one to make an accurate expectation for the future. Regarding financial literacy, additional studies should be conducted that ask for the individual's budgeting plan, including their income levels during college. Then, one can create a model that represents a correlation between a student's financial proficiency and their income levels. Finally, any events regarding the spread of false information by colleges should be immediately published in the news or made available elsewhere to bring attention to future students, thereby preventing them from attending such institutions and encouraging legal action to be taken by local authorities.

Works Cited

Flaherty, Colleen. "Stress Is Hurting College Students." Insider Higher Education, Inside Higher Education, 23 May 2023, www.insidehighered.com/news/student-success/health-wellness/2023/05/23/survey-stress-hurting-college-students. Accessed 22 July 2024.

FLEC. "Best Practices for Financial Literacy and Education at Institutions of Higher Education." U.S. Financial Literacy and Education Commission, 2019, home.treasury.gov/system/files/136/Best-Practices-for-Financial-Literacy-and-Education-at-Institutions-of-Higher-Education2019.pdf/. Accessed 22 July 2024.

Hanson, Melanie. "Student Loan Debt Statistics." Education Data Initiative,Education Data, 15 July 2026, educationdata.org/student-loan-debt-statistics. Accessed 22 July 2026.

Hanson, Melanie. "Average Student Loan Payment." Education Data Initiative, Education Data, 12 July 2026, educationdata.org/ average-student-loan-payment. Accessed 22 July 2026.

IGrad. "How Financial Literacy Boosts Retention and Graduation Rates." IGrad for Schools, 2024 iGrad, https://enrich.org/insights/1901/how-financial-literacy-boosts-retention-and-graduation-rates. Accessed 22 July 2024.

Hanson, Melanie. "Student Financial Literacy Goes beyond Student Loan Debt." IGrad for Schools, 2026 iGrad, https://enrich.org/insights/2032/student-financial-literacy-goes-beyond-student-loan-debt. Accessed 22 July 2024.

Marcus, Jon. "Colleges Provide Misleading Information about Their Costs." The Hechinger Report, Hechinger Report, 16 Sept. 2019, hechingerreport.org/eds-provide-info-on-colleges-they-say-will-help-consumers-avoid-bad-actors/. Accessed 22 July 2024.

"National Student Financial Wellness Study Key Findings Report." Center for the Study of Student Life Ohio State University, Office of Student Life College of Education and Human Ecology, Sept. 2015, cssl.osu.edu/posts/documents/nsfws-key-findings-report.pdf. Accessed 22 July 2024.

Ott, Molly. "Why For-profit College Enrollment Has Increased during COVID-19." Salon, Associated Press, 22 Nov. 2020, www.salon.com/2020/11/22/why-for-profit-college-enrollment-has-increased-during-covid-19_partner/. Accessed 22 July 2024.

"Percentage of Graduate Degree Completers with Student Loan Debt and Average Cumulative Amount Owed, by Level of Education Funded and Graduate Degree Type, Institution Control, and Degree Program: Selected Academic Years, 1999-2000 through 2019-20." National Center for Education Statistics, 2022, nces.ed.gov/programs/digest/d23/tables/dt23_332.45.asp. Accessed 22 July 2024

The Impact of Student Loan Debt." National Association of Realtors, Morning Consult, Sept. 2021, cdn.nar.realtor//sites/default/files/documents/

2021-the-impact-of-student-loan-debt-full-report-09-14-2021.pdf. Accessed 22 July 2024.

The Practice. "Debt Takes a Toll." Harvard Law School Center on the Legal Profession, Sept. 2023, clp.law.harvard.edu/knowledge-hub/magazine/issues/student-debt/debt-takes-a-toll/. Accessed 22 July 2024.

Press Office. "Department of Education Approves Borrower Defense Claims Related to Three Additional Institutions." U.S. Department of Education, 9 July 2021, www.ed.gov/news/press-releases/department-education-approves-borrower-defense-claims-related-three-additional-institutions. Accessed 22 July 2024.

Quintana, Chris. "Are For-profit Colleges Worth the Cost? Graduates Are Split on the Value of Their Degrees." USA Today, Gannett Seattle Information Network, 31 Jan. 2023 www.usatoday.com/story/news/education/2023/01/31/for-profit-college-education-student-loan-debt/11136095002/. Accessed 22 July 2024.

Schak, Oliver J., et al. "Student Debt and the Class of 2020." The Institute for College Access and Success, Creative Commons, Nov. 2021, ticas.org/wp-content/uploads/2021/11/classof2020.pdf. Accessed 22 July 2024.

The Impact of the African Growth and Opportunity Act (AGOA) on Sub-Saharan African Economies By Binh Minh Tran

Abstract

The AGOA, a nonreciprocal U.S. trade preference program enacted in 2000, had provided duty-free access to the U.S. for eligible Sub-Saharan Africa countries. While the policy was expected to be of significant aid to the growth of the eligible countries, many past studies have shown results in opposition to this expectation. In addition, there is limited research on the difference in growth of valid and invalid SSA countries after the enactment of SSA countries as well as analysing all markets and exporting items as a whole. Therefore, as the AGOA is coming to an end after several extensions, the paper aims to explore the growth rate of the SSA countries – both eligible and ineligible – and the influences of the act on the included countries. This paper employed a more holistic set of data compared to past papers from 1990 to 2023, estimating two-way fixed effects and event-study models of AGOA eligibility on GDP, GDP growth and trade openness. The analysis showed that there are no statistically significant effects of AGOA eligibility on GDP, GDP growth or trade openness. The event study showed that there is no significant pre-trend in GDP growth, but also no notable post-treatment effect. In contrast, the trade openness records a significant downward trend, invalidating the parallel trends assumption. To conclude, it seems that preferential market access is not enough to shift the aggregate GDP or growth for AGOA beneficiaries, and domestic structural factors are more important than trade-preference effects in influencing the economy of SSA countries at the aggregate level.

Introduction

The African Growth and Opportunity Act (AGOA) is a U.S. unilateral trade preference program - a one-sided program given to developing countries to foster their economies - enacted in 2000. The main target members of this act are 49 Sub-Saharan African (SSA) countries, of which 32 out of 49 countries are eligible for AGOA as of 2024. The AGOA is built around the U.S. Generalized System of Preferences (GSP) program, requiring countries to be eligible for GSP to qualify for AGOA. Additionally, there are five primary requirements for countries looking to apply for AGOA: economic (status of the AGOA country's market economy, economic reform, and elimination of barriers to U.S. trade); political (rule of law, political pluralism, and anti-corruption); poverty reduction; labor, child labor, and human rights; and terrorism and security (USITC, 2023).

Even before its enactment in 2000, the African Growth and Opportunity Act (AGOA) had been the topic of debate over its usefulness. Many people had argued that the policies might lead to detrimental effects for the U.S. domestic industries and that the AGOA is ineffective in improving the economies for SSA countries. On the other hand, others state that this is an important policy to open the way for SSA countries' exports as well as a method to mitigate the influence China has on Africa. This initiated many studies to investigate the effects that AGOA has on eligible countries.

One way to analyse the impact of the AGOA is through examining the export-led growth theory. This theory suggests that by fostering international trade, especially through export-oriented policies, countries can promote economic growth due to greater capital utilization, technological improvement through higher competition, creating jobs through foreign demand, and allowing countries to allocate resources efficiently based on their comparative advantage (Balassa, 1978). Subsequently, a country's level of trade openness is seen as an essential component in export-led growth theory. To make countries more open to trade, trade liberalization is one of the main tools used; trade liberalization works by cutting tariffs, eliminating trade barriers, increasing quotas, and deregulation, allowing freer exchange of goods and services. Therefore, trade liberalization is an important tool—specifically for the export-led growth theory—for the development of countries. This leads us to the African Growth and Opportunity Act (AGOA). The AGOA is a nonreciprocal U.S. trade preference program that allows duty-free access to the U.S. market for most exports by eligible Sub-Saharan African countries, integrating the Sub-Saharan African economies into the world trade system (Wong, 2023).

Literature Review

While the AGOA had the objective of improving trade and opening up opportunities between SSA countries and the U.S., there have been debates over its usefulness. Thus, many studies were conducted to see its actual effects on the exports of SSA countries.

Between the period 1995 to 2005, it seems as if the AGOA provided insignificant effects on trade to its members, which means AGOA did not achieve the effects it set out to have of increasing trade between the U.S. and Sub-Saharan African countries (Zappile, 2011). The main aims of AGOA—diversifying SSA economies and increasing their access to the U.S. market—have not been fully achieved; AGOA allows SSA countries to export over 1,800 duty-free product lines beyond standard GSP to the U.S. with no duties or customs, yet crude oil historically accounted for around 90% of total export value under the policy (Jacob et al., 2020). It is also mentioned that while AGOA has the effect of initializing trade between SSA countries and the U.S. in several product categories and sectors, its effect of increasing the overall volume of trade has been minimal, if not insignificant (Tadesse, 2008).

However, by employing a triple-difference-in-differences model, Frazer & Van Biesebroeck found out that SSA countries' GDP rose by 0.12% as a result of AGOA, a small but statistically significant number (Frazer & Van Biesebroeck, 2010). However, they also suggest that domestic factors of SSA countries are limiting the effectiveness of AGOA, stating that barrier removal can lead to significant impact (Frazer & Van Biesebroeck, 2010). For instance, poor infrastructure, low levels of per capita income, small geographical size, and domestic policies.

Furthermore, the SSA apparel sector gains significant benefits from the AGOA program, allowing countries to expand their manufacturing capacity and helping some in poverty reduction and job growth (USITC, 2023). In addition to that, the African Growth and Opportunity Act has

proven to raise trade flows between member countries and the U.S., and FDI inflows have been significant; however, the positive effects vary between countries and sectors, and challenges it set out to solve still persist, for instance, trade diversification and poverty reduction (Tadesse, 2024).

It should be mentioned that many past papers are failing to capture the full duration of AGOA and yielding differences due to the lack of data back in the days and the lack of cross-sectional coverage (Tadesse, 2008). To add to that, many studies focus the impact of the AGOA specifically on a single industry such as textiles, apparel, and agriculture (Zappile, 2011; Tadesse, 2024). Consequently, such studies do not give a comprehensive view of the overall impact of AGOA on total exports and aggregate economic growth, but rather evaluate effectiveness based on changes in those specific sectors. Past studies also rarely compare eligible and non-eligible SSA countries to see the actual baseline difference in benefits given by AGOA.

One more important thing to note is the fact that a 10-year extension for AGOA from 2015 to 2025 was not included in most early papers, meaning the current long-term effects AGOA may have on SSA countries are possibly different from those found in past studies. Therefore, such gaps in past studies inspired the creation of this paper.

Data

From the World Bank's World Development Indicators (WDI) database, several macroeconomic variables have been extracted. The list includes Gross Domestic Product (GDP) in current U.S. dollars, Population, GDP deflator, Official Exchange rate, as well as the exports and imports of goods and services as a percentage of GDP.

It is noteworthy that the variable Trade Openness is not directly acquired from the database, but rather calculated by adding the exports and imports of goods and services (as a % of GDP) together. In addition to that, data regarding the exports of SSA countries—both the treatment group and non-treatment group—into the U.S. were gathered from the UN Comtrade database.

As aforementioned, the effects of AGOA are not universal for all SSA countries. Even though the policy was enacted in 2000, many Sub-Saharan African countries became eligible at different time periods. For instance, while Benin and Botswana received eligibility in 2000, other countries like Burundi gained entry in 2006 or Côte d'Ivoire in 2011. Additionally, eligibility is reviewed annually by the U.S. government, meaning it can be revoked or reinstated. As a consequence, the treatment variable $AGOA_Eligible_{it}$ is not static, but rather a dynamic binary variable that changes between 0 and 1 based on eligibility in a specific year for country i. The information regarding SSA countries' eligibility was gathered from the Office of the United States Trade Representative (USTR) and the International Trade Administration (ITA).

The data span from 1990 up until 2023, showing a holistic view of the overall impacts of the African Growth and Opportunity Act and showing how SSA countries' growth differed pre- and post-enactment. This forms a country-year panel dataset to compare changes across years. However, as data for SSA countries in early years are limited, several countries have missing

years of data, leading to an unbalanced panel. Furthermore, as several countries have missing data, especially for key dependent variables like GDP or key control variables, those specific country-year observations are removed from the regression sample.

Defining Variables

The dependent variables, or outcome measures, of this study include Gross Domestic Product (GDP) for SSA countries measured in current U.S. dollars, annual GDP growth rate calculated as year-over-year percentage change, and Trade Openness—the sum of a country's total exports and imports of goods and services as a percentage of GDP—calculated by Equation (1):

$$Trade\ Openness_{\{it\}} = \left(\frac{\left(Exports_{\{it\}} + Imports_{\{it\}}\right)}{GDP_{\{it\}}}\right) \times 100 \quad (1)$$

The independent variables are the baseline treatment dummy *AGOA_Eligible_it* and the Event_Time variable. The *AGOA_Eligible_it* is a time-varying binary variable representing the eligibility of SSA country *i* in year *t*, with 1 representing beneficiary status and 0 stating the opposite. For countries in the control group, this variable remains 0 throughout the study. For countries in the treatment group, however, this variable changes to 1 in the year of admission and subsequent eligible years, reflecting different admission times. The *Event_Time* variable, by standardizing calendar years relative to the year of AGOA implementation for treated countries ($k = t - t_$entry), allows us to map the dynamic trajectory of policy effects up to 10 years pre- and post-policy.

Methods

In order to estimate the treatment effect of AGOA eligibility on the macroeconomic performance of SSA countries, a generalized Difference-in-Differences framework is employed using the Two-Way Fixed Effects (TWFE) specification, as shown in Equation (2):

$$Y_{\{it\}} = \beta_0 + \beta_1 \times AGOA_Eligible_{\{it\}} + y_i + d_t + \epsilon_{\{it\}} \#(2)$$

In which *Y_it* represents the dependent variable—GDP, GDP growth rate, or Trade Openness—for country *i* in year *t*. *AGOA_Eligible_it* is the binary variable representing eligibility for AGOA. *B_0* shows the baseline average outcome for the reference category. The coefficient of interest, *B_1*, shows the average impact of being an eligible member of AGOA. Next is *y_i*, the country fixed effects, controlling for time-invariant, unobservable characteristics unique to each country, such as geographic landlocked status. *d_t* represents year fixed effects, controlling for aggregate macroeconomic shocks that affect the entire SSA region simultaneously. Finally, standard errors are clustered at the country level to account for potential serial correlation within countries over time.

As aforementioned, the baseline DiD model assumes parallel trends for treated and control countries before AGOA enactment, meaning their economies were moving along similar trajectories. To verify this and observe whether the impact of AGOA on eligible SSA countries grows, shrinks, or stays constant over time, an event study specification is expressed in Equation (3):

$$Y_{\{it\}} = B_0 + \sum_{\{k=-10,\, k\neq -1\}}^{\{10\}} \times B_k \times D^k_{\{it\}} + y_i + d_t + e_{\{it\}} \#(3)$$

Here, *k* is the *Event_Time* variable, signifying years relative to the start of AGOA eligibility, with *k = 0* as the year a country became eligible. Because countries have different entry years, k follows event time rather than calendar time. *D_it^k* is a dummy variable equal to 1 if country i is *k* years away from its initial year of AGOA eligibility. The model window covers 10 years before and after beneficiary status (*k* in [-10, 10]) to maintain empirical consistency. The period *k = -1* (one year before entry) is excluded to serve as the baseline reference point (normalized to zero). Therefore, each estimated *B_k* identifies the average difference in the outcome variable compared to the baseline period, allowing us to visually inspect data for pre-treatment trends and observe differences in policy impact over the post-treatment horizon.

Results

1. Baseline TWFE Estimates

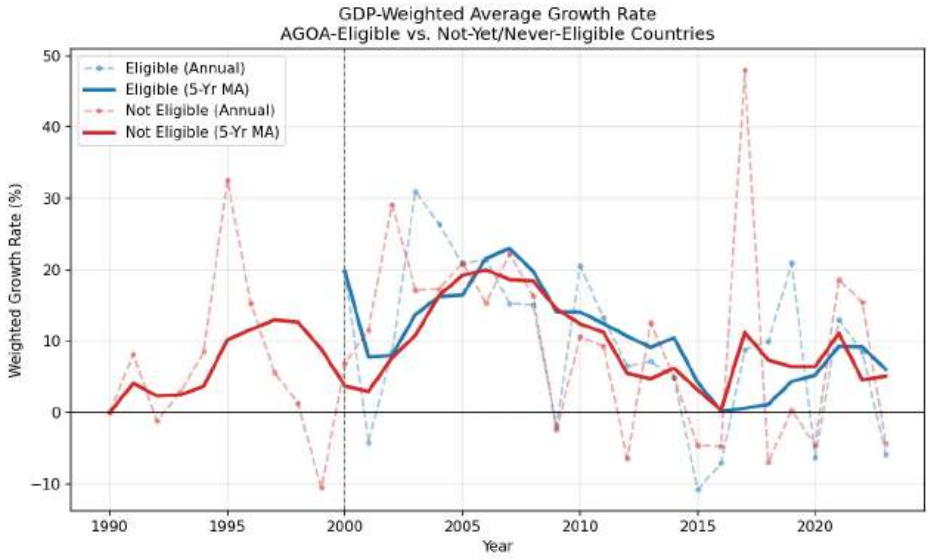
The three outcome variables are estimated independently using Equation (2), with standard errors clustered at the country level.

Table 1. Baseline TWFE Estimates (Equation 2)

Outcome	β₁ (AGOA_Eligible_it)	Std. Error	p-value
GDP (current US$)	+3.85 billion	6.01 billion	0.522
GDP Growth Rate (pp)	−1.76	2.72	0.518
Trade Openness (pp)	−3.58	9.06	0.693

As all three of them have p-values that surpass 0.05, the conventional significance threshold, none of the variables have a statistically significant relationship with AGOA eligibility. Two of the point estimates – GDP Growth Rate and Trade Openness – are negative, while there are large standard errors relative to the coefficients, this should not be considered as a negative effect but rather only as an estimate that can not rule out zero. While this differs from the 0.12% effect reported by Frazer & Van Biesebroeck (2010); that estimate comes from a different sample and method of triple difference-in-differences, so the two results are not compatible.

2. GDP-Weighted Average Growth Rate

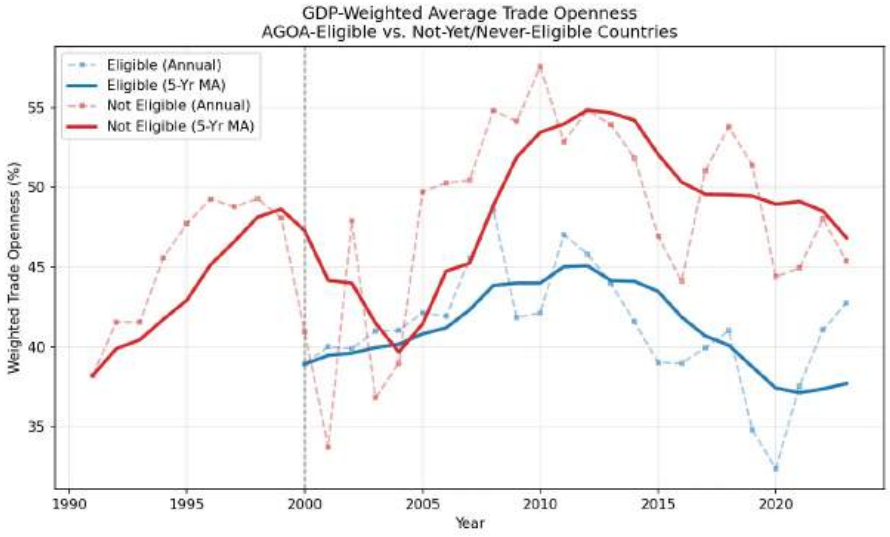


The period 1995 to 1999 witnessed a noticeable discrepancy in GDP growth rates. This inconsistency can be attributed to rising interest in U.S.–Africa trade as well as rumors circulating regarding drafts of AGOA, leading to an anticipation effect. Due to these drafts, reforms might have been undertaken by countries to fulfill the strict eligibility requirements of the policy, resulting in an upward change in growth rate. However, it is also possible that this surge was driven by other factors, such as price increases for key commodity exports supplied by eligible countries or domestic policy reforms aimed at general economic improvement. For this reason, it remains uncertain whether this surge can be considered a part of the policy's treatment effect.

In 2000, the annual growth rate for eligible countries surged, likely due to the official enactment of AGOA providing incentives for economic activity. In 2001, the growth rate for eligible countries dropped temporarily; the exact driver for this drop remains unclear as there were no unique major shocks affecting only the eligible group. Nevertheless, in subsequent years, growth rates for eligible countries recovered, which are comparable to the levels of non-eligible countries. While this may indicate an inconsistent growth gap, it is also possible that AGOA helped maintain growth rates that would have dipped without preferential market access.

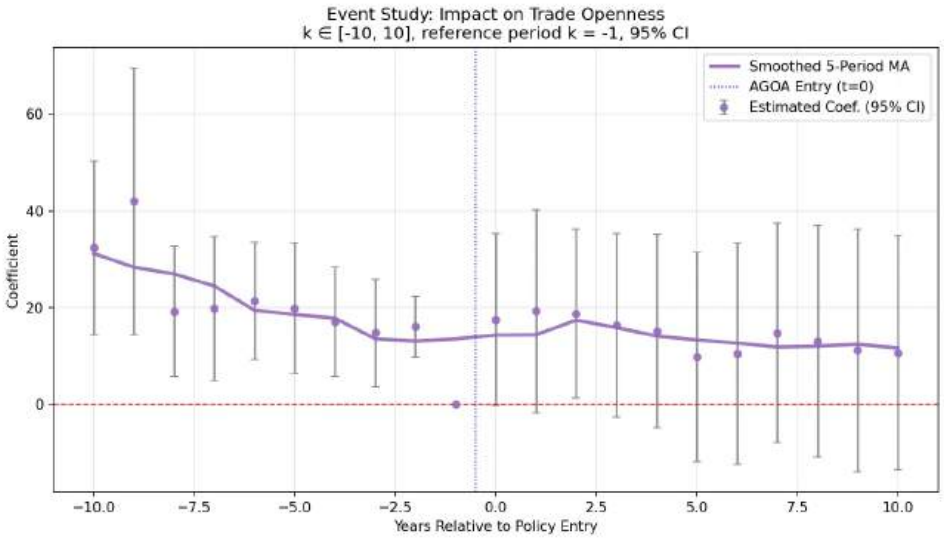
However, from 2008 onward, there are clear signs that the effectiveness of the African Growth and Opportunity Act, in the case that there are noticeable impacts, entered an exhaustion phase, as reflected in declining growth rates. Furthermore, growth rates for both groups began fluctuating in similar patterns. From 2010 to 2023, severe global economic shocks occurred—such as the global commodity price crash in 2015 and the COVID-19 pandemic in 2020. These external shocks overwhelmed any regional advantages provided by AGOA, canceling out differential effects and marking a normalization in growth rates between the two groups.

3. GDP-Weighted Average Trade Openness



Following the policy enactment in 2000, the eligible group recorded an increase in trade openness, reflecting expanded total transactions with foreign trading partners. Notably, while both groups experienced gains, the increase for the non-eligible group was noticeably steeper. In theory, AGOA should have driven greater relative export expansion among eligible SSA nations; however, the steeper increase in non-eligible nations suggests that broader global market factors, raw material demand, or import growth were playing significant roles across the region during this period.

4. Event Study: Impact on Trade Openness

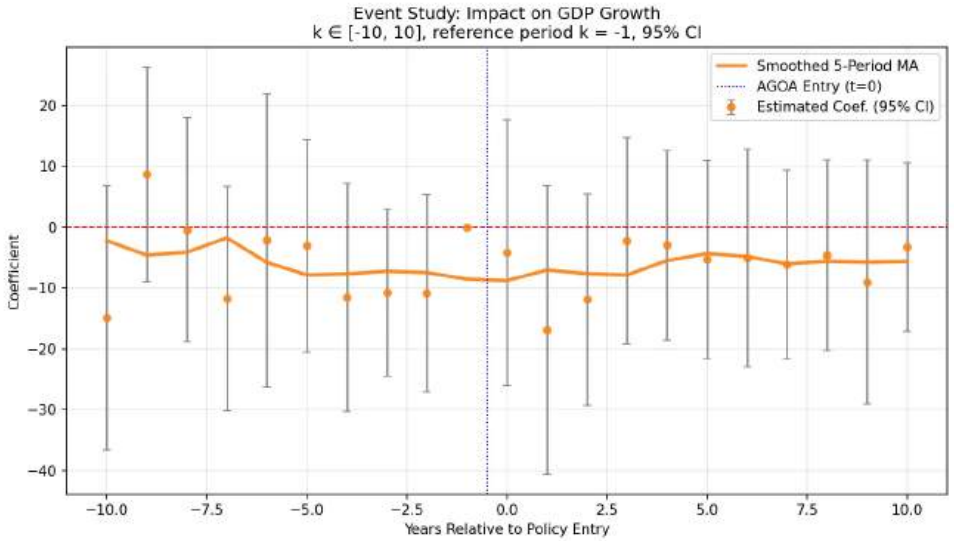


Prior to policy entry ($k < 0$), the estimated coefficients are consistently high, declining from 32 at $k = -10$ down to approximately 14 at $k = -2$. Therefore a joint F-test to see if the

coefficients ($k < -1$) are simultaneously equal to zero is strongly rejected ($F = 6.99$, $p < 0.001$). This means that there is a violation of the parallel-trends assumption which is crucial to the difference-in-differences design: eligible countries were on a different trajectory compared to non-eligible even before the treatment took effect. Hence, the model's core assumption is not plausible for trade openness. Post-policy coefficients continue hovering around levels similar to the pre-policy period, and even show periods of slight decline, continuing the pre-existing trend. As a result, the post-entry estimates cannot be interpreted causally, but rather only descriptively. Still, this graph provides us with two main insights:

- First, the graph suggests that there was no massive structural shock to the trade volume of eligible countries following enactment, pointing toward limited policy impact on total trade openness.
- Second, because the average trade openness coefficient for eligible countries was already high prior to $t = 0$, it indicates that the U.S. selected countries that were already highly trade-oriented into the program, meaning their open economic status was a pre-existing characteristic rather than a direct result created by AGOA.

5. Event Study: Impact on GDP Growth



Pre-policy, the estimated coefficients for impact on GDP growth stayed close to the zero line, remaining slightly below zero for most of the pre-treatment period. Following implementation ($k >= 0$), the coefficients did not record any significant upward shifts into positive territory. Instead, they remain close to the zero line, with the smoothed 5-period moving average flattening out between 0 and -5.

A joint F-test of whether all pre-treatment coefficients are simultaneously zero fails to reject the null ($F = 1.54$, $p = 0.161$), meaning no statistically significant pre-trend was seen for

this outcome. Therefore, it is safe to conclude that the long-term causal effects of the African Growth and Opportunity Act on annual GDP growth rates are statistically insignificant. This leads to the conclusion that preferential trade access alone, while beneficial for specific sectors like textiles and apparel, has not acted as a primary driver of aggregate macroeconomic growth across Sub-Saharan African economies.

Discussion

The results indicate that the AGOA has no statistically significant effects on any of the measured outcomes (GDP, GDP growth, trade openness). This aligns with the conclusions of Zappile (2011) and Tadesse (2008), showing that this is a consistent outcome which improves the validity of the study. However, there are contrasts with the results of Frazer & Van Biesebroeck (2010). This can be explained through a wider time range, including post-2015 extensions which were not included in past studies; additionally, the papers also have different empirical specifications, causing a difference in the final conclusion.

Regarding setbacks of this paper, one notable key limitation is the violation of the event study estimates for trade openness due to statistically significant downward pre-trend before the treatment ($F=6.99$, $p<0.001$). This shows that eligible and non-eligible countries were going through different trajectories prior to the AGOA, rendering the data to be incomparable and invalidating the assumptions of the difference-in-differences model. As a result, the data was analysed in a descriptive way instead of reaching a conclusion through the model causally. Furthermore, many countries have years of missing data, causing an unbalanced panel and even the exclusion of Djibouti as there was little record available for this country. In addition, as the two-way fixed effects method is being used, the data might have been biased in the presence of AGOA's staggered rollout across countries. Another significant setback is that this paper is analysing the whole eligible and non-eligible groups, not individual countries. Consequently, there are extreme values that can influence the groups' averages, calling great caution when it comes to examining the charts.

Through this, we can set a basis on the overall impacts of the African Growth and Opportunity Act to discuss if the renewal of this policy should be supported. In more than 20 years of preferential access, the AGOA has failed to move aggregate growth. This clearly indicates that domestic structural factors hold greater importance in the growth of SSA countries rather than increasing trade access. For AGOA to actually have an effect, significant changes need to be made internally by the government of the eligible countries, such as improving its infrastructures, revising domestic policies to create a more dynamic economy, and reducing regulatories for trading to fully enjoy benefits of the act.

For future researchers, it is recommended for them to rerun the analysis with other types of estimator to avoid biases, such as a staggered-adoption-robust estimator. Furthermore, if possible, the data should be analysed individually for each country to avoid outliers and also to see if different sectors as distinct industries may receive distinct effects and trends. This allows

us to verify whether the AGOA is actually insignificant on every aspect of the economy or only on industries with domestic limitations.

Conclusion

Taken together, these results suggest that AGOA eligibility, on its own, has not produced a statistically detectable improvement in aggregate GDP, GDP growth, or trade openness for Sub-Saharan African beneficiary countries between 1990 and 2023. The baseline fixed-effects estimates for all three outcomes were relatively not significantly different from zero, and the event-study failed to find any leap in GDP growth following a country's entry into the program. Trade openness, on the other hand, is a more complicated story: eligible countries were already diverging from non-eligible countries before treatment began, meaning any patterns or effects post AGOA cannot be attributed to AGOA with confidence due to parallel trend violation. These findings, however, do not imply that AGOA has been worthless — past sector-specific research points to real gains in apparel manufacturing and export; however, on the macroeconomic level, such effects can be considered insignificant. This suggests that preferential market access alone has not been sufficient to shift the broader economic trajectory of most beneficiary countries, even though some industries might have been affected. In order to fully take advantage of the African Growth and Opportunity Act, internal growth and policies changes are required for the eligible countries. Therefore, when considering future extensions, the U.S. government must re-evaluate the economic structures of eligible countries and consider if domestic factors, such as regulations, are supporting the AGOA or not. In the case where the AGOA is extended but some eligible countries' economies are not ideal for the AGOA to take place, efforts should be made to encourage changes in such countries.

Works Cited

Balassa, B. (1978). Exports and economic growth: Further evidence. Journal of Development Economics, 5(2), 181-189.

Frazer, G., & Van Biesebroeck, J. (2010). Trade growth under the African Growth and Opportunity Act. The Review of Economics and Statistics, 92(1), 128-144.

Investopedia. (2022). Export-Led Growth Strategy Through History. https://www.investopedia.com/articles/investing/011416/exportled-growth-strategies-through-history.asp

Jacob, O., Tadesse, B., & Fayissa, B. (2020). African Growth and Opportunity Act (AGOA) and Export Diversification: An Analysis. Working Paper Series.

Tadesse, B. (2008). Nonreciprocal trade agreements and trade: Does the African Growth and Opportunity Act (AGOA) increase trade? Journal of International Trade & Economic Development, 17(3), 411-440.

Tadesse, B. (2024). The Impacts of the African Growth Opportunity Act on the Economic Performances of Sub-Saharan African Countries: A Comprehensive Review. Research Progress Report.

United States International Trade Commission [USITC]. (2023). African Growth and Opportunity Act: Program Performance and Economic Impact (Publication No. 5419). USITC Investigation No. 332-588.

Wong, A. (2023). U.S.-Africa Trade Relations and AGOA Implementation. Congressional Research Service / U.S. Trade Representative Report.

World Bank. (1981). Effects of Export-Oriented Policies: Further Evidence. World Bank Staff Working Paper No. 680.

Zappile, T. N. (2011). The impact of AGOA on Sub-Saharan African exports to the US. Journal of Agricultural & Food Industrial Organization, 9(1), 1-28.

Beyond the Wage Gap: Explaining the Persistence of the Gender Wealth Gap in the United States By Hudson Waiting

Introduction

In the long-standing fight for equality between men and women, the gender wage gap has remained a persistent disparity, yet it may not be the most significant economic divide separating the sexes. The Organization for Economic Co-operation and Development (OECD) defines the gender pay gap as the median difference in earnings between working men and women. The U.S. Census Bureau first quantified this gap in the 1960s, when women earned only about 61% of men's wages (Kollar and Scherer 42). Since then, the gap has narrowed considerably, with a recent report finding that women's earnings have risen to approximately 83% of men's (Kollar and Scherer 10). Despite this progress in earnings, women remain disadvantaged by another persistent inequality: the gender wealth gap. The gender wealth gap, defined as the measure of disparity in accumulated assets, investments, and net worth between men and women, has been growing even as wages have become more equal (Ruel and Houser). This raises an important question: Why does the gender wealth gap persist despite a narrowing gender wage gap in the United States? Although the narrowing gender wage gap has improved women's earning potential, it has not closed the gender wealth gap because wealth accumulation depends on more than income. Factors such as caregiving-related career interruptions, investment and retirement savings disparities, occupational segregation, and historical inequalities continue to limit women's ability to build long-term wealth.

Existing Research on Gendered Earnings and Wealth

Scholarship on gender inequality has increasingly expanded its focus from differences in annual earnings to disparities in long-term wealth accumulation. The U.S. Census Bureau's annual income report provided nationally representative data confirming that a measurable earnings gap between men and women persists (Kollar and Scherer 10). The Bureau of Labor Statistics' *Highlights of Women's Earnings* report expanded on this analysis by comparing earnings differences across occupations, industries, education levels, and racial groups, showing that the size of the gap varies depending on sector and credentials. For example, women with less than a high school diploma earn approximately 82% of what similarly educated men earn, whereas women with a bachelor's degree or higher earn only about 78% as much as their male counterparts ("Highlights of Women's Earnings in 2023"). Examining these patterns over time, a Pew Research Center analysis found that occupational segregation and the earnings penalties that come with parenthood and caregiving are among the strongest perpetuators of the wage gap, even as wages approach parity (Kochhar). These sources show that although the overall gender wage gap has narrowed, substantial disparities persist across education levels, occupations, and family circumstances, confirming a broader trend toward widening earnings inequality.

While researchers generally agree that the gender wage gap is gradually shrinking, the gender wealth gap has not followed the same trajectory. Sociologist Angela Lee examined how

differences in earnings, employment patterns, and socioeconomic factors contribute to long-term wealth inequality. While her study of individual-level wealth finds that lifetime earnings differences account for much of the gap, she notes that men have more access to employer benefits, some of which include retirement benefits, insurance, and stock options (Lee). Her findings suggest that earnings alone cannot fully explain the trend, and that there are other more structural issues at play. Using a different methodology, a fifty-year longitudinal study reached a similar conclusion: lower lifetime earnings for women reduce their wealth accumulation, and differences in investment behavior further widen the gap (Ruel and Houser). Further research extends this argument by showing that income inequality alone cannot explain differences in wealth and instead points to structural factors such as caregiving responsibilities, divorce, and unequal access to investment opportunities (Chang). These findings suggest that narrowing the wage gap alone will not eliminate gender disparities, and that addressing the wealth gap will require greater attention to disparities in saving, investing, caregiving, and employment benefits.

One structural factor that emerges repeatedly across these studies is family and household structure. Despite relying on different research methods like cross-sectional analysis, longitudinal data, and qualitative synthesis, researchers have consistently found that caregiving responsibilities and household composition impact long-term wealth accumulation. Sociologists Dalton Conley and Miriam Ryvicker found that lower savings rates account for much of the financial disadvantage experienced by female-headed households relative to male-headed households. Their work challenges the assumption that income alone determines wealth accumulation, reinforcing Chang's argument that even though earnings are relatively similar, caregiving responsibilities and the financial consequences of divorce reduce women's ability to build assets. Similarly, research by Rakesh Kochhar has found that the economic penalties associated with parenthood fall disproportionately on women, reducing both lifetime earnings and opportunities for advancement (Kochhar). These researchers' findings suggest that family-related financial penalties create a compounding disadvantage, reducing what they can retain, invest, and accumulate over time.

Because family and caregiving responsibilities can limit the money available for saving, the next question is how women invest the resources they are able to retain. Yet the gendered nature of investment behavior remains underdeveloped in the existing literature. Ruel and Hauser's fifty-year longitudinal study identifies investment decisions as a possible contributor to the gender wealth gap but does not investigate the mechanisms behind these differences extensively. Moreover, because its sample consists only of graduates from a single Wisconsin high school, its findings may not be broadly generalizable. Vanguard's *How America Saves* report provides large-scale evidence that differences in retirement-plan participation and contribution rates contribute to disparities in long-term investment growth ("How America Saves 2026" 95). However, because the report analyzes only Vanguard-administered retirement accounts, its findings may not represent workers without employer-sponsored retirement plans or those investing through other financial institutions. Even though existing research shows that disparities in retirement-plan participation, contribution rates, and access to investment

opportunities can compound into substantial differences in wealth, it does not fully explain what produces these disparities or why they persist among men and women with comparable incomes.

Methodology

This review synthesizes evidence from peer-reviewed journal articles and reports published by government agencies and nonprofit research organizations to examine why the gender wealth gap persists despite progress toward wage equality. Sources were identified through Google Scholar and government databases using search terms such as "gender wealth gap," "gender wage gap," "wealth accumulation," "retirement savings," and "investment behavior." Sources were selected based on their relevance to the relationship among earnings, saving and investment practices, structural barriers, and women's long-term wealth accumulation.

Sources were included if they focused on the United States and directly addressed at least one factor connecting gender to long-term financial outcomes, including earnings, employer-sponsored benefits, household structure, caregiving, saving, retirement-plan participation, or investment behavior. Foundational studies from 2005 to 2012 were retained because they provide longitudinal or structural explanations of wealth accumulation, while more recent reports from 2022 to 2026 were used to assess whether these patterns remain visible in current earnings and retirement data. Sources were excluded if they focused exclusively on short-term wage differences without considering their implications for wealth, examined countries with substantially different labor or retirement systems, did not report findings by gender, or lacked a clearly identifiable data source or methodology.

Selected sources include government statistical reports, cross-sectional analyses, longitudinal studies, and qualitative or structural analyses. Studies were organized around recurring themes: lifetime earnings, access to employer-sponsored benefits, family and caregiving responsibilities, household structure, saving practices, and investment behavior. Comparing research findings across these areas made it possible to identify areas of agreement while accounting for differences in data collection processes and research design.

How Economic Disadvantages Accumulate

The central pattern across the sources is one of cumulative disadvantage. In other words, relatively modest inequalities at different stages of women's working lives compound into substantial disparities in long-term wealth. Wealth disparities also appear to result from the interaction of multiple economic and social factors, rather than income differences alone. Across nearly every source, caregiving responsibilities emerged as one of the strongest contributors to persistent wealth inequality. Younger mothers are less likely to be employed, and if they are, working reduced hours. On the other hand, younger fathers are more likely to be employed, and work more hours each week, aggravating the wealth gap (Kochhar). Chang finds that women also have less access to the "wealth escalator," translating to reduced lifetime earnings, limited opportunities for promotion, and decreased retirement contributions (Chang). Differences in

savings strategies and investment also contribute to the issue (Ruel and Hauser). Even relatively short interruptions can compound over decades by reducing investment growth through missed employer retirement matches and lower compounding returns.

Retirement savings and investment participation provide a key mechanism through which early financial disadvantages compound over time. Lower earnings often result in smaller retirement contributions, while differences in investment behavior and access to investment opportunities can further limit growth. Conley and Ryvicker, using longitudinal data from the Panel Study of Income Dynamics, found that differences in savings rates between female- and male-headed households accounted for much of the gender gap in net worth, even after accounting for differences in income (Conley and Ryvicker). Similarly, Lee, using person-level data from the Survey of Income and Program Participation, determined that while the gender income gap had narrowed, the gender wealth gap actually widened from the mid-1990s to the mid-2010s, with women's median wealth falling from approximately 90% of men's to 60% (Lee). These effects are intensified by occupational segregation, family structure, divorce, and historical inequalities in asset ownership. Because of these factors, narrowing the wage gap is necessary but insufficient to eliminate wealth inequality. Long-term wealth depends on continuous employment, consistent investing, access to retirement benefits, and opportunities to accumulate appreciating assets over time.

Toward Equal Opportunities for Wealth Accumulation

Although women now earn a substantially higher proportion of men's earnings today than they did several decades ago, evidence indicates that reducing the wage gap alone will not close the gender wealth gap. Understanding this distinction matters because wealth, unlike income, provides long-term financial security, supports retirement, enables intergenerational mobility, and offers a buffer during periods of economic uncertainty. Existing research consistently demonstrates that caregiving-related career interruptions, disparities in retirement savings and investment behavior, household structure, occupational segregation, and the compounding effects of historical inequality continue to restrict women's ability to accumulate wealth over their lifetimes, patterns often rooted in broader societal expectations about gender roles.

Encouragingly, recent policy developments have begun to address one of the structural conditions that contribute to these disparities. Thirteen states and Washington, D.C. now require employers to offer paid family leave to both parents rather than the mother alone, reducing the disproportionate share of caregiving-related financial burden historically held by women ("Highlights of Women's Earnings in 2023"). Even so, significant research gaps remain. In particular, more work is needed on how investment behavior affects wealth accumulation among individuals with comparable incomes, as well as on the roles of financial literacy, risk tolerance, employer-sponsored benefits, and unequal access to investment opportunities. A clearer understanding of these factors would help policymakers and financial institutions design strategies that address not only income inequality but also the deeper structural drivers of persistent wealth inequality.

Works Cited

Chang, Mariko Lin. "The Women's Wealth Gap." Shortchanged, 2010, pp. 1-17, https://doi.org/10.1093/acprof:oso/9780195367690.003.0001.

Conley, Dalton, and Miriam Ryvicker. "The Price of Female Headship: Gender, Inheritance and Wealth Accumulation in the United States." Journal of Income Distribution®, 1 Jan. 2005, https://doi.org/10.25071/1874-6322.1307.

Highlights of Women's Earnings in 2023. Report no. 1111, U.S. Department of Labor, 2024. U.S. Bureau of Labor Statistics, www.bls.gov/opub/reports/womens-earnings/2023/home.htm.

"How America Saves 2026." Vanguard, 25th ed., workplace.vanguard.com/content/iig-transformation/pdf/how-america-saves-2026.html.

Kochhar, Rakesh. "The Enduring Grip of the Gender Pay Gap." Pew Research Center, 1 Mar. 2023, www.pewresearch.org/social-trends/2023/03/01/the-enduring-grip-of-the-gender-pay-gap

Kollar, Melissa, and Zach Scherer. Income in the United States: 2024. Report no. P60-286, Government Publishing Office, Sept. 2025. U.S. Census Bureau, www2.census.gov/library/publications/2024/demo/p60-282.pdf.

Lee, Angela Wang. "The Gender Wealth Gap in the United States: Trends and Explanations." Social Science Research, vol. 107, Sept. 2022, p. 102745, https://doi.org/10.1016/j.ssresearch.2022.102745.

Organisation for Economic Co-operation and Development. "Gender Wage Gap." www.oecd.org/en/data/indicators/gender-wage-gap.html.

Ruel, Erin, and Robert M. Hauser. "Explaining the Gender Wealth Gap." Demography, vol. 50, no. 4, 21 Dec. 2012, pp. 1155-76, https://doi.org/10.1007/s13524-012-0182-0.

U.S. Department of Labor. "Paid Leave." www.dol.gov/agencies/wb/featured-paid-leave.

AI and the Changing Role of Human Workers By Raunit Sharma

Abstract

Artificial intelligence is changing not only how quickly work gets done but also which parts of jobs humans perform. This paper examines that change using published research and four semi-structured interviews with professionals in product management, software engineering, telecommunications, and network technology. In the interviews, AI was most useful for repetitive work, information processing, coding support, and initial analysis, while humans remained largely responsible for verification, relationships, and final judgment. The evidence shows that AI can make a single employee far more productive while also reducing the number of human workers needed for specific tasks. Entry-level positions may also face pressure because many traditional junior-worker responsibilities overlap with tasks AI excels at. A further concern is skill loss if workers outsource too much reasoning to AI. Overall, this paper argues that AI is reshaping jobs by redistributing tasks rather than eliminating them, increasing productivity while placing pressure on entry-level roles and requiring human judgment, critical thinking, and AI expertise to be of most importance.

Introduction

Artificial intelligence has changed swiftly from experimental software to an everyday tool. Workers are using it to write large amounts of software, summarize documents, analyze data, create presentations, and answer customer questions. Survey evidence suggests these tools can create large amounts of productivity gains. Tracy Brower reports that 77 percent of responses in a 2025 workplace survey reported that AI helped them accomplish more work in less time, while 73 percent said they produced better-quality work with less effort (Brower). Yet Brower also describes a more complex pattern: workplace AI use was nearly evenly divided between augmentation, in which AI assists workers, and automation, in which AI performs work the employee would have done (Brower). Therefore, AI is not just helping people perform jobs faster; it is beginning to redistribute work within those jobs.

That redistribution can produce very different outcomes depending on the occupations. Pew Research Center found that 19 percent of U.S. workers were employed in occupations considered highly exposed to AI, with exposure particularly high among workers with a bachelor's degree or higher (Kochhar). Specifically in radiology, AI is used to prioritize scans, improve overall image quality, and assist with reports while physicians continue to make diagnoses and other decisions (Eadicicco). In the customer service industry, Reuters reports that some companies use AI agents to handle large amounts of routine interactions and reduce the number of workers required for those inquiries (Vengattil and Kalra). Both examples can be true at once: AI can strengthen a worker’s capabilities in one context and reduce labor demand in another.

This paper examines that tension through published research and four interviews with professionals whose work is already affected by AI. Rather than asking only whether AI will

"replace jobs," a more useful question is what happens inside a job when AI can perform more of its tasks. The interviews point to a common transformation: AI is increasingly performing repetitive execution, Pattern recognition, gathering information, and analysis. Human workers are pushed toward judgment, domain expertise, verification, and accountability. This transition may reduce some roles and can weaken worker expertise if employees are overly dependent on the technology. Artificial intelligence is therefore reshaping human-centered work less by uniformly eliminating entire professions than by redistributing tasks within them.

Methodology

During the summer of 2026, I conducted four Zoom interviews with professionals in technology-intensive roles. Gurpratap Thiara is a product manager at Amazon. Krishna Iyer is a Chief Technology Officer for the communications business at GlobalLogic. Rahul Tyagi is a principal software engineering manager at Microsoft working specifically with Azure core compute. Sumeet Prakash is Sr. Director, software development at T-Mobile. Each interview addressed the tasks for which AI is most useful, the value and limitations of AI at work, whether its functions are more useful as an assistant or a replacement, the future of entry-level work, and the skills future employees should develop. The interviews were transcribed and compared for recurring themes. They are used here as qualitative evidence, not as a representative sample of the labor market.

Automation and the Shift from Executing to Judging

The clearest point of agreement across the interviews is that AI is especially effective at repetitive, predictable, and information-dependent portions of jobs. Gurpratap described AI as useful across product development, including research, brainstorming, and creating prototypes. He gave the example of monthly business updates that once took up employee time can now be partially automated. The benefit is not only that the update is produced faster; the saved time can now be redirected towards creative or strategic work (Thiara). This is a useful model for how AI often enters a profession and does not replace the role; instead, it learns repeatedly followed steps as a pattern.

Iyer described the same pattern in telecommunications. His work involves reviewing RFPs that can reach 30-80 pages in length. AI can summarize those documents and identify their central points more quickly than reading every page from the start. However, the summary is useful only when combined with Iyer's knowledge of telecommunications and the customer problems. AI can compress the information, but he still determines what matters and how it should be implemented (Iyer). Tyagi described AI producing baseline code, unit testing, document drafting, log analysis, and initial incident diagnostics. Instead of manually searching through large quantities of technical logs, an engineer can use AI to identify patterns and organize an initial explanation, then spend more time investigating the root cause of the problem (Tyagi). Prakash similarly described AI supporting presentation drafts, testing scripts, generating code, and reviewing that code (Prakash).

The same effect appears outside of the technology profession as well. In radiology, AI can quickly examine large amounts of medical imaging, prioritize scans, and help prepare reports, while doctors remain fully responsible for diagnosis and patient care (Eadicicco). These examples show a broader shift from execution to judgment. When AI handles more searching, formatting, and pattern recognition, workers become more responsible for implementing the results and deciding what happens next.

These changes make human expertise more important, not less. Iyer emphasized that AI can provide information quickly but cannot blindly be trusted to make decisions without context. He identifies security, legal questions, ethical concerns, and customer relationships as important areas where human involvement remains crucial (Iyer). Prakash made a similar point about data analysis: AI can help a team understand customer pain points much faster, but employees still need to verify the numbers and create guardrails for accuracy (Prakash). Thiara also distinguished between tasks such as writing code, where AI may take on a growing share of work, and winning or maintaining customer relationships, where human involvement remains crucial (Thiara). The human contribution is therefore not simply a final approval click; it includes accountability, context, trust, emotional understanding, and the professional knowledge required to recognize when an AI-generated answer is inadequate.

Employment and Entry Level Workers

Human judgment will remain necessary even when fewer human workers are needed. Tyagi explained that work that once required several engineers can increasingly be completed by an engineer who understands the business problem and knows how to use AI effectively and efficiently (Tyagi). Reuters mentions a more dramatic version of the same with Indian call centers. LimeChat says its AI agents allow clients to reduce the number of workers that are needed to handle 10,000 monthly queries by 80 percent, and the company reports automating thousands of jobs (Vengattil and Kalra). Amazon CEO Andy Jassy has also said that AI-driven efficiency will mean fewer people performing some existing jobs while other types of work grow, and he predicted that these gains would reduce Amazon's corporate workforce (Halverson). A tool can therefore augment the employees who remain while simultaneously reducing the number of human employees required.

However, lower labor demand in one area does not mean AI will eliminate jobs overall. Thiara argued that many teams already have more potential work than they can complete. If AI gives the same teams greater capacity, a company can use the productivity gain to build more instead of reducing their headcount (Thiara). He also expects new products, projects, and companies built around AI to create work that does not currently exist. Iyer likewise expects many jobs to be repurposed rather than being lost. An employee who once handled basic customer questions may now instead focus on difficult cases and customer retention (Iyer). AP News similarly reports that economists remain cautious about attributing large layoffs directly to AI because many early gains happen first to workers, and organizational-level changes take time (O'Brien).

The transition may be the most difficult for entry-level workers. Reuters reports concern that young workers entering call-center and back-office jobs may experience the greatest short-term effects of AI (Vengattil and Kalra). Iyer expects basic research, data gathering, basic analysis, and first-level customer service to become increasingly automated, even if many junior positions are transformed rather than removed (Iyer). Tyagi is more cautious about entry-level jobs specifically for software engineers. Junior engineers have traditionally been hired to write smaller pieces of code after senior engineers break a project into manageable pieces for them. Because AI can now produce most of that code, fewer workers may be needed for that same function (Tyagi). This creates a paradox: the tasks easiest to automate are often the tasks through which beginners historically learned a profession. The problem is not only whether an entry-level job survives, but whether the transformed job still gives a worker enough meaningful practice to develop the judgment expected later in their careers.

Skill Loss and Human Oversight

The entry-level problem connects to another risk: skill loss. Tyagi described professional intuition as something that is built through repeated problem-solving. He compared it to a mechanic diagnosing the problems in a broken car. A mechanic who listens to an unusual sound, tests different possibilities, and rules out multiple causes develops knowledge that extends beyond just one repair. If an AI system immediately provides the diagnosis every time, the task may be completed faster, but the mechanic gains less experience that builds future judgment. Tyagi believes the same risk exists in software engineering. A new engineer who continually accepts AI-generated code may produce functioning software without understanding the concepts well enough to recognize when the solution is wrong (Tyagi).

Brower describes a closely related "supervision paradox." Workers need expertise to evaluate whether AI output is accurate, yet delegation of too much work to AI can erode the very skills needed for that supervision (Brower). She also cites MIT research finding substantially lower cognitive engagement among participants who relied heavily on AI tools (Brower). This does not mean AI use automatically weakens thinking. Rather, the main concern is that convenience can reduce opportunities to practice important independent reasoning skills. Human oversight only works when the human is knowledgeable enough to challenge the system. For AI augmentation to remain sustainable, schools and organizations need to preserve opportunities for workers to learn fundamentals, solve problems by themselves, and understand the systems behind AI-generated outputs.

Skills in an AI-Integrated Workforce

When the four interviewees discussed future workers, their answers converged around a similar combination of needed skills. Thiara emphasized curiosity, systems thinking, independent thought, and what he specifically described as AI proficiency. He distinguished real proficiency from simply knowing how to use a chatbox; a stronger worker would understand how to build with AI and use it as part of daily larger systems (Thiara). Iyer emphasized domain knowledge,

communication, relationships, critical thinking, and responsible use. In his view, knowing many AI tools matters significantly less than having the knowledge to intelligently apply them with enough knowledge to identify hallucinated outputs (Iyer). Tyagi similarly expects technical fundamentals to become more important because AI can now produce software even for someone with limited understanding. When a system fails, the worker who understands the concepts behind it will be better able to diagnose the problem. He also emphasized communication, emotional intelligence, and adaptability (Tyagi). Prakash focused on problem-solving: how a worker breaks down tasks, communicates the problem to AI, and evaluates the proposed solution (Prakash). Across the interviews, the pattern is very consistent. As AI reduces the value of routine execution, it increases the relative value of the abilities required to direct that execution smartly. The strongest worker is neither the person who avoids AI nor the person who depends on it for everything, but the person who combines AI fluency with their own expertise.

Limitations and Conclusion

This study has a couple of important limitations. The interview sample is small and concentrated in technology-intensive roles, so it cannot establish how the same patterns will appear across the entire labor market. The participants' comments about future hiring are informed professional predictions rather than statistical forecasts. The technology is also changing quickly, so claims about current model limitations may age faster than the broader questions raised by AI adoption. Future research should include workers outside technology and compare how early-career employees develop expertise under different levels of AI use.

Even with those limits, the interviews and secondary research point toward the same central conclusion. Artificial intelligence is neither simply replacing workers nor making existing work faster; rather, it is changing the division of responsibility between people and technology. AI is increasingly used for repetitive execution, information processing, pattern recognition, and initial analysis. As those tasks move toward AI, human value becomes more concentrated in judgment, context, expertise, accountability, and the ability to recognize when a machine-generated answer is inadequate. Employment consequences will vary: some organizations will use AI to expand output with the same workforce, while others will operate now with fewer people. Entry-level work may be especially unstable because the tasks traditionally assigned to beginners are often the easiest to automate. The long-term challenge is not only preserving jobs, but preserving the process by which people become knowledgeable enough to perform the higher-level responsibilities that remain in human hands. AI is already replacing pieces of human work while expanding what individual workers can accomplish. The future of work will depend on using that power without surrendering the knowledge and judgment that make human oversight especially valuable.

Works Cited

Brower, Tracy. “How AI Is Changing Work and the Human Experience.” Forbes, 22 Dec. 2025, https://www.forbes.com/sites/tracybrower/2025/12/22/how-ai-is-changing-work-and-the-human-experience/.

Eadicicco, Lisa. “This Job Has Become the Ultimate Case Study for Why AI Won’t Replace Human Workers.” CNN Business, 9 Feb. 2026, https://www.cnn.com/2026/02/09/tech/ai-replacing-jobs-concerns-radiology.

Halverson, Alex. “Amazon CEO Expects AI Will Reduce Workforce.” The Seattle Times, 18 June 2025, p. A1. Seattle Times Collection, NewsBank, https://infoweb-newsbank-com.us1.proxy.openathens.net/apps/news/document-view?p=NewsBank&t=pubname%3ASTIW%21Seattle%2BTimes%252C%2BThe%2B%2528WA%2529&sort=YMD_date%3AD&hide_duplicates=2&fld-base-0=alltext&maxresults=60&val-base-0=AI%20replacing%20jobs&docref=news/1A14B2B412385A38.

Iyer, Krishna. Personal interview. Summer 2026.

Kochhar, Rakesh. “Which U.S. Workers Are More Exposed to AI on Their Jobs?” Pew Research Center, 26 July 2023, https://www.pewresearch.org/social-trends/2023/07/26/which-u-s-workers-are-more-exposed-to-ai-on-their-jobs/.

O’Brien, Matt. “Some Companies Tie AI to Layoffs, but the Reality Is More Complicated.” AP News, 2 Feb. 2026, https://apnews.com/article/ai-job-impacts-layoffs-amazon-pinterest-dow-7736d042172743301dd7e494813a885d.

Prakash, Sumeet. Personal interview. Summer 2026.

Thiara, Gurpratap. Personal interview. Summer 2026.

Tyagi, Rahul. Personal interview. Summer 2026.

Vengattil, Munsif, and Aditya Kalra. “Meet the AI Chatbots Replacing India’s Call-Center Workers.” Reuters, 15 Oct. 2025, https://www.reuters.com/world/india/meet-ai-chatbots-replacing-indias-call-center-workers-2025-10-15/.

The Market Potential and Commercial Value of Accessible Design: Business Opportunities in Products and Services for People with Disabilities By Yuzheng Dong

Abstract

Accessible design is widely regarded as necessary accommodations provided to a relatively limited amount and range of users, yet this way of thinking undervalues the scale of the disability market and the overall market impact of accessible products, which are not only simpler but more easily perceived, understood, and used for customers. This paper explores accessibility standards, accessible design research, and case studies from the public health and corporate sector to discuss the value of accessible design for businesses. It argues that making things more accessible would help grow a firm's addressable market, increase product mainstream usability, boost customers' trust in the firm, reduce costs for later remediation, and better prepare the firm for stricter requests from regulatory bodies. Further, the market opportunity of accessible design is not just for people with disabilities, because a change in attention, age, a temporizing injury or environmental factors can also impact consumer interactions with a product and/or service. Such benefits can also be achieved when companies involve disabled people in their research processes, who co-create products, participate as testers, and engage in governance decisions. Examples can be found in business cases like OXO Good Grips, the Xbox Adaptive Controller, and digital accessibility practices. The paper further points out that accessible design is not a money spinner; rather, it has expected payback which depends on the ability to implement it in a non-disaster-oriented way with measurable goals and, of course, long term organizational involvement.

Introduction

Technically speaking, a product may be available to the general public, but practically unusable for its particular segments. An application where you can't use a screen reader, a package which needs strong grip force, a video with no captions, or a ticket machine that communicates only through colors could present substantial barriers between business and customers. Often, these barriers are evaluated as issues integral to business compliance or social responsibility, yet they also point to the failures of the markets. If a customer is unable to find a product, assess its value, buy it, use it or get support for it, then the company has very likely missed a sale, and may have lost the respect and trust of customers.

The scale of such impacted customers is substantial. The World Health Organization (WHO) estimates that about 16 percent of the world's population (or about 1.3 billion people) experience significant disability (World Health Organization, *Global Report* 1). By definition, there is not one single category for disabled consumers. It encompasses individual differences in vision, hearing, mobility, dexterity, speech, cognition, learning, and mental health, along with combinations of these experiences and more. Given that there are differences between people and their particular situations, the population is so big and multifarious that it cannot be considered a mere niche of the fringe of the economy. Accessibility is also a growing issue, as it

has been found that the proportion of people with disabilities rises with age (World Health Organization and World Bank 34). Moreover, persons without permanent impairments typically face temporary, situational constraints: If someone has a broken arm, she or he might not be able to use the phone as they used it before the injury; bright sunlight can affect the visibility of computer screens; background noise can affect the audio; or stress can affect attention and working memory. Creating products with disabled users as stakeholders can thus expand the usability of products which benefit many other users on the continuum of handicaps, either permanent, temporary or situational (Microsoft, *Inclusive 101 Guidebook*).

How may accessible design boost commercial value and broaden market opportunities for businesses? The paper does not employ surveys or interviews, but relies on qualitative secondary research. The analysis draws on disability and design scholarship, international standards, public reports and business cases. It emphasizes that value creation can happen at the intersection of five related mechanisms: market expansion, usability spillover effects, innovation, trust, brand differentiation, and a reduction in life cycle risk. However, such advantages of accessibility only occur when it is built into business strategy and product development. A simple list of 'check' boxes, an annual "special" product look-over or a one-off audit might meet neither the needs of customers with disabilities nor the business goals of firms.

The paper first attempts to clarify the linkages between the concepts of accessibility, universal design and inclusive design. It then assesses the market opportunity in terms of market size and structure, describes the benefits of access with regard to business returns, considers relevant case studies, and concludes by exploring the organizational hurdles that inhibit access implementation. The last three sections offer a practical approach to implementing accessibility as a long-term capability for businesses.

Accessible Design as Market Infrastructure

Accessible design is achieved when a product, service, environment or communication is used by persons with a disability to the same degree, with the same efficiency and in the same manner as persons without a disability. For instance, digital accessibility can entail functional features such as being operable by a keyboard, providing alternatives to images in a text, having a sufficient color contrast, captioning, having clearly marked headings, having predictable navigation and compatibility with assistive technologies. For physical products, it can mean low force controls, labelling that can be read, labelling that is tactile (can be felt), gripping surfaces that do not slip, and packaging that does not require special strength or dexterity to open. As such, accessibility is not a single feature; it permeates and is integral to a customer's full experience of a product.

Overall, three terms are currently employed in scholarship and industry to capture the meaning and significance of accessibility design: universal design, inclusive design, and accessible design. While all three terms convey the same concepts and ideas, each highlights a facet of the issue. The goal of universal design is to create products and/or environments that do not require special adaptation to be used by the broadest possible user base. Inclusive design, on

the other hand, puts particular focus on the difference between users. By highlighting exclusion as a design cue, it aims to draw attention to exclusion, its impact, and its potential resolution. A third concept, accessible design, aims to eliminate barriers to the disabled and attain specific standards. Persson and his co-authors suggest that these ideas have a common aim, but differ in focus in terms of historical and methodological concerns (509-12). As with anything in the business world, it is more important to know how to operate rather than discern terminology differences: human diversity should not be considered only when a “normal” user identity has been established, but should be taken into account in the initial design process.

There is another principle that to some extent alters the understanding of disability. A medical perspective locates disability within a person's body. Yet a social and relational perspective is concerned with a question of how the environment, systems and assumptions are creating the disabling conditions. The limited mobility of wheelchair users is not the only reason they are excluded--when stairs are not replaced by a wheelchair path, wheelchair users are excluded from its access. A blind customer is excluded not only by blindness but also by an unlabeled digital button. Access, Hamraie demonstrates, has long been an issue with universal-design movements, as they have challenged the notion of how "access" is a kind, generous addition to what is an otherwise full, complete environment (18–24). From the commercial perspective, this is not because many customers are lost customers, it is because they are lost customers as a result of design decisions that can be taken up or done away with.

Accessibility is also an infrastructure for the market. The customer should be able to go through the comprehensive stages of discovery, comparison, payment, delivery, usage, and support with reasonably manageable barriers. Incorporating accessibility only in a single stage might not resolve any potential commercial issues. For example, if you have to use visual puzzles with an identity verification on your website that are inaccessible, then the website is not accessible. If the instructions to an easy-open package are not readable, then it is not an easy-open package. If the only way of communicating with the customer service is by voice telephone, there is no access in a captioned advertisement. It is therefore not enough to assess systems in terms of their interfaces; businesses need to assess systems. Particularly, this system's perspective is fundamental to the commercial case and revenue is driven by full journeys and not the individual compliant components.

Estimating the Market Opportunity

It is easy to undervalue the market opportunity for accessible design, if measured with just one single number based on world-wide revenue. The definition of disability is different in various surveys and jurisdictions and so is the commercial potential dependent on the product category, income, infrastructure and sectors. Nevertheless, three different types of evidence indicate that the opportunity of accessible design is important: population scale, unmet demand, and spillover demand (from individuals who would benefit from the same features for other purposes).

The first indicator is the population scale. This figure of 1.3 billion--according to WHO--is bigger than the population of any single country, and it reaches a cross section of consumers in all income and age groups (*Global Report* 1). It fails in one other way, however, in that it also reports on the number of purchase decisions that are affected by accessibility as being less than the total number of purchase decisions. Family members, care takers, schools, employers, health systems and public agencies choose and/or pay for a product for the disabled user. In a business-to-business setting, software that can be readily used can impact a procurement's eligibility since an employer, university or government entity may require software that can be used by the entire business population or service population. The market is therefore not restricted to a single class of disabled people who purchase – covering the single buyer, but also institutions and households that have an access need which influences the purchase decisions.

That leads to the second indicator: unmet demand. The knowledge about the gap is illustrated by digital access. An automated evaluation in 2024 by Web Content Accessibility Guidelines (WCAG) for one million of the most common home pages detected faults in WebAIM's automated evaluations of home pages in 95.9% of those pages (WebAIM). While automated testing can't uncover all the barriers, nor satisfy complete legal requirements, when failure is this high it is clear that inaccessible design is commonplace. Not all the defects mean a lost sale; the implication for business is that there is still fundamental friction in a lot of markets which are able to be taken off by the competition. Accessibility can be clearly explained in a retailer's clear headings, accessible labels for links that make sense, accessible controls that are keyboard friendly, a checkout process that is as easy to complete as possible, and a retailer might garner customers not with a luxury to accessibility, but through a transaction that is easier to complete.

The third indicator is that of spillover demand. Creating features around the people with disabilities can make them mainstream conveniences. This effect that Blackwell calls the “curb-cut effect”—a change that aims to remove one group's barrier for another group reaps broader benefits for the public. Physical curb cuts provide access for people using wheelchairs, but also allow parents with strollers, travelers with luggage, and delivery drivers to access the sidewalk easily (Blackwell). In the commercial design, the same conditions apply. Captions are used for deaf and hard-of-hearing people and for certain noisy situations such as while in transport, quiet environments like in an office, learning languages, or for search. With the use of high contrast text, people with low vision will be able to read it and when used with a screen in the sun, it will improve readability. Voice control can be used to assist a person with limited dexterity when they cook, or to help a cook whose hands are otherwise occupied in busy routines. Simple instructions can help some people who have cognitive disabilities, and minimize errors for almost all users.

This further extends the strategic horizon as a result of population ageing. Functional limitations of vision, hearing, dexterity, balance and memory may result from a gradual change over the course of a lifetime, and thus many consumers may not be handicapped even if they

have one of these functional limitations. Products that are sold on the basis of being designed just for “special needs” might not be selling to customers who even though would like the same effectiveness but are turned off because of the stigma attached to the terms “special needs.” Rather, this can be rectified through the universal and inclusive design approach by integrating access into the mainstream products. It arguably matters less whether or not there is any business opportunity to simply sell more expertise-specific devices, but it does matter to maintain and improve the client relationship as time goes on and as customers' capabilities and contexts evolve.

But the size of the market must not be equated with market demand for any product designated accessible. Consumers with disabilities are not a monolithic group and “accessibility” is not the same as “price”, “quality”, “aesthetic”, “security” or “convenience”. The absence of a good design does not mean that the product is competitive even if it consists of only one accessibility feature. There is a need for businesses to therefore specify its access barriers, user groups, purchasing situations, and value propositions. One market's opportunity can be that same core product which is available to all in the other market. It can also be a user customizable interface, an accessory that adapts to the user, customer support that can be easily accessed and prepared to help other companies with their procurement needs. In fact, the best market analysis firstly starts with some concrete market exclusions instead of with abstractions such as “the disability market is large.”

How Accessible Design Creates Commercial Value

Accessible design can create commercial value through several mechanisms that work together. The simplest one is market expansion. A barrier when removed increases the independence of access to a desired behavior by an individual. In the e-commerce space, it can be a search and compare, payment and delivery tracking. In financial services, it could be account opening, authenticating transactions and so on. In entertainment, it can be in the form of game control and game knowledge. Some methods of measuring the financial value include conversion, retention, transaction completion, and support demand. Commercialization is made possible by changes to these outcomes—things become accessible when they become commercially meaningful.

A second mechanism is something that is more at ease for businesses to use. Information hierarchy, easy navigation, obvious focus (showing which element is selected), bigger targets, easily understood error messages and other options to interactions that take a fixed amount of time can often be essential to accessible design. These practices eliminate any type of friction to the mind and operation. According to the Web Accessibility Initiative (W3C) “Business Case,” accessibility is tied to other benefits for users, such as better user experience, more widely accessible market opportunities, innovation, and less risk. This makes perfect sense: a form which clearly indicates an error can be useful to the screen-reader user, but is useful to a distracted customer as well. A bigger touch target helps people with tremor and minimizes the

number of taps, when a user is on a moving train. Accessibility can consequently enhance metrics other than the one initially driving the change for the group it was put in place for.

A third mechanism is innovation. When designers are working for a wide range of customers which may well experience circumstances at the extremes, naturally they must consider variety and diversity in things as input and feedback, or personalization and control. Indeed, authors including Clarkson believe that inclusive design should take advantage of the diversity as a source of design needs, not just statistical noise (10–14). This can result in a variety of ways of interaction between designers and users, and more flexible products, and even new service models. Observing a disabled user isn't enough to produce an innovative solution; it takes a congruence between the individual, their task(s), and their environment. This end result can then be used in a larger customer base.

Trust and brand differentiation is a fourth mechanism. Accessibility is a way of telling if businesses are working to make themselves accessible to disabled people as customers, employees or members of the public. Better access can enhance user loyalty as they are reluctant to move their business to different products due to the time invested in learning to use them in a compatible way. On the other hand, making changes that eliminate support for assistive technology or that hinder or prevent effective use of assistive technology will no doubt erode trust in a blink. Public communication campaigns about brands should be undertaken with care, as it may seem as if they are exploiting the idea of accessibility design, if brands do not in fact provide access to the products and support. This has been historically common in the United States even after the passing of the Americans with Disabilities Act (ADA). Ronald L. Mace, the proponent of universal design, had elaborated on "seamless design", where design attention to disability should not burden designers but blend naturally and seamlessly in overall plans for all users. But such an idea had been mostly underachieved by companies and media, despite its popular attention (Williamson, 2019). Therefore, a consistent approach to product, marketing, purchasing and service can indeed set a company apart from their competitors, who by contrast still tend to view disability as an exception.

A fifth mechanism is lower lifecycle cost and risk. It can be very difficult and costly to integrate accessibility as an afterthought since defects can reside in design systems, the code, content, vendor agreements and support processes. According to Lazar, Goldstein and Taylor (143–70), in addition to fixing individual technologies, considerations such as organizational policy, procurement, training and evaluation are also required for sustainable digital accessibility. Early integration enables requirements to be incorporated into the research, architecture, component libraries, quality assurance and into acceptance criteria. In no way does this make accessibility "free", but it transforms accessing from remediation to reusability. It also eliminates the risk of complaints, unsuccessful procurement reviews, last minute redesign and the damage to brand reputation.

The environment of regulations supports this business rationale. Accessibility, as a precondition to participation, is a key element of the UN Convention on the Rights of Persons with Disabilities and is addressed by States in their obligations to make information,

communications, facilities and services available to the public (United Nations, art. 9). The Web Content Accessibility Guidelines are guidelines for digital content that have been recognized internationally as technical criteria, and the European Accessibility Act extends the requirements of disability accessibility by linking to a broad range of different categories such as e-commerce, banking services, electronic communications and e-books (European Parliament and Council). The standards do not establish liability and there are various legal requirements for each jurisdiction. Despite all this, there is a clear idea of the direction of policy: accessibility is going further into product markets and supply chains. Companies can build their expertise internally which is preferable to having a complaint or requirement for a contract.

It's also suggested that disability inclusion leads to better organizations, but correlation does not automatically mean causation. During the time period analyzed (Accenture 6–7), the companies that were deemed to be disability-inclusion leaders achieved improvements across a range of financial metrics compared with a sample of other companies in the U.S.. These results may not support that these initiatives alone have generated increased revenue or profit. While some firms are doing well on a number of measures, others could be doing better with their management. High performance is possible alongside disability inclusion, and disability inclusion can be a facet of greater talents, innovations, customer insight and good governance.

This means commercial value must not be taken for granted, but rather measured. Some of the useful metrics are task completion metrics with assistive technology, checkout abandonment, number of contacts with customer service due to accessibility issues, accessibility defects per release, time to fix accessibility issues, retention of users of AT and number of users eligible for AT contracts. It's also important to gather qualitative evidence since a technically correct interface can become confusing, tiring and embarrassing. The intent of the measure is to not reduce access to revenue only. It is also important to find out if accessibility investments are leading to equity of use and improvements in business performance.

Business Cases: From Specialist Need to Mainstream Value

One of OXO Good Grips's noted success stories in inclusive physical-product design is the history of the brand. OXO’s co-founder and renowned industrial designer, Sam Farber, noticed that his wife, Betsey, found a conventional metal vegetable peeler painful to use because of arthritis, so he developed an easier-to-use peeler. The resulting Good Grips handle was bigger, softer and easier to control, but not offered as a medical aid (OXO). Instead, it is a well-designed kitchen tool. This is the case because it shows that a usability weakness in accessing parts of a system can be exposed by a case and help show that the whole system could be weaker. Thin metal handles are excellent for precision grip, and are not very comfortable in repetitive activities. For some users, the introduction of a new handle design might be a helpful option due to arthritis, reduced dexterity, damp hands or for being simply more comfortable.

But OXO's lesson is not that all the concepts involving disability will go to the masses. What stands out in this case is the series of decisions that would have had to be made. It was decided to exclude something that was observed; it was decided what was the underlying

functional problem; it was decided to integrate the solution into the mainstream product; and it was decided that it would be communicated as performance not charity. There was no requirement for the product to state that it is a "disability product". This lifted the stigma, and helped to broaden market demand. The case provides a glimpse for businesses of how accessible design can help achieve the differentiation of a premium product when the end result can be easily seen by many consumers.

There's another case with Xbox, which is its Adaptive Controller. While using the traditional hand-held controller involves using the buttons, triggers, thumbsticks and both hands, the traditional controller must be used with a coordinated combination of all those elements. These assumptions and the most likely features are not applicable to many players with less mobility. The Xbox Adaptive Controller has simplified, centralized controls and ports to enable external switches, buttons, mounts and joysticks to be set up around an individual player (Microsoft, “Xbox Adaptive Controller”). It is particularly important for the commercial applications due to its modularity. The product isn't the controller that will fit all of the bodies; it's a hub in a greater system.

Co-design is also a crucial aspect of the controller and its impact. If the product is to be adaptive, the assumption of disability has to be solid; that is, it must not be universal or generic but rather identify specific disabilities. Hands, feet, shoulders, head moves or assistive switch may all be used, as well as needing to change depending on the different players. Working from the start with disabled users and specialist groups is the key to inclusive design, throughout development and testing. This product could expand the type of parties that can be engaged in gaming, enhance the program's ranking, and establish demand for suitable add-ons and settings for the software. During the same period, there are some constraints: adaptive equipment can be costly, there may be a need to use technical expertise for set-up, and game software itself needs to offer alternatives like the ability to remap, adjust difficulty, include subtitles, and even offer visual alternatives. Customer access to hardware is just one aspect of the customer journey.

The third and more broad business case is digital accessibility. Digital systems are used in virtually every sector, and accessibility is not only a key component of a website and application, but also a vital aspect of every self-service system—particularly in today's digital era. The Web Content Accessibility Guidelines are guidelines around four principles to keep content accessible (World Wide Web Consortium, WCAG 2.2): content should be perceivable, operable, understandable and robust. These are all commercial questions, answers to which should be based on these principles. Will it be possible to acquire the product information across to the customer range? Can interface be operated without using a mouse? Is the instruction and error easily understood? Is the service compatible with currently available assistive technologies and are there any assistive technologies being planned for development? If any of the following fails, then the transaction will not take place.

Whereas a single adaptive product is only a component of a digital accessibility, digital accessibility is a continuous production process. However, barriers can be brought back due to content changes, software updates, third party widgets and vendor updates. This is particularly

the case with regard to governance. An organisation may need design components that require accessibility features, editorial advice, purchasing provisions, automated testing, expert review, and testing with disabled users. There are tools that can be automated to lookup for missing labels, contrast problems, structural issues, etc., but they cannot determine whether there is usable alternative text, or whether a workflow is cognitively manageable. The micro-economic upsides are repeatability: with a more robust system, the access to an experience could be provided across numerous products at more marginal costs to fix it in an emergency.

There are three cases, reflecting the various ways to the value. If the access need is the basis of a specific choice of product differentiation, then OXO demonstrates mainstream product differentiation. The Xbox Adaptive Controller represents a user-diversity-principled platform that's customizable. Digital accessibility demonstrates an organization's capability throughout the whole service system. Of the three, the clear winner is accessibility that influences the basic design as opposed to just being tacked on as a lesser version. They also display that there is no uniformity in accessibility. There are some needs that can be provided by general-purpose, non-specific provision; others that either must be tailored to the individual, compatible and/or require special equipment. Any strategy that is going to be commercially successful needs to be used in both of these ways.

Implementation Barriers and Strategic Risks

Many organizations underinvest in accessibility, and even as part of a business case. The first reason may be that it is hard to detect excluded customers in traditional analytics. If a person is not able to register, he or she will only be shown as a "disconnect" or an “abandoned session”. Inaccessible updates can be made by the user and if the user exits this without accessing it, it may not be reported. Customer-service teams may have to find their own solutions to their issues without sharing with product teams to identify patterns. When there are no complaints, this can lead to confusion that it means that there are no barriers.

Fragmentation of organizations is the second barrier. If it fails, it is the customer service's responsibility; if it is required, it is the role of the product teams to control interfaces; if it is required, that is the responsibility of the marketing control campaign. This is to say that the interface is controlled by product teams, the requirement is interpreted by the legal teams and the vendor is selected by the procurement teams. These can happen between the functions of accessibility. A website can be audited and PDFs not accessible; a product can be accessible to use during the setup process, but not while using a PDF. If the ownership and standards for the local improvement activities are unclear, local improvement doesn't result in a journey that is accessible. Thirdly, symbolic inclusion is a risk. The inclusion of disabled consumers in advertisements and the absence of disabled employees from product research and development and/or from leadership roles is common in businesses. They can set up an accessibility tab or form without adding the necessary tasks. However, Hamraie cautions that universal-design language can turn into political hackneyed-speak if it does not draw on the lived experiences and rights of disabled people (95–101). Symbolic action is very fragile in the business world, as the

user is able to compare the promotion with what is experienced in practice. Co-design should not only be consultative, but also participatory and include research as a key component, and have a meaningful impact on decision making, rather than just after a significant decision.

The fourth barrier is the assumption that compliance is equal to quality. Standards have to be there as it helps provide measurable requirements, but that doesn't ensure a good experience. Technically correct headings which may be in an odd order can be placed on a page. Alternate text might be available, but may be describing something that has no particular relevance. Captions may be present but inaccurate. An extension control that is hard to locate, may be included in a time limit. Standards should be adopted as a benchmark and tasks and user research should be used in conjunction with them.

Lastly, in some cases, companies may have a difficult time showing quick ROI. Some benefits accumulate over time, like reduced risks, improved retention and re-usability capability. Accessibility work also sheds light on some technical debt that may have been lurking in the background, which just makes it seem like it's more expensive to do projects at the beginning, versus routine projects. Those in charge should understand that accessibility doesn't require a huge expense, whereas remedying years of non-accessible choices does. They should also not require a disabled customer to establish a separate proof of any extra gain or loss for the same provision as other mainstream design features do not need to be proven to justify the same provision. The best approach is to join the right to the responsibility to business measurements, not oppose them.

A Strategy for Converting Accessibility into Business Capability

The first rule of businesses looking to commercialize the potential of accessible design is to start out with a clear customer journey and a clear market hypothesis. There are many ways to improve access, and it is hard to prioritize them all and direct investment. A more robust objective may be to decrease the abandonment rate in an online account opening scenario for keyboard users and screen readers, or make a software product institutional-procurement ready due to reduced grip strength of the user. Specific objectives enable the link between the needs of users, the design decisions and the outcomes for the business.

The second step is to involve disabled people in research and development. Recruitment of participants should be geared to the likely consumers of the product, and should offer various access methods rather than view one participant as being representative of all disabilities. The information used for research, the consenting process, meeting platforms, prototypes and payment systems need to be accessible. Co-design doesn't imply users are expected to be co-creative with the final solution. It involves putting knowledge of barriers into practice when addressing requirements, trade-offs and testing.

Thirdly, organizations must incorporate accessibility into most systems that are reused. Digital teams have the ability to develop user-friendly components, coding conventions, content designs, and test processes. Physical-product teams can offer input on the matters of force, reach, tactile feedback, labelling and packaging. There may be evidence requirements for procurement

teams for accessibility and procedure to fix defects. The training needs of executives are different from that of designers and developers, and writers and quality-assurance teams also require role-specific training.

Fourth, monitoring of performance should be done in terms of access and business indicators. Conformance findings are useful, but they should be linked to actual tasks. Firms can monitor if a disabled user can perform their high value journeys, the amount of time it takes for disabled users to complete these journeys and where support is needed, as well as if regressions are caused by releases. Conversions, customer trust, remediation time, retention, support cost and procurement success are some of the business metrics. Results should be checked on a regular basis and results honestly reported. Developing accessibility statements that identify known limitations, workarounds, and contact channels, and planned improvements, are more credible.

Last but not the least, accessibility needs to be considered as a product feature as opposed to a campaign. The goal is not the ideal end goal, as technologies and the context of users change. The goal is an institution that recognizes and addresses exclusion at an early stage, and enhances without any crises incurred while responding to feedback. That capacity can enable market expansion, drive innovation and resilience, and pay respect to disabled people as customers with equal rights to the market.

Conclusion

Many current products still contain unnecessary barriers. The disabled population is large and diverse, and accessible features may also benefit users beyond the groups for whom they are designed. All these factors mean that accessible design has great potential. The benefits of this commercial value are increased. Usability, innovation, and trust are boosted. Remediation cost is reduced, and procurement and regulatory readiness are improved. These mechanisms might function in a variety of different ways, as evidenced by the cases of OXO Good Grips, Xbox Adaptive Controller, and mature digital-accessibility programs.

It is important to not create a business case that is justified by the false premise that accessibility always equals profits. Disabled consumers are wide-ranging, it takes investment to enable implementation, and poor products are not going to be competitive with their symbolic features. Durable value is achieved when businesses identify how some customers are being excluded, consult with disabled customers and incorporate accessibility needs into the fabric of their service, or track full customer experiences. While standards are an important reference point, a product or service will be commercially successful if it is usable, reliable and organizations are accountable.

The broader implication is conceptual. In valuing people with disabilities, they should not be considered as an add-on to an "otherwise complete" market. They are consumers and employees of that market, producers, and decision makers. By using their experiences, design weaknesses can be identified, which are not captured by the conventional assumptions. Companies that take such actions can create a more equitable, more flexible and sometimes more competitive product and service.

Works Cited

Accenture. Getting to Equal: The Disability Inclusion Advantage. Accenture, 2018.

Blackwell, Angela Glover. “The Curb-Cut Effect.” Stanford Social Innovation Review, Winter 2017.

Clarkson, P. John, et al., editors. Inclusive Design: Design for the Whole Population. Springer, 2003.

European Parliament and Council. “Directive (EU) 2019/882 of 17 April 2019 on the Accessibility Requirements for Products and Services.” Official Journal of the European Union, L 151, 7 June 2019, pp. 70–115.

Hamraie, Aimi. Building Access: Universal Design and the Politics of Disability. University of Minnesota Press, 2017.

Lazar, Jonathan, Daniel F. Goldstein, and Anne Taylor. Ensuring Digital Accessibility through Process and Policy. Morgan Kaufmann, 2015.

Microsoft. Inclusive 101 Guidebook. Microsoft Inclusive Design, 2021.

Microsoft. “Xbox Adaptive Controller.” Xbox, www.xbox.com/en-US/accessories/controllers/xbox-adaptive-controller.

OXO. “Our Story.” OXO, www.oxo.com/our-story.

Persson, Hans, et al. “Universal Design, Inclusive Design, Accessible Design, Design for All: Different Concepts—One Goal? On the Concept of Accessibility—Historical, Methodological and Philosophical Aspects.” Universal Access in the Information Society, vol. 14, 2015, pp. 505–526. doi:10.1007/s10209-014-0358-z.

Story, Molly Follette, James L. Mueller, and Ronald L. Mace. The Universal Design File: Designing for People of All Ages and Abilities. Center for Universal Design, North Carolina State University, 1998.

United Nations. Convention on the Rights of Persons with Disabilities. 13 Dec. 2006.

WebAIM. The WebAIM Million: The 2024 Report on the Accessibility of the Top 1,000,000 Home Pages. 2024, webaim.org/projects/million/2024/.

Williamson, B. (2019). Accessible America: A history of disability and design. In Accessible America. New York University Press.

World Health Organization. Global Report on Health Equity for Persons with Disabilities. World Health Organization, 2022.

World Health Organization and World Bank. World Report on Disability. World Health Organization, 2011.

World Wide Web Consortium. “The Business Case for Digital Accessibility.” Web Accessibility Initiative, www.w3.org/WAI/business-case/.

World Wide Web Consortium. Web Content Accessibility Guidelines (WCAG) 2.2. W3C Recommendation, 5 Oct. 2023, www.w3.org/TR/WCAG22/.

The Effects of the Padel Industry on Pakistan's Sporting Industry

By Khazain Hassan Mansha

Abstract

This paper examines the rapid growth of padel in Pakistan as both a sporting activity and an emerging business industry. It asks whether padel can develop into a sustainable part of Pakistan's sports-service economy or whether rapid investment could create a short-term market bubble. The study uses a descriptive case-study approach based on secondary data from international padel bodies, government institutions, industry reports, commercial operators, and market listings. It also applies an illustrative financial model to a two-court facility to compare setup costs, operating expenses, revenue, profitability, and break-even conditions. The findings indicate that padel benefits from strong global growth, increasing institutional recognition in Pakistan, and a business model capable of generating multiple revenue streams. However, profitability depends heavily on court utilisation, pricing, land costs, and repeat demand. The study concludes that padel has long-term potential in Pakistan, while poorly planned facilities remain exposed to oversupply. Policy priorities include wider access, public-private partnerships, junior development, coaching, and careful infrastructure planning.

Keywords: padel; Pakistan; sports industry; business; investment

1. Background and Rationale

1.1 Introduction

This case study will look at Pakistan's padel industry not only as a form of entertainment or physical activity but as a business. While most of us see sports as a way to keep entertained or stay fit, in many countries it is also seen as an industry. In many countries, sport is treated as an important economic sector because it creates jobs, attracts investment, supports tourism, develops infrastructure, and gives young people opportunities to build skills. Today in Pakistan however the main focus of the people and the government when it comes to sporting is cricket. Cricket dominates media coverage, sponsorships, public attention, and investment. Due to this other sports struggle to grow however in more recent time there has been an outlier to this trend this sport is padel. Padel has become a massive sport very quickly not only across the globe but in Pakistan which is rare when cricket is so dominant therefore padel is important because it shows how a new sport can enter the market and create business opportunities in a country where the sports industry is still developing as controlled by one sport.

This case study will therefore examine how padel is changing Pakistan’s sports industry through private investment, court construction, club memberships, coaching, tournaments, and digital bookings. It will also look at whether padel can become a wider sport in Pakistan or whether it will remain limited to wealthier urban areas.

1.2 Background

Firstly we will look at the Pakistani sporting industry as a whole. Though Pakistan's history has been an important part of the community however throughout that same history Pakistan has been dominated by one sport at a time with hockey being the dominant sport up until the 1990s with Pakistan winning 3 Olympic gold medals and winning 4 world cup (Olympics.com, 2026). After the 1990s cricket became the dominant sport with 3 world cups and 2 asia cups. Cricket still dominates the country on television rights and sponsorships to academies, merchandise, and local tournaments, cricket has shaped the way sport is commercialised in the country. However, this heavy dependence on one sport has also limited the growth of other sporting markets. Football, tennis, squash, boxing, athletics, and other sports have often struggled to receive the same investment, attention, and infrastructure.

Besides just the cultural aspects of sports in Pakistan it also plays a big part in the economy. Pakistan exported around $444 million worth of sports equipment in 2024, making it one of the world's significant sports equipment exporters (OEC, 2024). More recent government linked data also showed that Pakistan exported $207.8 million worth of sports goods during July–December 2025-26, showing an 11.1% increase compared to the same period the previous year (APP, 2026). These are big figures for the Pakistani economy. In fact, Pakistan famously produces 70 percent of the world's football in just one city (Business Insider, 2026).

This shows that Pakistan is already important in the global sports economy, but mostly as a manufacturer rather than as a country with a strong local sports service industry. Pakistan can make and export sports goods, but locally the business side of sport is still limited. For example, sport in Pakistan has not fully developed into a large industry of paid clubs, coaching centres, private leagues, sponsorship events, and recreational venues. This is the gap where padel becomes important.

These figures show that Pakistan is already involved in the global sports business, especially through manufacturing. There is a gap when it comes to the local sports service industry which includes clubs, coaching, court rentals, fitness spaces, and tournaments all of which is under developed. Padel can help fill this gap due to this rapid growth and strong interest.

Padel is different from many sports because it can work as a direct business model. A cricket ground or football field usually needs a large space and many players, but padel can be played on a much smaller court. This makes it easier for private investors to build facilities in urban areas where land is expensive. A standard padel court is around 20 metres by 10 metres, meaning that several courts can fit into a relatively small sports complex (FIP, 2025). This makes padel attractive in cities such as Lahore, Karachi, and Islamabad, where private sports spaces are growing but land costs remain high.

1.3 Introducing Padel

Padel is a racket sport that combines elements of tennis and squash. It is usually played in doubles on an enclosed court with glass walls. One reason for its global success is that it is easier

to learn than many traditional racket sports. Part of its rapid growth is that it is easy to play. A beginner can usually start enjoying the game quickly, while tennis or squash often require more training before people can participate freely. This leads to interest from people of all walks of life who may not have much time and want an easy way to stay fit or a fun social activity. According to the International Padel Federation, padel now has more than 35 million players globally, with a 16.1% increase in clubs, a 15.2% increase in courts, and a 42% increase in registered federation members compared to the previous year (FIP, 2025). These figures show that padel is not just a short term trend. It is becoming a fast growing global sports industry.

The global business case for padel is also strong. The Global Padel Report 2026, produced by Playtomic and Strategy& by PwC, estimated that there were around 58,300 padel courts, nearly 20,900 clubs, and 19.4 million players worldwide in 2025. The same report projected that the number of courts could reach 91,000 by 2028 (Playtomic and PwC, 2026). These figures matter because they show that padel is not only growing as a sport, but also as a commercial market. The report also found that successful padel clubs in established markets can reach around 70% court utilisation, 55% gross margins, and a payback period of around three years (Playtomic and PwC, 2026). This means that padel can be studied like a proper investment, not only like a game.

This rapid growth has now come to Pakistan where padel has moved from being almost unknown to becoming one of the fastest-growing private sports trends in major cities. Unlike cricket, hockey, or squash, padel in Pakistan has not grown mainly through schools, government teams, or national sports programmes. Instead, it has grown through private groups (Padelverse, 2026; Pakistan Sports Board, 2026).

This is important because it shows a new route for sport in Pakistan. Instead of a sport first becoming famous through a national team, padel is growing through businesses. Entrepreneurs are building courts, creating clubs, charging hourly fees, offering memberships, hiring coaches, and organising tournaments. In this way, padel is becoming part of Pakistan’s private recreational economy. It is similar to how gyms, football turfs, and indoor sports arenas became popular in major cities, but padel has the advantage of being new, social, and connected to a global trend.

The growth of padel is Pakistan has allowed private groups and the sporting industry to gain over all and has helped fill a critical gap in Pakistan's sporting sector. It has also given the pakistan people a new accessible way to a sport which was not the case in previous sports in Pakistan such as hockey and cricket. The clearest evidence of padel’s growth in Pakistan is the spread of dedicated padel clubs and courts in cities such as Lahore, Karachi, and Islamabad. Padelverse, for example, advertises itself as having seven courts and describes itself as the largest panel facility in Pakistan (Padelverse, 2026).

2. Research Questions and Objectives

Research question: How is the growth of padel affecting Pakistan's sports-service industry, and under what conditions can padel facilities operate as sustainable businesses rather than a short-term investment bubble?

Objectives: This paper aims to assess the growth of padel in Pakistan; examine the cost, revenue and operating-profit structure of a padel facility; evaluate the risk of overinvestment and oversupply; and identify policy measures that could support wider and more sustainable participation in the sport.

3. Review of Literature

To understand how padel is affecting Pakistan's sports industry we need to look at what other people have said about it. We have to consider three things: how padel is doing in Pakistan, how it is growing around the world and if padel clubs can make money in the long run. From what we can see, padel is different from sports in Pakistan because people pay directly to play, take lessons and buy equipment. Padel is not about playing games or watching them, it is also a business where people pay for services.

3.1 Pakistan's Sports Economy and the Emergence of Padel

Pakistan already has a sports industry but most of the money comes from making sports equipment. In 2024 Pakistan sold $444 million worth of sports equipment to other countries. From July to December 2025 Pakistan sold $207.803 million worth of sports equipment, which's 11.10% more than the year before. This shows that Pakistan is good at making sports equipment. It does not mean that people in Pakistan are spending a lot of money on sports. Padel might be changing this. There are companies like Padelverse that have courts to teach lessons, rent out equipment and allow people to book online. There are also websites like Playtomic that list padel facilities in Pakistan, such as Padel in Pakistan and Padel One in Lahore. The Pakistan Sports Board is also getting. Has asked for proposals to build panel facilities in Islamabad, Lahore and Karachi. Padel is helping to grow Pakistan's sports economy by getting people to spend money on sports services (OEC, 2024; APP, 2026; Padelverse, 2026; Playtomic, 2026; Pakistan Sports Board, 2026).

3.2 Global Growth and Evidence of Long-Term Demand

If we look at what's happening in other countries we can see that padel is getting very popular. The International Padel Federation says that there are over 35 million amateur players and that the number of clubs, courts and players is growing fast. Another company, Playtomic says that there are about 19.4 million players, 58,300 courts and 20,900 clubs around the world. They think that the number of courts will reach 91,000 by 2028 (FIP, 2025; Playtomic and Strategy&, 2026).

In Britain the number of people playing padel grew from 15,000 in 2019 to one million in 2026. This shows that padel can become very popular if there are facilities, lessons and competitions.

Many international sources are saying that padel is growing fast both in terms of people playing and the number of facilities.

3.3 Padel as a Commercial Business

There is also a lot of interest in investing in padel. Some companies say that padel clubs can make a lot of money if they are run well with up to 70% of courts being used and profit margins of around 55%. However, the location of how many people use the courts. The prices are all very important (Playtomic and Strategy&, 2026).

In Pakistan it can cost around PKR 10-15 million to build a court and PKR 12-18 million to build a premium court. Some companies are already making money from padel with prices ranging from PKR 4,000 to PKR 5,000 per hour. Padelverse is an example of a company that makes money from a combination of court rentals, lessons, equipment and other services (Unicorn Flooring, 2025; Pure Padel, 2026; PlayPro, 2026; Padelverse, 2026).
The potential for making money from padel is high. It also requires a big initial investment and there are significant costs to consider.

3.4 Institutional Development in Pakistan

To see if padel can become a part of Pakistan's sports industry we need to look at if there are any institutions being set up. The International Padel Federation has recognized the Pakistan Padel Federation as a member. They are now keeping track of rankings and organizing competitions. Pakistan has also entered the competition structure with events like the FIP Promises Karachi and the FIP Bronze Pakistan I (FIP, 2026; Pakistan Padel Federation, 2026). The fact that there are now institutions, rankings and international tournaments in Pakistan shows that padel is becoming established.

3.5 Sustainability and the Bubble Argument

Just because padel is growing fast does not mean that every investment will be successful. In fact there is a risk that many courts might be built leading to a surplus and some clubs closing down. This has already happened in countries like Sweden. In Britain the number of courts grew from than 500 to over 1,800 in just two years while the number of players passed one million. This shows that the growth of padel can be very fast. It also means that there is a risk of overinvestment. For Pakistan the key to sustainability will be if the growth in courts is matched by growth in repeat players. Facilities that offer a range of services including lessons, competitions and social activities might be better protected from declining demand, than those that just rely on court rentals. Padel seems to be a sport but there is still a risk that rapid investment could lead to oversupply in the market (Financial Times, 2026; LTA, 2026).
This case study has examined the growth of padel in Pakistan not simply as a new sport, but as an emerging business industry. The evidence shows that padel has entered Pakistan at a time when the country’s sports economy has traditionally been dominated by cricket and by the manufacturing and export of sporting goods. Pakistan already has a strong position in the global

sports economy through exports, producing hundreds of millions of dollars worth of sporting equipment each year (OEC, 2024; APP, 2026). However, the domestic sports-service industry, including private clubs, coaching, court rentals, tournaments and recreational facilities, has remained relatively underdeveloped. Padel is beginning to fill this gap.

4. Methodology

4.1 Research Design

This paper uses a descriptive case-study design. It combines market and institutional evidence with an illustrative financial scenario for a two-court padel facility in Pakistan. The purpose is not to estimate a guaranteed return on investment, but to examine how the business model changes under different assumptions about costs, pricing and court utilisation.

4.2 Data Sources

The analysis relies on secondary data. Sources include the International Padel Federation, Pakistan Padel Federation, Pakistan Sports Board, Pakistan Bureau of Statistics data reported by APP, the OEC, Playtomic and Strategy& by PwC, the LTA, NEPRA, the Punjab Labour Department, commercial court operators, property listings and court-installation suppliers. No primary survey or interview data were collected for this paper.

4.3 Theoretical Framework and Empirical Strategy

The paper treats padel as a sports-service business and evaluates it through unit economics and market-sustainability analysis. The empirical strategy compares fixed and variable costs with booking revenue under different utilisation scenarios, then uses international evidence from expanding and oversupplied padel markets to assess whether growth in Pakistan is likely to be sustainable.

4.4 Variables and Measurement Approach

The main variables are court construction cost, required space, rent, staffing, utilities, maintenance, hourly booking price, booked hours per court per day, monthly revenue, operating expenses, operating profit and break-even utilisation. The financial model uses the assumptions stated in the Results and Discussion section, including a two-court facility and a premium PKR 10,000 hourly booking price.

5. Results and Discussion

5.1 Padel as a Business

We can all clearly see the rapid growth of Padel in pakistan but now the big question is can it be a sustainable business? Unlike several other sports padel has a very distinct and clear commercial model. There are several ways through which a padel club can generate income such

as hourly court hire, memberships, coaching sessions, hosting tournaments and competitions, corporate events, partnerships and sponsorship, club merchandise and equipment sales, food and beverages and more. This makes a padel facility more than simply a place to play, but rather a fully operational business with the potential to generate various streams of income.

This is a particularly salient aspect in Pakistan as the local sports services industry is yet to mature. While Pakistan is a major player in the international sports sphere due to sports manufacturing, the business of private sporting clubs, private leagues, coaching programs, and recreational services, is currently a minor player in the economy. Padel's business model helps to fill this gap by transforming sport into a service offering. Instead of relying solely on the sale of merchandise to spectators or the performances of national teams, a padel club monetizes customer interaction by charging a fee for access to the sport.

The distinct commercial model of padel explains the interest that has begun to be displayed by private investors. Data from global reports indicates that high court utilization rates and cost efficiencies can result in a high performing business for padel clubs (Playtomic and PwC, 2026). In Pakistan the enthusiasm displayed by private clubs, as well as government entities like the Pakistan Sports Board, illustrates that padel has begun to be seen as a valuable investment opportunity rather than a niche hobby (Pakistan Sports Board, 2025). Therefore, this section will focus on padel as a business by examining revenues, costs, potential profit, risks, and the future viability of the market within Pakistan.

The first major financial requirement when establishing a padel business in Pakistan is the cost of building the courts. A Pakistani court supplier estimates that a standard padel court can cost between PKR 10 million and PKR 15 million, while a premium court made with higher-quality glass and steel may cost between PKR 12 million and PKR 18 million (Unicorn Flooring, 2025). Therefore, a basic two-court facility could require between PKR 20 million and PKR 30 million for the courts alone. A premium two-court facility could cost between PKR 24 million and PKR 36 million.

These estimates might not cover some costs such as preparing the land, drainage work, building foundations, creating parking spaces, adding changing rooms, washrooms, reception areas, seating areas, cafés, generators, air conditioning systems and other equipment. When all these costs are added, a good estimate for a two-court facility would be around PKR 25 million, to PKR 46 million not including the cost of buying the land.

Space is another factor to think about when it comes to money. A normal padel court is 20 meters long and 10 meters wide. The Pakistan Sports Board says that one court needs 2,486 square feet when the area where people play and the space around it are all counted (Pakistan Sports Board, 2026). Two courts would need at least 4,972 square feet.. When you add walkways, a reception area, storage rooms, washrooms, seats, parking spaces and maybe even a café, a business club with two courts might need between 7,000 and 10,000 square feet.

The cost of this space can have a major effect on profitability. One Lahore commercial property listing advertised a 10,000-square-foot space in Johar Town for approximately PKR 1.6 million per month (Zameen, 2026). This is one example and not the national average because

rent changes a lot based on the city and the place. A club in DHA, Gulberg, Clifton or another high-end area might have higher rent, than a place built in an industrial area or a less central spot. This means people who already have the land have a big financial advantage.
have skills is now PKR 40,000 per month and this started in July 2025 according to the Punjab Labour Department in 2025. A small club with two courts needs a lot of staff like a manager, two court attendants, a security staff person to work at the reception and one or more coaches who teach people how to play padel. Even if you pay them salaries that're not too high it can still cost a lot of money, around PKR 300,000 to PKR 600,000, per month. The exact amount of money you have to pay depends on how you pay the coaches whether you give them a fixed salary or a part of the money they make from teaching people how to play padel (Punjab Labour Department, 2025).

Electricity is also very important because a lot of people want to play padel in the evening. Usually you need eight lights that use 200 watts each to light up one court and this uses about 1.6 kilowatts of electricity per court according to Olympia Courts in 2026. If you turn on the lights for two courts for eight hours every day just the lights alone will use a lot of electricity.

$$1.6\ \text{kW} \times 8\ \text{hours} \times 30\ \text{days} \times 2\ \text{courts}$$
$$= 768\ \text{kWh per month.}$$

The cost of utilities for a padel club does not include things like fans, air conditioning, reception lighting, water pumps, café equipment, security systems or backup power.

Commercial electricity charges in Pakistan can change because of things like tariff adjustments, fixed charges, taxes and fuel price adjustments as stated by the National Electric Power Regulatory Authority in the year 2026 (NEPRA, 2026).

A good budget for utilities for a padel club with two courts could be between PKR 100,000 and PKR 250,000 per month.

This really depends on whether the padel club's outdoor, covered or fully indoor.

Padel clubs also need to think about maintenance costs.

Padel courts need things like cleaning the glass, brushing the turf, checking the net, cleaning the drainage, fixing the lights and taking care of the steel structures.

The artificial turf on a padel court may need to be replaced after five to eight years.

This depends on how much the court's used, the weather and how well it is maintained according to I-Padel in the year 2026.

So a padel club should put aside some money each month for maintenance and replacement.

This could be around PKR 50,000 to PKR 150,000 per month, for a padel club.

The main way a padel club makes money is from renting out courts by the hour. For this example we will say that the average cost to rent a court is PKR 10,000 per hour. This is a price. It is based on the idea that the facility's, in a rich area has good courts and offers good services to support the courts. Padel club owners can make a lot of money from court rental at a padel club.

A two-court facility may have the following estimated monthly operating-cost structure:

Rent: PKR 700,000
Staff salaries: PKR 350,000
Electricity and utilities: PKR 150,000
Maintenance reserve: PKR 100,000
Marketing, booking software and administration: PKR 150,000
Security, repairs and other costs: PKR 100,000

This creates estimated monthly operating expenses of approximately PKR 1.55 million. These figures are assumptions used to demonstrate the business model rather than official national averages. Actual costs will depend heavily on the city, location, building design, staffing structure, and quality of the courts.

If both courts are booked for five hours each day, monthly revenue would be:

$$\text{PKR } 10{,}000 \times 5 \text{ hours} \times 2 \text{ courts} \times 30 \text{ days}$$
$$= \text{PKR } 3 \text{ million per month.}$$

After deducting PKR 1.55 million in operating expenses, the business would make an estimated operating profit of:

$$\text{PKR } 3 \text{ million} - \text{PKR } 1.55 \text{ million}$$
$$= \text{PKR } 1.45 \text{ million per month.}$$

This would equal approximately PKR 17.4 million in annual operating profit, assuming demand remained stable throughout the year.

If each court is booked for seven hours per day, revenue would be:

$$\text{PKR } 10{,}000 \times 7 \text{ hours} \times 2 \text{ courts} \times 30 \text{ days}$$
$$= \text{PKR } 4.2 \text{ million per month.}$$

After deducting operating costs, estimated operating profit would be:

$$\text{PKR } 4.2 \text{ million} - \text{PKR } 1.55 \text{ million}$$
$$= \text{PKR } 2.65 \text{ million per month.}$$

This would equal approximately PKR 31.8 million in annual operating profit.

If each court is booked for eight hours per day, revenue would increase to:

$$\text{PKR } 10{,}000 \times 8 \text{ hours} \times 2 \text{ courts} \times 30 \text{ days}$$
$$= \text{PKR } 4.8 \text{ million per month.}$$

Estimated operating profit would then be:

$$\text{PKR } 4.8 \text{ million} - \text{PKR } 1.55 \text{ million}$$
$$= \text{PKR } 3.25 \text{ million per month.}$$

This would equal approximately PKR 39 million in yearly operating profit if demand remained stable throughout the year.

A stronger-performing club with ten booked hours per court each day would generate:

$$\text{PKR } 10{,}000 \times 10 \text{ hours} \times 2 \text{ courts} \times 30 \text{ days}$$
$$= \text{PKR } 6 \text{ million per month.}$$

Estimated operating profit would be:

$$\text{PKR } 6 \text{ million} - \text{PKR } 1.55 \text{ million}$$
$$= \text{PKR } 4.45 \text{ million per month.}$$

This would equal approximately PKR 53.4 million in annual operating profit
Under this model, the break-even point is approximately 2.6 booked hours per court per day. At this level, the revenue produced by the courts would approximately cover the PKR 1.55 million monthly operating cost. Any bookings above this level would begin to generate operating profit. This shows that court utilisation remains one of the most important measurements for a padel business, even when the club charges a premium price.

Overall, padel can be a highly profitable business in Pakistan under a premium pricing model, but only under the correct conditions

5.2 Padel as a Bubble

Padel is growing fast in Pakistan.. We have to ask if this growth will last or if the industry is just a big bubble that will burst. A bubble happens when people invest in something just because it is popular and seems like it will make a lot of money. This causes people to build many padel courts too quickly. If this happens some courts might struggle to stay open when people lose interest in the sport.

There are some reasons to think that this could happen in Pakistan. Building a padel club costs a lot of money. Around PKR 25 million to PKR 46 million for two courts. People might invest in padel because they think they can make a lot of money from it. If a court is booked for five hours a day it could make around PKR 3 million per month.. This only works if people keep paying PKR 10,000 per hour to play.

This price is very high for people in Pakistan. If more clubs open they might have to lower their prices to compete. If the price goes down to PKR 5,000 per hour the court will make as much money.. The costs of running the court like rent and electricity will stay the same. So a court that seems profitable now might not be in the future.

Another thing to consider is that padel is mostly popular in areas of Lahore, Karachi and Islamabad. There are investments being made in padel like Padelverse with its seven courts. The Pakistan Sports Board is also building courts.. Just because more courts are being built it does not mean that more people are playing padel regularly.

With these risks padel is not necessarily a bubble. It is part of an international trend. There are over 35 million padel players worldwide. The number of courts and clubs is growing fast. In Asia the number of players and courts grew by around 20% each year (FIP, 2025). This suggests that padel in Pakistan is part of a term of global growth, not just a local fad.

Padel has some advantages that might help it last. It is easy to learn. People of all ages can play. It is also a sport that combines exercise with fun. Pakistan is starting to organize competitions, which can help create a community of players.

So padel, in Pakistan has elements of both an industry and a possible bubble. The sport will probably stay. Not all clubs will succeed. Clubs that rely on prices and do not have a strong community might struggle.. Clubs that have lower costs, strong communities and multiple sources of income are more likely to survive. The main bubble might not be padel itself. The idea that every new padel court will automatically make money.

6. Conclusion and Policy Implications

6.1 Policy Implications

The growth of padel in Pakistan should be celebrated as it has helped to break a long standing trend of one sport being prominent In Pakistan while also encouraging physical activity in Pakistan which is critically important when you look at the lack of options for physical acuity in Pakistan and its effects on pakistanis health.

Demand for sports other than cricket exists and is growing. Padel, a recreational activity of a private individual, can be facilitated by the government by allocating land for sports facilities, simplifying approval process and encouraging public-private partnerships. The involvement of Pakistan Sports Board in development of padel already indicates the possibility of cooperation between public and private sectors (Pakistan Sports Board, 2026).

On top of this when talking about padel in pakistan Making Padel more accessible is key to greater policy impact. Padel is currently played in private clubs and most courts are located in areas of high average income. To have greater social impact, Padel courts need to be gradually introduced to schools, universities, public sports complexes and community centers. There are examples of how institutional support, investment in facilities and in coaches can lead to a huge rise in participation numbers, such as in Britain (LTA, 2026).

Thirdly, invest in coaching, junior development, referees and in creating competitive structures. The Pakistan Padel Federation is already doing this. Introducing national rankings and organizing competitions for example is a good start. Pakistan is hosting FIP-recognized events (Pakistan Padel Federation, 2026; FIP, 2026). With government support, these programs can be further developed.

Uncontrolled investment in facilities such as uncontrolled proliferation of padel courts when there are not enough regular participants of the sport is also undesirable. A policy of promoting sustainable participation in padel through proper planning, affordability and demand for facilities in the long term would be more appropriate.

6.2 Conclusion

The report analyses the growth of the sport of padel in Pakistan. Not only a sport to play but also an emerging business industry. Pakistan's sports economy has traditionally revolved around the sport of cricket. Pakistan is a major manufacturer and exporter of sports goods which fetch hundreds of millions of dollars every year in exports (OEC, 2024; APP, 2026). Therefore, the domestic sports-service industry which includes private clubs offering sports and recreation services, coaching, court rentals, tournaments and other sports facilities, remains underdeveloped in Pakistan. Padel is filling this void.

Sooner or later, there is also need of development of national sports of Pakistan, in the framework of participation of Pakistan Sports Board, International Padel Federation and Pakistan Padel Federation, and in the light of National Rankings of the country and holding of International FIP events in Pakistan (Pakistan Padel Federation, 2026).

Globally, the growth of padel is very rapid and the International Padel Federation estimates that there are currently more than 35 million amateur padel players all over the world (FIP, 2026). According to a study conducted by Strategy& by PwC, in collaboration with Playtomic, currently there are tens of thousands of courts around the world and over thousands of clubs. The number of padel courts and clubs is expected to grow in the next few years (Strategy& by PwC, in collaboration with Playtomic, 2026). The growth of padel can create a huge commercial market for the sport which is currently relatively unknown, as has already been observed in countries such as Britain, where the number of padel players increased from 15,000 in 2019 to one million in 2026.

In the analysis of a potential business opportunity, it is essential to determine why a particular business opportunity would be good for investment. For the case of padel courts, the owner can charge by the hour. The facility under consideration is a two-court facility. Two courts of a two-court facility can be booked for 8 hours each day. Using a financial model, the estimated monthly revenue of PKR 4.8 million can be generated. After deducting the estimated operating expenses of PKR 1.55 million from the monthly revenue, the facility would make an operating profit of PKR 3.25 million per month.

I want to emphasize that the numbers that I have presented are not indicative of the return on investment of a padel club. Instead, these are the potential earnings of someone who is running a padel club, dependent upon a number of variables, including the price that a customer would pay to play padel, and then how much they can charge for playing on different days of the week. The operator would also look at how much time he can get out of each of the courts in terms of number of hours, each day. So in the example that I gave before, a two court club with two 600 square meter courts, and each court able to be used for 8 hours a day, that means that there are 16 hours of use of the courts each day. Depending on the quality of the club, and the start-up costs of the club, a single padel court could cost as much as PKR 10 million to PKR 18 million (Unicorn Flooring, 2025). Therefore, a two-court club would require an initial investment of tens of millions of rupees before any consideration of the cost of the land on which the club is to be built, or of any rent that may have to be paid.

While the growth of padel has and continues to generate great interest regarding new facilities being opened everywhere, there is considerable concern regarding a ‘bubble’ forming amongst poorly thought out investment in padel facilities by inexperienced investors in the UK, whilst others are left struggling in an over supplied market in Sweden where the growth of padel experienced an initial and rapid growth.

A Bubble! The above question has been answered on a balanced basis. Padel is not a temporary trend of some sort. All over the world there are millions of amateur padel players. The competitive structure of padel is becoming more established. In Pakistan, the Pakistan Sports Board and the Pakistan Padel Federation are promoting the sport. There is an FIP event being held in Pakistan (FIP, 2026; Pakistan Padel Federation, 2026).

Padel is currently being played as a recreational activity in clubs by wealthier customers. However, to create a greater impact on Pakistan’s sporting culture, padel will need to graduate from a club activity to be played in schools, universities, in community centers and above all as a sports activity that can be played in facilities that are affordable for most Pakistanis.

In summary, padel has immense potential for growth in the sporting industry of Pakistan. The sport can generate investment, create employment, increase consumer spending on sporting goods, and build sporting infrastructure. In the long run, the most successful padel facilities and clubs would be those that have managed to keep a lid on their land costs, and maintained high levels of court utilization. In addition, such facilities would also offer coaching and organize tournaments. The padel enthusiasts could develop into a loyal following of the sport. Padel can become a permanent feature of the sports economy of Pakistan provided the industry is developed in a prudent and thoughtful manner.

Works Cited

Associated Press of Pakistan. (2026, January 26). Export of sport goods increased 11.10% to $207.803 mln. https://www.app.com.pk/business/export-of-sport-goods-increases-11-10-to-207-803-mln

Business Insider. (2026). The one city making 70% of the world's football. https://www.businessinsider.com/fifa-footballs-manufacturing-world-cup-testing-factory-2026-6

Financial Times. (2026). Britain's padel boom attracts novice investors. https://www.ft.com/content/26b507ee-e5db-4101-a13b-dd8447bbf302

International Padel Federation. (2025). FIP World Padel Report 2025. https://www.padelfip.com/world-padel-report-2025/

International Padel Federation. (2025, November 7). Passion for padel in Asia, steady growth (+20%) among courts and players. https://www.padelfip.com/2025/11/passion-for-padel-in-asia-steady-growth-20-among-courts-and-players/

International Padel Federation. (2026). FIP Bronze Pakistan I. https://www.padelfip.com/events/fip-bronze-pakistan-i-2026/

International Padel Federation. (2026). Pakistan Padel Federation. https://www.padelfip.com/federation/pakistan/

I-Padel. (2026). Replacing artificial grass on padel courts: Lifespan, maintenance and renovation. https://i-padel.nl/en/replacing-artificial-grass-on-padel-courts-everything-you-need-to-know-about-lifespan-maintenance-and-renovation/

Lawn Tennis Association. (2026). 1,000,000 padel players in Britain. https://www.ltapadel.org.uk/news/2026/may/1000000-padel-players-in-britain/

National Electric Power Regulatory Authority. (2026). Tariff and electricity-sector information. https://www.nepra.org.pk/

Observatory of Economic Complexity. (2024). Sports equipment in Pakistan. https://oec.world/en/profile/bilateral-product/sports-equipment/reporter/pak

Olympia Courts. (2026). OC lights. https://olympiacourts.com/lighting/oc-lights/

Olympics.com. (2026). Origins and a tale of two hockeys. https://library.olympics.com/

Padelverse. (2026). Padelverse. https://padelverse.net/

Pakistan Padel Federation. (2026). Pakistan Padel Federation. https://www.pakistanpadelfederation.org/

Pakistan Sports Board. (2025). Padel tennis development proposal. https://www.sports.gov.pk/

Pakistan Sports Board. (2026). Padel tennis courts: Tender and development information. https://www.sports.gov.pk/

PlayPro. (2026). Prime Padel Haus: Court availability and pricing. https://playpro.pk/court-availability/prime-padel-haus/

Playtomic. (2026). Padel in Pakistan, Lahore. https://playtomic.com/clubs/padel-in-pakistan-lahore

Playtomic & Strategy& by PwC. (2026). Global Padel Report 2026. https://playtomic.com/global-padel-report
Punjab Labour Department. (2025). Minimum wages notification. https://labour.punjab.gov.pk/minimum-wages-notification
Pure Padel. (2026). Pure Padel. https://purepadel.pk/
Unicorn Flooring. (2025). Premium quality padel court in Pakistan. https://epoxyflooring.pk/premium-quality-pedal-court/
Zameen.com. (2026). Commercial properties for rent in Johar Town, Lahore. https://www.zameen.com/Rentals_Commercial/Lahore_Johar_Town-93-1.html