Macalester Journal of Economics

Volume 34 - Spring 2026

Editors

Kudzhai Dhewa
Bailey Lengfelder
Adaye Yvan N'guettia
Linh Nguyen
Anna Settels Palacios
Alessia Velásquez Nitti

Omicron Delta Epsilon

Bailey Lengfelder, *President*
Anna Settels Palacios, *President*

Faculty Advisor

Felix Friedt, *Associate Professor of Economics*



Contents

Foreword
Felix Friedt 3

Fertility and Child Care Affordability: Evidence from CCDF Eligibility Threshold
Teagan Church and Audrey Mackay 5

Teacher Efficacy & Student Success in a Developing Country Context
Emma Runchey Smalley 19

Income as a Determinate of Mortgage Choice in the Conforming and Jumbo Markets
Daniel Wiltse 39

Medicaid Expansion and Private Health Insurance: Evidence on Crowd-out and Discontinuities at the 138% FPL Eligibility Threshold
Dong Nguyen 54

The Impact of Marijuana Legalization on Adolescent Drug Use and Educational Outcomes
Annie Hsu 73

How do Severe Winter Cold Shocks Impact Midwest Electricity Systems?
Jerry Zheng 86

How did the Covid-19 Pandemic Affect Household Entertainment Expenditure in the US?
Anna Fratto 102

Investigating Industry Sensitivity to Citywide Speed Limit Reductions
Aram Petrosyan 116

The Promise of Designation: Long-Term Impact of Opportunity Zone Designation on Employment Outcomes
Patrick Angelo Sursano 132

Foreword

The Macalester College chapter of Omicron Delta Epsilon, the international honors society in economics, proudly presents the thirty-fourth volume of the *Macalester Journal of Economics*. This year's editorial team—Kudzhai Dhewa, Bailey Lengfelder, Adaye Yvan N'guettia, Linh Nguyen, Anna Settels Palacios, and Alessia Velásquez Nitti—has selected nine papers that reflect the breadth and ambition of undergraduate research in the Economics Department at Macalester College. Paper topics range from child care and education to health insurance, housing finance, electricity systems, and local economic development. Across these diverse topics, the authors pursue a common question: how do public policies and other changing conditions shape the economic realities of households, labor markets and industries?

One recurring theme is the distance between a policy's intended objectives and its observed consequences. Teagan Church and Audrey Mackay investigate whether expanding eligibility for Child Care and Development Fund subsidies increases fertility. Although greater affordability could reduce the costs of raising children, their analysis does not establish a clear effect of eligibility thresholds on recent births. Patrick Angelo Sursano reaches a similarly cautious conclusion about the employment benefits of Opportunity Zone designation. He finds no statistically significant increase in the number of jobs, while an apparent increase in the share of resident workers is complicated by trends that preceded designation.

Relatedly, Annie Hsu investigates the effect of marijuana legalization on adolescent use of marijuana. She finds an insignificant relationship between marijuana legalization and adolescent use after accounting for persistent differences across states and that educational attainment is negatively associated with adolescent use in several specifications. These studies remind us that expanding eligibility or offering financial incentives does not necessarily lead to the intended behavioral responses or that marijuana legalization leads to increased use among adolescents.

Dong Nguyen's study of Medicaid expansion extends this discussion by showing that policy can change the composition of an outcome as well as its overall level. His results suggest that expansion reduces private insurance coverage primarily through the employer-sponsored channel and that policy evaluation should consider this trade-off, rather than just changes in the uninsured rate.

A second theme is the role of economic resources in shaping households' responses to risk and disruption. Daniel Wiltse finds that higher-income borrowers are more likely to select adjustable-rate mortgages, with a stronger association in the jumbo market. His analysis suggests that the same financial product can present different opportunities and risks depending on a household's resources. Similarly, Jerry Zheng identifies a related income gradient in Minnesota's electricity demand during severe winter cold. He finds that electricity demand responds more strongly in higher-income areas. Anna Fratto's examination of household entertainment expenditure offers another perspective on the importance of resources. Income is the most consistently significant predictor of spending in her analysis, while measures of COVID-19 severity do not yield statistically significant effects. Combined, the papers point to an important issue: consumption choices reflect both the need and the capacity to meet that need. Mortgage selection, heating demand, and discretionary spending reflect both preferences and the financial capacity to respond to changing circum-

stances.

This volume of the MJE is rounded out by two excellent studies in the fields of education and transportation economics. Emma Runchey Smalley finds a positive association between teachers' beliefs in their ability to influence student performance and student test scores in Ethiopia, with a stronger relationship for English than mathematics. Her study directs attention toward the classroom conditions that support learning, while acknowledging the difficulty of establishing causality.

Aram Petrosyan finds that citywide speed limit reductions are associated with improved performance in leisure and hospitality sectors, without statistically significant adverse effects in the other industries examined. His findings suggest that safer streets and local economic activity may be compatible, although concurrent safety measures and differences across cities warrant caution.

On behalf of my colleagues in the Economics Department, I am delighted to present the work of these talented students and to thank the editors for bringing it together. Their research exemplifies the contribution of a liberal arts education to economic inquiry: asking consequential questions, examining evidence critically, and taking causal inference seriously. I hope readers will find in this volume both new insights into its many subjects and an appreciation for the careful reasoning that links them.

Felix Friedt
Associate Professor of Economics

Fertility and Child Care Affordability: Evidence from CCDF Eligibility Threshold

Teagan Church and Audrey Mackay

ECON 431: Economics of Public Policy (Advisor: Sarah West)

Introduction

In a span of 42 years, from 1960 to 2002, the rate of children born, or the fertility rate, decreased by 45% in the US (7). Currently, the fertility rate in the US is well below the replacement level of 2.1 children per woman — as of 2023 it was only 1.6 children per woman (22). A lower fertility rate results in a lack of population replacement and an aging population. This aging population phenomenon is associated with less innovation, slower economic growth, fiscal planning problems, and reduced funding of social welfare programs (17). A rapidly rising older population without replacement poses significant challenges from structural unemployment, cultural changes, spatial mismatch, and increasing total healthcare risks and costs. Maestas et al. (14) find that "each 10 percent increase in the fraction of the population aged 60+ decreased per capita GDP by 5.5 percent. One-third of the reduction arose from slower employment growth; two-thirds due to slower labor productivity growth."

There are two main accepted theories for lower fertility: the quantity-quality tradeoff of investing more money into fewer children (2), and the opportunity cost of rising wages compared to the time cost of raising children (9). Even with changing relationships of gender roles and compatibility of career and having children, both theories suggest that cheap and available child care would reduce the time cost of having children by changing the dynamics of work and cost of children.

Childcare prices are extremely high. Families in the US are spending 8.9% to 16% of median income on childcare, and that is only for one child (21). These high prices could be a contributing factor to the decrease in fertility in the US — 43% of the people surveyed cited insufficient money as one of their barriers to having children (1). According to the American Family Survey in 2025, 71% of Americans believe that the cost of raising children is not affordable, which is 13 percentage points higher than in 2024 (1).

Public policy exists with the goal of combating these rising costs, including federal and state assistance, sliding fee scales, scholarships, and discounts. Federal assistance via subsidization of childcare is commonly used; the Child Care and Development Fund (CCDF) is the largest of these. The CCDF is focused on helping low-income working families access child care (3). This type of subsidy allows for a progressive burden shifting, making the difference in percentage of income from lower incomes to higher less stark by effectively decreasing the cost of childcare for lower income households, but like many subsidies, there

[0]Teagan Church is a Senior majoring in Economics and Political Science. Correspondence concerning this article should be addressed to tchurch@macalester.edu. Audrey Mackay is a Senior majoring in Economics. Correspondence concerning this article should be addressed to amackenz@macalester.edu.

are eligibility thresholds. A higher cap on eligibility would widen the net and allow for more families to access childcare per state. Federal eligibility tends to depend on the child (a) being younger than 13 years of age, (b) living with parents or guardians who are working, or needing protective services and residing with a family whose (c) total income does not exceed 85% of their State's median income for a family of the same size and (d) whose assets do not exceed $1,000,000 (20). The eligibility threshold therefore refers to the maximum dollar amount that different policies and states use to qualify families of different sizes, and whether the threshold changes for families who are first applying for assistance. In this paper we endeavor to explore the potential link between childcare subsidization and fertility. We ask what the effect of eligibility thresholds for childcare is on the fertility rate at the state level.

In order to investigate these effects, we use data from the US Census Survey on Income and Program Participation (SIPP) and the Child Care and Development Fund policy data. We compare them as cross-sectional panel data over the years 2014–2023 at the state level using two-way fixed effects models with event studies with different treatment identifications as robustness checks for the effect of the changes in eligibility thresholds in different states and years. These estimates are based on the following model in Section II of optimal decisions with respect to having an additional child based on maximization of utility and income. In Section III, we explore the existing literature and our place within it as we are exploring an identification to strengthen evidence for the causal link between fertility and childcare availability. Section IV discusses our data from the SIPP and CCDF database and their descriptive characteristics and analysis. Sections V and VI present and evaluate our empirical results that there is limited effect on recent births from the CCDF eligibility threshold changes and compare this to our predictions that higher eligibility maximum thresholds will increase fertility, based on our theory.

Theory

In order to better understand the relationship between child care prices and fertility, we consider a theoretical model based on Doepke et al. (8). We assume that fertility decisions are a joint decision for couples and they seek to maximize their combined utility, which they derive from consumption, c, and the number of children. Following Doepke et al. (8) we model utility using a logarithmic function for concreteness, where δ is the utility derived from children. For simplicity's sake we assume that couples are made up of two partners, m_1 and m_2, who earn after-tax wages w_{m_1} and w_{m_2} respectively, where $w_{m_1} > w_{m_2}$. The time endowment is normalized to one, which also means that the potential income for each individual is the same as their wage. Child rearing costs are split into two types: fixed time costs ϕ, and monetary costs ψ (e.g. spending on clothing), where time costs greatly outweigh monetary costs. However, with the marketization of child care, parents have the choice to purchase some share of childcare, s, on the market with the associated price p_s, or to provide childcare on their own. Due to having the lower wage, optimal specialization suggests that m_2 ought to be the one to exchange working hours for childcare hours, resulting in the higher-earning partner using their entire time endowment to work in the labor market while the lower-earning partner has the time constraint: $l_{m_2} + n\phi \leq 1$. This leaves us with the following maximization problem:

$$\max_{c,\,n,\,s} \log(c) + \delta \log(n) \quad \text{s.t.} \quad c + \psi n + s\, p_s\, n\phi = w_{m_1} + w_{m_2}[1 - (1-s)n\phi] \tag{1}$$

where s exists such that $s \in [0, \bar{s}]$, and $\bar{s}$ is the maximum amount of childcare for each child that can be provided by marketized childcare agencies. If $s = 0$ then no marketization is possible and if $s = 1$ then full marketization is possible. This relates the cost of child care against the possible wage that the lower-earning partner would receive, where a greater p_s represents a greater opportunity cost for working.

Here, we can finally get to the crux of Doepke et al. (8)'s contribution for our model: the addition of public family policies to parents' fertility decisions. The public partial provision of childcare would reduce the cost of having/raising children which we posit could, in turn, increase fertility. We adapt Equation (1) to allow the government to provide a share g (a policy parameter) of total childcare. With this, our couple's optimization problem becomes:

$$\max_{c,\,n} \log(c) + \delta \log(n) \quad \text{s.t.} \quad c + \psi n = w_{m_1} + w_{m_2}[1 - (1-g)n\phi] \tag{2}$$

with the following equation as the derived optimal fertility choice:

$$n = \frac{\delta}{1+\delta} \cdot \frac{w_{m_1} + w_{m_2}}{\psi + (1-g)w_{m_2}\phi} \qquad \text{where} \qquad \frac{\partial n}{\partial s} = \frac{\delta}{1+\delta} \cdot \frac{(w_{m_1} + w_{m_2})\, w_{m_2}\phi}{\left[\psi + (1-g)w_{m_2}\phi\right]^2} > 0 \tag{3}$$

As discussed, our research centers around changes in eligibility thresholds for families to receive CCDF subsidies for childcare. When more families are eligible for subsidies, it reduces the cost of having/raising children for more families. This can be read as an effective wage increase as the government provision reduces the cost of working, changing partner m_2's gains from labor from $w_{m_2} - p_s$ to $w_{m_2} - p_s + g$. An increase in effective wage has two pathways to impact n — the income effect and the substitution effect. The income effect says that for normal goods, which we assume children are, an increase in perceived income will increase the demand for children. If the income effect dominates, then we would expect the number of children to increase. On the other hand, the substitution effect tells us a greater effective wage actually makes not working — in this case by taking care of children — relatively more expensive, which would have a negative effect on the demand for children. If the substitution effect dominates, then we would expect the number of children to decrease.

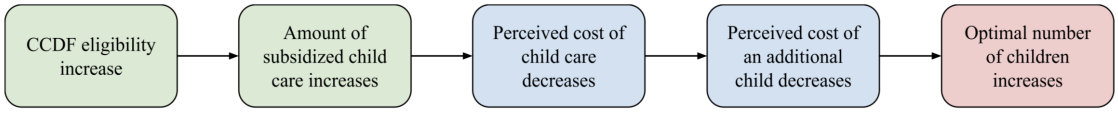
By using this model we hope to add to this broader understanding of the connection between family planning and financial support, increasing the maximum childcare available and effectively increasing parental wages w, increasing the expected financial stability of being able to work and also have children. In theory, this expected stability acts as the connection between financial support and the decision to have more children that we predict to see with increased subsidies.

Empirical Literature Review

The goal of this paper is to explore the potential link between childcare subsidization and fertility rates. Based on Section II and the Doepke model, empirical analysis should be

evaluative of the connection between eligibility thresholds and fertility. To do this, we use the basic causal flow to frame our research, shown below in Figure 1.

Figure 1: *Causal Framework*



The literature tends to cover a few main causal effects that we seek to unite via the framework outlined in Figure 1: the effect of income or wages on fertility and labor force participation, with the former often a sidenote. Many papers look at subsidies or incentives of formal child care on parental labor force participation. By comparison, it is much more rare to look at the effect of subsidies on fertility, especially subsidies of childcare. Our contribution is to expand on this limited subfield of literature by specifically using CCDF eligibility thresholds as our subsidization identification.

The first subsection of literature that we find helpful studies income on both labor force participation and fertility. The Blau and Robins (4) and Blau and Robins (5) papers are frequently cited by subsequent literature for their hazard function cross-sectional empirical analysis, using data to find effects for all four combinations of the outcome variables: employment status and additional children. Most useful for their study framework, based on leisure and consumption models with the driving motivation of the Becker and Lewis (2) quantity-quality model. The 1988 paper finds that nonwage income reduces labor force participation while the 1989 paper finds that each dollar increase for child care costs reduces the rate of entering employment by 3% and decreases the birth rate by 2%.

The next subsection of literature uses subsidies as the independent variable, but only looks at labor force participation as the outcome, rather than fertility. One example of this is Pepin (18), which focuses on formal child care participation after subsidies using CDCC benefits, and finds an unequal distribution of benefits to households, but that overall a 10% increase in CDCC benefits resulted in a 4% increase in paid childcare use. Herbst (12) and Herbst (11) give a comprehensive review of CCDF policies and their outcomes and note that CCDF subsidies are likely to lead to higher employment rates among disadvantaged mothers, but have unclear implications for any other measure of success. These papers emphasize the first few connections in Figure 1 where increased subsidies for child care leads to decreasing perceived cost of child care.

A third framework uses fertility as the main outcome of interest, but has various independent variables. Del Boca (6) focuses more on the opportunity cost explanation of the causal connection between childcare subsidies and family outcomes, finding that income has a positive relationship with fertility and a negative relationship with labor force participation in Italy. Although these papers look at income and availability rather than child care subsidization, they support the theory that if the perceived cost of child care decreases, the optimal number of children may increase.

The final subsection is the most limited, but the closest to our research question; both Mörk et al. (16) and Haan and Wrolich (10) look at decreases in child care prices all

using fertility as the outcome of interest.

Mörk et al. (16) find that a reduction of child care prices increased first births by 9.8%. The basis of their theory is more of an opportunity cost model with a staggered rollout quasi-experimental method. Notably, there was no overall significant increase to fertility, only specifically with couples who previously had no children. One concern here is that they limited their data to only married couples, which weakens the analysis of first births as over 55% of children in Sweden are born out of wedlock (13), but the similar results of fertility increases in other papers increases confidence in the validity of these results.

Haan and Wrolich (10) have a smaller sample size and less ideal experimental design than Mörk et al. (16), but likewise find a negative relationship between cost of child care and fertility. Haan and Wrolich (10) find that increasing child care subsidies conditional on employment increases the labor supply of all women as well as the fertility of the formerly childless and highly educated women.

By looking at the CCDF funding, this paper attempts to narrow down the impact of broadening eligibility for subsidized childcare on fertility choices, in order to give context within both quality-quantity and opportunity cost theoretical models. By focusing only on fertility, we seek to add clarity to a subtopic that often gets added as an afterthought.

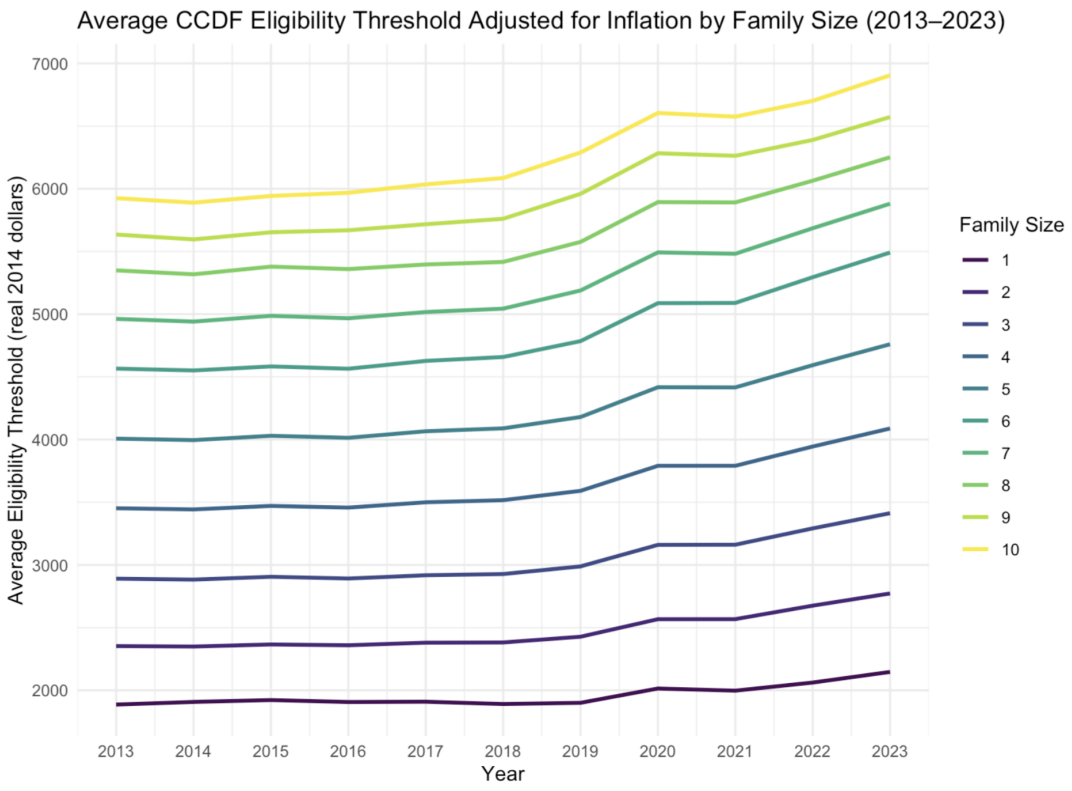
Data Description & Summary Statistics

Our state-level policy data on CCDF eligibility thresholds comes from the CCDF Policies Database and covers 2013 to 2023. The maximum income eligible for CCDF benefits varies over date, state, and number of household members. Below is a plot of average CCDF thresholds by family size from 2014 to 2023, all of which have been generally increasing over time, even when adjusted for inflation. As CCDF sets its goal threshold as a percentage of the poverty threshold, a real increase is important — we use the inflation-adjusted threshold in all future references.

We chose to look at household level data covering 2014 to 2023 (due to lagged effects) from the Survey of Income and Program Participation (SIPP), which follows tens of thousands of households each panel in a series of waves, resulting in monthly level household data. This massive sample is eventually filtered down to female fertility decision makers, which we approximate to be women between the ages of 21 and 45 that are married or cohabitating with a partner. Unfortunately SIPP data has limited fertility variables, so instead we consider fertility using a created, binary `recent_birth` variable that equals 1 if any member of the household has an age at last birthday of 0, which indicates that an infant was born to the household in the last year. In order to prevent counting a single child numerous times in the case that there are numerous women in one household that fulfill these characteristics, we choose to filter out households that have more than one fertility decision maker, as this group of households only makes up approximately 3% of households in our sample and we prioritize controlling for individual-level demographics. In the final filtered sample we have 323,435 observations.

While SIPP does have a family size variable, it is not comparable to the family size that CCDF uses to determine eligibility — due to this we create our own. We create a family size variable that counts the individual, adds 1 if the individual has a spouse or a partner they cohabitate with, and adds 1 for every reference number that exists for their child(ren); this is then applied to all members of the family and allows us to match individuals to the

Figure 2: *Average CCDF Eligibility Threshold by Family Size 2014–2023*



CCDF threshold they would be subject to (we refer to these as one's own threshold). While this does assume that if a woman has a spouse then said spouse lives in the same household, which isn't necessarily always the case, we are comfortable making this assumption in order to be able to include non-married cohabitating partners.

According to the US Census Bureau, 74.8% of Americans are White alone, 13.7% are Black alone, and 6.7% are Asian alone (19). Looking at Table 1 it is clear that our proportions are rather similar in early years with more deviation in later years. The rates of high school and college graduation are on par with the national sample (89.7% and 35.7% respectively across America as a whole). Our sample does have much greater average incomes across the entire time frame, especially in later years, which could be a concern.

As fertility is not instantaneous and we are looking at births, not pregnancies, we lag CCDF thresholds by one year in order to capture the threshold an individual was more likely subject to when the decisions are made. We do this by creating a leaded month variable in the CCDF dataset that is equivalent to the month that we would expect to see fertility decision results and joining the two dataframes by this new month variable in order to make full use of available data.

Table 1: *Summary Statistics*

Variable	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
Average Household Yearly Income ($)	84699.54	86648.12	87206.23	95659.81	107070.82	108186.59	125458.56	123559.76	142059.83	136928.71
Average Age	34.52	34.88	35.20	35.49	34.46	34.47	34.94	34.99	34.90	34.92
Average # Children	1.81	1.85	1.94	2.01	1.71	1.74	1.66	1.59	1.67	1.65
Average Family Size	3.80	3.87	3.93	3.94	3.72	3.50	3.66	3.55	3.64	3.51
Recent birth (%)	12.79	15.76	18.90	20.52	10.53	4.70	10.53	9.01	9.72	7.62
High School Graduate (%)	89.89	90.01	89.82	90.18	90.99	90.07	92.83	92.79	92.59	92.29
College Graduate (%)	37.08	38.57	39.82	41.35	42.75	41.40	46.18	47.97	47.15	46.51
White (%)	81.71	81.96	82.10	80.95	79.96	82.27	80.59	78.08	78.12	77.78
Black (%)	9.00	8.53	8.33	8.59	8.14	5.94	7.70	7.06	7.16	6.54
Asian (%)	6.25	6.52	6.65	7.16	7.89	7.80	7.83	9.83	10.60	10.75
Other Race (%)	3.04	2.99	2.92	3.30	4.01	3.99	3.88	5.03	4.12	4.93

Summary Statistics by Year.

Empirical Results – TWFE

The ideal experiment would entail a treatment group which is given CCDF funding to a household and the control does not receive CCDF funding, studied at the individual level, holding income variability fixed, where the control and treatment are otherwise exact copies, and not having made any decisions about future fertility and have only been influenced by the conditions of the funding. Unfortunately, this is not possible, which leads us to our actual estimation strategy.

As previously mentioned, since we do not have any observations for fertility itself, we choose to approximate the relationship using recent births as proxies. Two-way clustering standard errors at the state and month levels, in our main specification we estimate the equation

$$Y_{ist} = \alpha + \beta \cdot \text{Threshold}_{st} + X'_{ist}\,\gamma + \delta_s + \theta_t + \varepsilon_{ist} \tag{4}$$

where Y is the probability of a recent birth for individual i in state s at time t. Threshold is the CCDF threshold of interest (either one's own or the threshold of a two-person family depending on the specification) while X is a vector controlling for age, education level, household income, and race. We also include state and time fixed effects, represented by δ_s and θ_t respectively. In this equation β is our coefficient of interest and represents the change in probability of a recent birth from the CCDF thresholds increasing by $1 (real).

Table 2 displays results from our lagged two-way fixed effects model using the household data from SIPP against the CCDF thresholds. One variable of interest, one's own CCDF threshold, is significant at the 1% level, meaning we could reject the null hypothesis of one's CCDF thresholds having no effect on recent births. Despite the significant results, the AIC value is very large in all models and the R-squared is less than 0.1, suggesting that CCDF thresholds do not explain much variation in recent births. When controls are added to this linear probability model the effect increases, along with the significance — with a $1 increase in CCDF eligibility threshold increasing the probability of recent births by 0.003 percentage points. While the results may seem minor, the estimated coefficient of interest implies that a $1,000 increase in the CCDF eligibility threshold is associated with a 3 percentage point increase in the probability of a recent birth. A $1 increase in CCDF thresholds would suggest 750 additional recent births — which is likely to be an overestimation. These results are most likely driven by reverse causality, as each additional birth would increase the CCDF eligibility threshold one is subject to by increasing the family size.

We therefore also chose to look at the effect of increases in the CCDF eligibility threshold of a two-person household on recent births as well, which was positive but not significant. We suggest that this lack of significance is more likely accurate than the large significant results found in the first two columns of Table 2.

Table 2: *Linear Probability Model Results*

	(1)	(2)	(3)
VARIABLES	TWFE	TWFE With Controls	2 Person Threshold with Controls
Own CCDF Threshold (t-12)	0.000021***	0.000030***	
	(0.0000004)	(0.000006)	
2 Person CCDF Threshold (t-12)			0.000004
			(0.000003)
Black alone		−0.006638*	−0.004668
		(0.002969)	(0.002938)
Asian alone		0.000175	−0.001883
		(0.005087)	(0.004548)
Other		0.000874	0.004378
		(0.005664)	(0.005898)
Age		−0.005379***	−0.004717***
		(0.000992)	(0.000880)
Graduated college		0.021109***	0.013345***
		(0.004668)	(0.003577)
Monthly household income		−0.000000	−0.000000
		(0.000000)	(0.000000)
Num. Obs.	323,435	323,435	323,435
R^2	0.059	0.083	0.071
AIC	−82069.4	−90340.9	−86358.9
BIC	−80637.4	−88844.8	−84862.8
RMSE	0.21	0.21	0.21
FE: state	X	X	X
FE: month	X	X	X

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$, **** $p < 0.001$.

The true accuracy of these estimates is nevertheless in question; results may reflect the direction and the general magnitude of fertility decisions, but recent births are not a true measure of fertility. While we attempted to control for heteroskedasticity by clustering at the state and month levels, given the implausible magnitude of our regression results we believe it is likely that our standard errors are still highly biased. Given our limited variables related to other fertility determinants, these results could also be affected by omitted variable bias which could be assigning the effect of an omitted variable to CCDF thresholds, falsely inflating our estimate. Linear probability regression models also inherently assume that the probability relationship is linear, and, while common in econometrics, can result in biased estimates, which is another likely cause of our unrealistic results. To further investigate our results we also tested with a logistic regression model, however the estimated effect was even greater and therefore less likely, failing to disprove our hypothesis that these results are driven by endogeneity.[1]

[1]Results for logarithmic regression available upon request.

Empirical Results – Secondary Estimations

One important aspect of our research is the way in which eligibility thresholds increase: within some states, there are multiple increases sometimes affecting only some family sizes, with different treatment times. With a stacked event study, we can aggregate data to the state level, stacking policy changes and looking at months before and after to find out if there is any clustering of probability of having a child. This also aids in eliminating some of the endogeneity bias, both by grouping at the state rather than household level and by limiting the estimation to a more limited before-and-after treatment period. Although some states have multiple eligibility threshold changes within a short time period, some event studies can still be applied, especially by limiting the window of study to states which have only a change affecting all family sizes, or by looking at the change for only one family size so as not to pick up the reverse causality of eligibility thresholds changing as family size increases.

The first event study estimation aggregates at the state level and fixes the years so that the policy change is the time of each eligibility threshold increase, in the following equation:

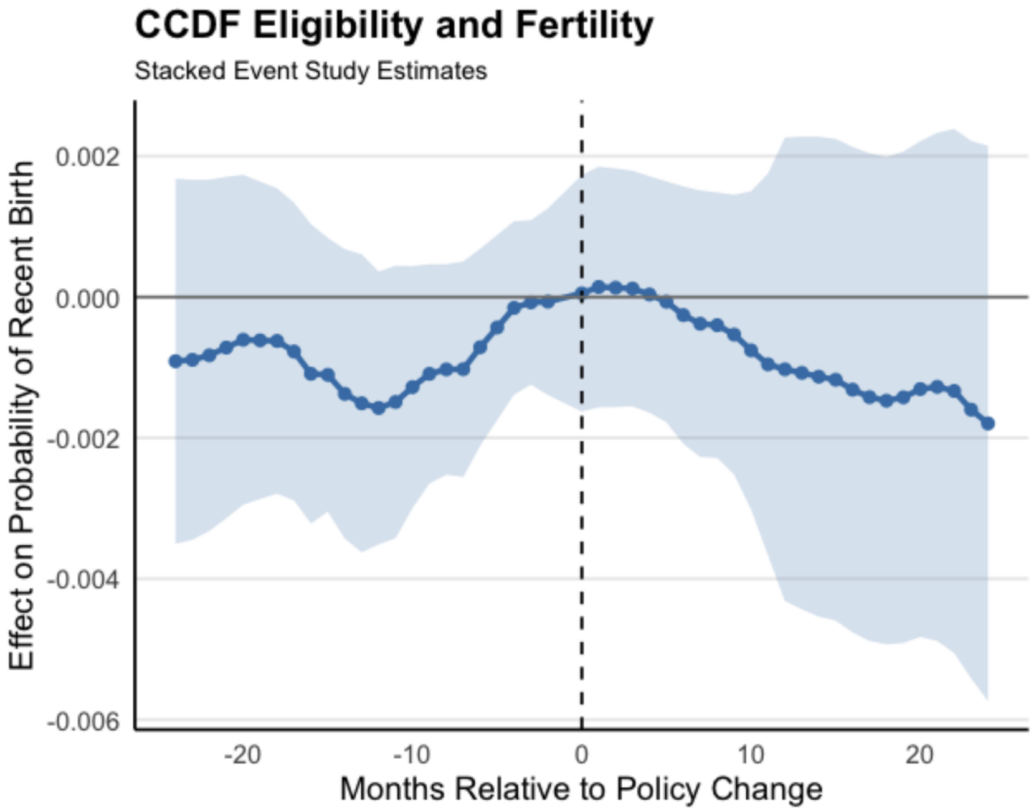
$$Y_{ist} = \alpha_i s + \lambda_t + \sum_{k \neq 1} \beta_k D_{ist}^k + \gamma X_{it}' + \mu_s + \lambda_t + \varepsilon_{ist} \tag{5}$$

in which k is years relative to treatment and D_{it}^k indicates each event time for β. The results, as seen in Figure 3, indicate a positive effect on the probability of a recent birth for the first 5 months after eligibility thresholds changed, after which point the effect becomes negative and decreases to pre-trend levels. Post-trend coefficients are small in magnitude and almost completely insignificant. Positive coefficients range from 0.0003 to 0.00045 and negative coefficients range from -0.0009 to -0.0001. Each month before and after the treatment, the standard errors are larger than the actual estimate of the effect, and even controls are statistically indistinguishable from zero. The lack of statistical significance comes in part from the definition of the treatment — due to the simultaneous increasing eligibility in different states and increasing eligibility at different family sizes, it is difficult for this stacked event study to get a clear before and after treatment, because there are many options for what constitutes a treatment, given the data. This is especially evident in the lack of clear pre-trends and the low overall fit of the estimate, the R-squared value indicating only 29% of the variation can be explained by the treatment timing. Overall, the lack of statistical significance and the potential underestimation due to the potential for overlapping treatment periods, even when stacked, indicate little evidence of the CCDF increased eligibility having an effect on recent births.

The next estimation is a Sun-Abraham estimator (15), which is another estimation that treats bias with staggered rollout. In this, we attempted to eliminate some of the endogeneity of increasing family sizes by only looking at 12 cases where all of the potential family sizes in the state increased in eligibility at the same time, as specified below:

$$Y_{ist} = \alpha_s + \sum g \sum_{g \neq 1} \delta_{gl} \cdot \mathbf{1}[G_s = g] \cdot D_{ist}^l + X_{ist}' \gamma + \mu_s + \lambda_t + \varepsilon_{ist} \tag{6}$$

Figure 3: *Stacked Event Study*

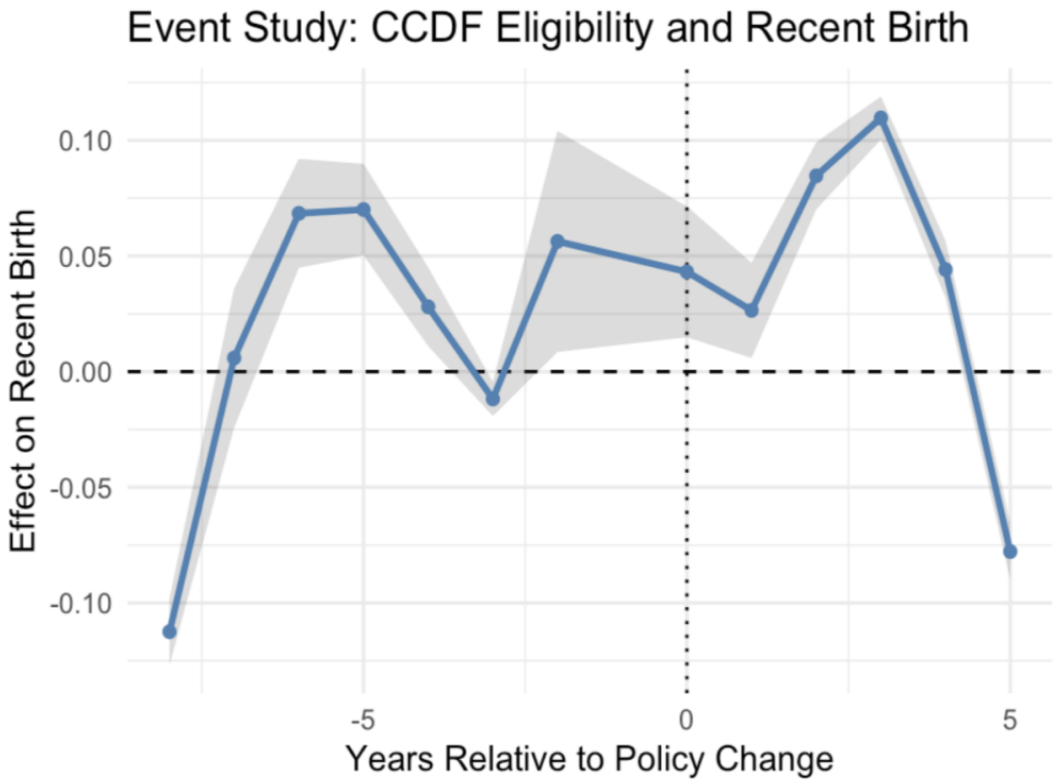


where D^l_{ist} represents the Sun-Abraham weighted interaction and G is the time-treated cohort, grouped by the first change in eligibility threshold, with $\delta_{g,l}$ as the main coefficient of interest.

The result of this is a much more statistically significant, but reduced sample. As before, this estimate fixed national fertility trends, while comparing within states, but this time, the potential for overestimating because of reverse causality is much reduced. Once those cases were selected, we actually see a very small decrease in probability of a recent birth. Given the other results and theory, it seems likely that the change in sign might be attributed to the limited sample and the very small R-squared value of 0.015, indicating that little of the estimation is actually captured by the time treatment variation. Importantly, this takes a very different shape from the above results, and from the more standard TWFE.

Finally, we estimated a third continuous event study, a DDiD that more accurately groups the original household data by state and month and uses the start, end, and survey dates to create a difference-in-differences estimate in which multiple treatments per state are allowed for; although there is likely some overlap, the treatment period is more accurately represented throughout the states and it covers a longer before-and-after period than the first event study which only covered 2 years, while this one covers all the years in the

Figure 4: *Sun-Abraham Study*



dataset.[2] The estimating equation for this is

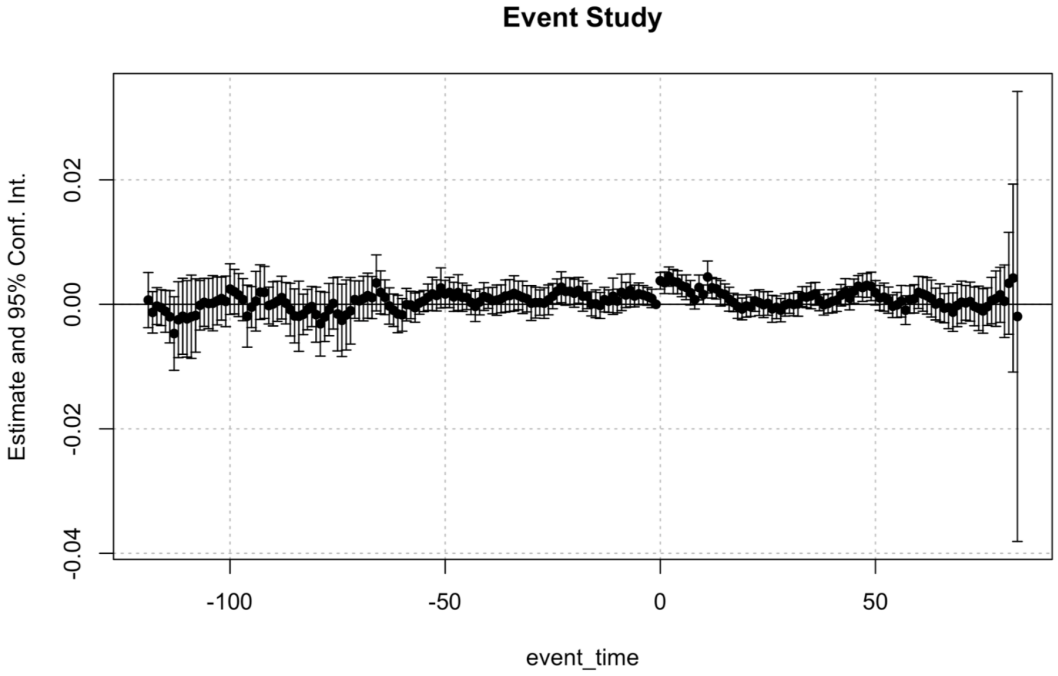
$$Y_{ist} = \alpha + \sum_{k \neq 1} \delta_k (\text{Thresholds} \cdot \mathbf{1}[t - T_s = k]) + X'_{ist}\gamma + \mu_s + \lambda_t + \varepsilon_{ist} \tag{7}$$

where Threshold_s indicates the state-month dummy variable with k, created by the dates of treatment and the date of the survey, and X stands for controls as with the TWFE. There is far less variation in these final estimates, but at the treatment time there is a clear break at $t = 0$ before it returns nearer to zero, as shown in Figure 5. Otherwise, the pre and post trends are very similar and have small changes and no statistical significance.

Similar to the above estimate, pre-trends present a challenge to the estimation, as they are large relative to the rest of the model and not constant, signifying the likely possibility of omitted variable bias and pre-trends causing more of an effect than the independent variable. In all, although two-way fixed effects are often seen as less robust, the difference-in-differences parallel assumptions show the overlapping of the CCDF eligibility changes timing to such a degree that there may be too much covariance to come to a conclusion about CCDF eligibility effects using these event studies.

[2]Full results tables for all three event studies are available upon request.

Figure 5: *Dynamic Differences-in-Differences Estimate*



Conclusion

We are unable to conclude that CCDF eligibility thresholds have an effect on recent births, although several models show a positive effect with varying significance. These results are relatively in line with existing literature, which although generally indicates limited effects of subsidization of child care on fertility, is very limited while other inputs such as child care availability and child care cost show more of a positive effect. Although part of our contribution is adding context to a limited and contested topic, results also conflict with several studies as well. It makes sense that results are limited, given the unclear treatment period of the CCDF funding and the changing dynamics between wage and child bearing.

In the future we look towards continued study on the relationship between fertility and childcare subsidies. Particularly, further data on larger time periods would be extremely useful in determining long-term fertility trends and life-cycle choices. To further evaluate the CCDF and the subsidies' effectiveness, data that enables a treatment intensity rather than event study, as well as more information on uptake of the program and childcare expenditures post-subsidy is needed, and we look forward to future papers that investigate this. Moreover, while we limited our sample to individuals who cohabitate with a partner (or spouse), as social norms about fertility change and having a partner becomes less necessary to having children, we would like to see further research on single individuals' decisions to have children.

References

[1] American Family Survey (2025). American family survey 2025: American families in an era of rapid change. Wheatley Institute.

[2] Becker, G. S. and Lewis, H. G. (1973). On the interaction between the quantity and quality of children. *Journal of Political Economy*, 81:279–288.

[3] Bipartisan Policy Center (2019). History of federal funding for child care and early learning.

[4] Blau, D. M. and Robins, P. K. (1988). Child-care costs and family labor supply. *The Review of Economics and Statistics*, 70(3):374–381.

[5] Blau, D. M. and Robins, P. K. (1989). Fertility, employment, and child-care costs. *Demography*, 26:287–299.

[6] Del Boca, D. (2002). The effect of child care and part time opportunities on participation and fertility decisions in italy. *Journal of Population Economics*, 15:549–573.

[7] Dilmaghani, D. et al. (2024). Decreasing fertility rate in the united states: Demographics, challenges, and consequences. *Mayo Clinic Proceedings*, 99(11):1693–1697.

[8] Doepke, M., Hannusch, A., Kindermann, F., and Tertilt, M. (2022a). The economics of fertility: A new era. Working Paper 29948, National Bureau of Economic Research.

[9] Doepke, M., Hannusch, A., Kindermann, F., and Tertilt, M. (2022b). The new economics of fertility. International Monetary Fund F&D Magazine.

[10] Haan, P. and Wrolich, K. (2009). Can child care policy encourage employment and fertility? Evidence from a structural model. IZA Discussion Paper 4503, Institute for the Study of Labor.

[11] Herbst, C. (2008). Do social policy reforms have different impacts on employment and welfare use as economic conditions change? *Journal of Policy Analysis and Management*, 27:867–894.

[12] Herbst, C. (2023). Child care in the united states: Markets, policy, and evidence. *Journal of Policy Analysis and Management*, 42(1):255–304.

[13] Jacobs, F. (2023). In these 11 european countries, births out of wedlock are in the majority. Big Think.

[14] Maestas, N., Mullen, K., and Powell, D. (2023). The effect of population aging on economic growth, the labor force, and productivity. *American Economic Journal: Macroeconomics*, 15(2):306–332.

[15] Metricgate (2026). Sun-Abraham estimator.

[16] Mörk, E., Sjögren, A., and Svaleryd, H. (2011). Childcare costs and the demand for children – Evidence from a nationwide reform. *Journal of Population Economics*, 26:33–65.

[17] Munnell, A. H. (2025). Low fertility is a really big problem – with an easy answer. Center for Retirement Research.

[18] Pepin, G. (2020). The effects of child care subsidies on paid child care participation and labor market outcomes: Evidence from the child and dependent care credit. Upjohn Institute Working Paper 20-331, W.E. Upjohn Institute for Employment Research.

[19] U.S. Census Bureau (2026). QuickFacts: United states. Accessed April 29, 2026.

[20] U.S. Department of Health and Human Services (2025). Understanding federal eligibility requirements. Office of Child Care: Child Care Technical Assistance Network.

[21] U.S. Department of Labor (2025). National database of childcare prices. Women's Bureau.

[22] World Bank (2025). Fertility rate, total (births per woman) — United States. UN Population Division.

Teacher Efficacy & Student Success in a Developing Country Context

Emma Runchey Smalley

Department of Economics, Macalester College
ECON 426: International Economic Development Capstone (Advisor: Amy Damon)

Introduction

"Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all" (United Nations Department of Economic and Social Affairs). The United Nations' Sustainable Development Goal 4 is a global call to action. Education has long been a focus of economists, policymakers, and other stakeholders. How teacher effects such as beliefs, pedagogies, classroom practices, and other characteristics impact students has been a focus of researchers because of their centrality to student learning. One measure of teacher effects is called teacher efficacy. Teacher efficacy is defined as the "extent to which the teacher believes he or she has the capacity to affect student performance" (2, p. 137). Though some research has been conducted in developed countries since at least the 1970s (22), there is still a distinct lack of research regarding teacher efficacy in developing countries. This article attempts to empirically measure the impact of teacher efficacy on student achievement outcomes in a developing country context.

Ethiopia has made significant progress in educational enrollment. Despite more recent issues including displacement and the COVID-19 pandemic, 88% of students remained enrolled in 2021 (UNICEF Ethiopia). As a result, focus is increasingly shifting to school progression and achievement. In Ethiopia, only 33% of students make it to secondary school, and the vast majority of students are unable to complete tasks at grade level (UNICEF Ethiopia). This means the need for trained, qualified, and motivated teachers is more important. This study is of particular importance given the dearth of empirical research in developing countries like Ethiopia.

The term teacher efficacy and its measurement emerged in 1976. Based in a theoretical framework of expectancies based on perceptions of internal versus external control by Rotter (19), the RAND Corporation conceived the idea of teacher efficacy through the addition of two questions to a survey sent out to teachers in the United States (24). Since then, several conceptual theories and measures have developed (1; 9).

Literature Review

Since the late 1960s and early 1970s (22), research on teacher effects has gained emphasis in education economics research. A 1968 summary report from the then-US Department of Health, Education, and Welfare kicked off debates around teacher influence

[2]Emma Runchey Smalley is a Senior majoring in International Studies and Economics. Correspondence concerning this article should be addressed erunchey@macalester.edu.

on student success (7). Since then, there has been research on the various teacher characteristics and beliefs that influence student achievement and outcomes. In an early study, Hanushek (11) looked at the quality of schools via the characteristics of its teachers. He aggregated findings from 147 studies conducted across the United States, finding mixed results in terms of correlation between teacher and classroom characteristics, such as teacher-pupil ratio, teacher experience, and teacher salary, and student performance. The majority of studies had statistically insignificant results. Hanushek suggests this could be for a variety of reasons, including missing or low-quality data, administrators' actions in schools, and other factors.

Sanders et al. (20) analyzed the test scores of third, fourth, and fifth grade students in Tennessee in relation to several teacher and classroom effects. They found that the most important factor is the teacher, and that there is wide variation in the effectiveness of a teacher. Additionally, they found that teacher effectiveness holds with all achievement levels and that classroom variable effects are minor. Another study by Rockoff (18) looked specifically at teacher quality using panel data from schools in New Jersey. Teacher quality is notoriously difficult to quantify, though teacher experience was used as a primary factor. Rockoff found that a one standard deviation increase in teacher quality leads to a 0.1 standard deviation increase in both reading and math test scores.

A 2017 study by Blazar and Kraft (3) moved beyond simply student achievement to look at student behaviors. Using data, classroom observation, and a survey conducted in fourth and fifth grade classrooms on the East Coast of the United States, they found that high-quality, effective teachers raise test scores and improve student behaviors. A similar study was conducted by Cents-Boonstra et al. (6) in the Netherlands. Through classroom observation and a survey, they found that teachers who displayed more "motivating teaching behaviors" (e.g. asking for participation, providing positive feedback) in their lessons had more highly-engaged students.

Research into teacher effects in developing countries has also occurred, though not to the extent it has in developed countries. Research often focuses on teacher training (13) or the ambiguous "teacher quality" (21) or teacher knowledge (4). One study that uses all of these is Cueto et al. (8), who looked at pedagogical content knowledge (PCK) — "knowledge of content and students" (8, p. 333) — using the Young Lives Peru study. They find that PCK is positively associated with student achievement, but only at a certain threshold.

While early research looked at a broad swath of teacher effects, teacher efficacy is the aspect of teacher influence of interest in this paper. There have been two primary strands of research on teacher efficacy. The first is concerned with how teacher efficacy is developed, relying on the assumption that teacher efficacy has an effect on students. The second looks at teacher efficacy's impact on student achievement (24). Findings have varied, though generally they find some positive correlation. In one of the earliest studies on teacher efficacy, Moore and Esselman (15) looked at teacher self-efficacy and the factors that contribute to the development of this feeling in schools in Kansas City, Missouri. Self-efficacy is defined as the belief that an individual has the ability to impact student outcomes. They found that "personal and teaching efficacy were highly, although inversely, related" (15, Abstract). They also found that efficacy was related to the school atmosphere and to teachers' perceived ability to contribute to school decision-making.

The second strand produces studies such as Muijs and Reynolds (16) and Kim and Seo (14). Muijs and Reynolds (16) found that teacher beliefs in the United Kingdom influence student behavior indirectly, and that classroom factors were the highest predictors of student achievement, measured by test scores. They also found that teacher self-efficacy is linked to better test results. Kim and Seo (14) used Gibson and Dembo's (1984) scale to measure teacher efficacy in a meta-analysis. In an analysis of 16 studies, they found that teacher efficacy had a small, but significant effect on student achievement. However, the findings across studies were generally inconsistent.

A third aspect of research on teacher efficacy comes in the form of collective teacher efficacy. Collective teacher efficacy is "the perceptions of teachers in a school that the efforts of the faculty as a whole will have a positive effect on students" (10, p. 480). After creating a collective teacher efficacy scale, they ran a pilot study of 70 teachers across five states. They found that individual teacher efficacy had a positive, significant coefficient. They also found that collective teacher efficacy had a positive but insignificant coefficient. In a 2004 study by Tschannen-Moran and Barr (23) conducted in Virginia, they found a significant positive relationship between collective teacher efficacy and student test scores in math, writing, and English. However, when socioeconomic status was included, only the writing score had a significant relationship with collective teacher efficacy. Since both efficacy and test scores were aggregated at the school level, any observation of individual effects was eliminated.

One of the few examples of literature specifically on teacher efficacy conducted in a developing country context is Paisun et al. (17). The study engages in a qualitative descriptive analysis of 38 teachers and administrators in a district of Java, Indonesia. They found a positive correlation between teacher self-efficacy and student achievement. However, this being one of the only studies on teacher efficacy in a developing country context goes to show the lack of empirical evidence in particular.

Two major gaps in the literature emerge. First, even empirical research tends to be aggregated or very limited in scope, such as to one state or even one school district. Second, nearly all the research was conducted in developed countries. However, given the influence that teachers are thought to and have been shown to have on student achievement, they should not be neglected in a developing country context. Additionally, the lack of strong empirical study at the individual level leaves room for research. This paper attempts to address both of these gaps.

Economic Theory

Education economics as a field has grown substantially since the early 1990s (5). Education is considered to be a public good and has been shown to impact economic outcomes (12). There are two primary economic theories at play. The first is Human Capital Theory, in which education is viewed as an individual investment decision. Brewer et al. (5, p. 194) note that "knowledge and skills acquired through educational investments increase human productivity." Though one may incur expenses in pursuit of education, this pays off in the long run, particularly in the form of income. One way this is seen is the correlation between education and earnings in developed countries.

A second theory is Education Production Theory. Since education is known to lead to better economic outcomes, it raises the question of how education is produced. There

are two forms. The first is "to treat education as a production function", and the second views education as a "web of interpersonal contracts" (5, p. 196). The view of education as a production function in which inputs lead to outputs provides the grounding for thinking about teacher efficacy and how it impacts student test scores. In keeping with theory and the findings of the literature, if teacher efficacy is higher, test scores should also be higher.

Empirical Strategy

While there are many conceptual theorizations of teacher efficacy, the one used here is most closely related to the original RAND survey. In the RAND survey, two items were summed to get a teacher efficacy score: "When it comes right down to it, a teacher really cannot do much because most of a student's motivation and performance depends on his or her home environment" and "If I really try hard, I can get through to even the most difficult or unmotivated students" (24, p. 204).

This paper will use an Ordinary Least Squares (OLS) regression analysis. Student test scores in math or English are the variables of interest, denoted by y_i. These test scores are taken from the second wave of data collection at the end of the school year. The independent variable is teacher efficacy. Teacher efficacy is measured by taking the average of a teacher's responses to the statement "Teachers are not a very powerful influence on student achievement when all factors are considered" and the statement "If I really try hard, I can get through to even the most difficult or unmotivated students" (29). There will also be many teacher and student control variables from both teacher and student surveys added, which come from both the beginning and end of year surveys. Lastly, fixed effects for locality (urban or rural), region, and individual school effects will be included. Neither the controls nor fixed effects are noted in the regression below due to the number included.

$$y_i = \beta_0 + \beta_1 \text{ Teacher Efficacy}_i + \varepsilon_{it} \tag{1}$$

Data Description

The Young Lives Study follows the lives of 12,000 children from four countries: Ethiopia, India, Peru, and Vietnam (28). The study follows two age groups: one group born in 1994–1995 and the other born in 2001–2002. The data were collected in seven rounds over twenty years, and is considered to be one of the most comprehensive studies on poverty. For the purposes of this study, the 2016–2017 Young Lives School Survey, Ethiopia is used (29). This survey examines 12,182 students in Grades 7 and 8 at multiple rural and urban sites in seven regions of the country: Addis Ababa, Amhara, Oromia, the SNNPR, Sidama, Tigray, Somali, and Afar. The data were collected in two waves, at the beginning and end of the school year.

Most of the classrooms surveyed include at least one child from the Young Lives Survey cohort, though the geographical scope did increase for the School Survey, so not all classrooms include one. The variables of interest are test scores and teacher efficacy. This paper looks at the impact of teacher responses to the statements "Teachers are not a very powerful influence on student achievement when all factors are considered" and "If I really try hard, I can get through to even the most difficult or unmotivated students" on student survey scores in English and mathematics classes. Teachers ranked their response from

Strongly Disagree (1) to Strongly Agree (4). The scale was flipped for the first statement for operationalization and accuracy, making the new statement "*Teachers are a very powerful influence on student achievement when all factors are considered.*"[3] The responses to the two statements were then averaged to create a teacher efficacy measure. Access to the datasets is provided by the UK Data Service (25).

Table 1: *Summary Statistics*

Statistic	N	Mean	St. Dev.	Min	Max
Math Teacher Efficacy	11,408	2.853	0.600	1.000	4.000
English Teacher Efficacy	11,471	2.790	0.504	1.500	4.000
Math Score	9,733	19.226	7.229	1	40
English Score	10,180	20.187	7.115	2	40

Table 1 shows summary statistics for teacher efficacy and test scores. Both math and English teacher efficacy scores average around 2.8, on a 4-point scale. However, the lowest math teacher efficacy is 1 whereas the lowest English teacher efficacy is 1.5, giving the two slightly different medians. Math and English test scores also average near each other, around 19 or 20 points out of 40 on average. The spread for test scores is very large.

Summary statistics were also run by region and by years of teacher experience. Tables 7 and 8 in the Appendix show summary statistics by region. Teacher efficacy scores for both math and English are largely similar across regions, with more variation in the math teacher efficacy scores than the English ones. Test scores have a bit more variation, though no region is such a large outlier as to be concerning. Tables 9 and 10 in the Appendix show summary statistics by the teacher's years of teaching experience. Teacher efficacy does appear to increase with more years of experience for both math and English, though not significantly so. Test scores largely remain around the same mean across years.

Table 6 in the Appendix shows summary statistics for the same variables standardized as z-scores. Figures 3–**??** in the Appendix show the kernel density estimates for each of the four variables of interest. There are slightly skewed distributions for each, with test scores sitting below zero and teacher efficacy scores mostly splitting around zero. This is expected given that performance at grade level is a major issue in Ethiopia, which likely explains why most of the scores are below average.

Figures 1 and 2 represent the relationship between English and math teacher efficacy and English and math test scores respectively. Because teacher efficacy is an average of two categorical variable responses, the test score results are siloed. However, with the box plot overlaid, one can get a better idea of the averages within the responses. There is a slight upward correlation between teacher efficacy for both English and math teachers with the average test scores of their students.

Results

Table 2 shows the regression models for English teacher efficacy. The first regression is the basic OLS regression. The second regression includes control variables. The third regression includes fixed effects for locality (urban or rural), region, and individual school.

[3] Italics added.

Figure 1: *English Teacher Efficacy & Test Scores*

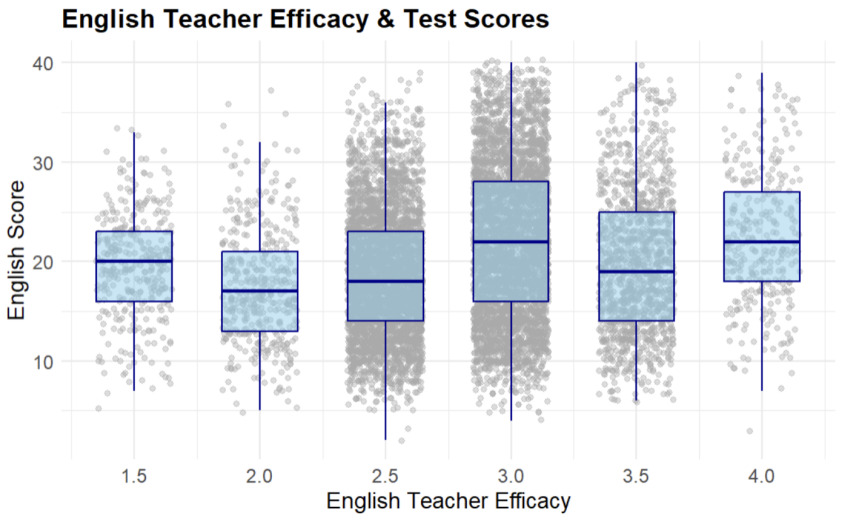
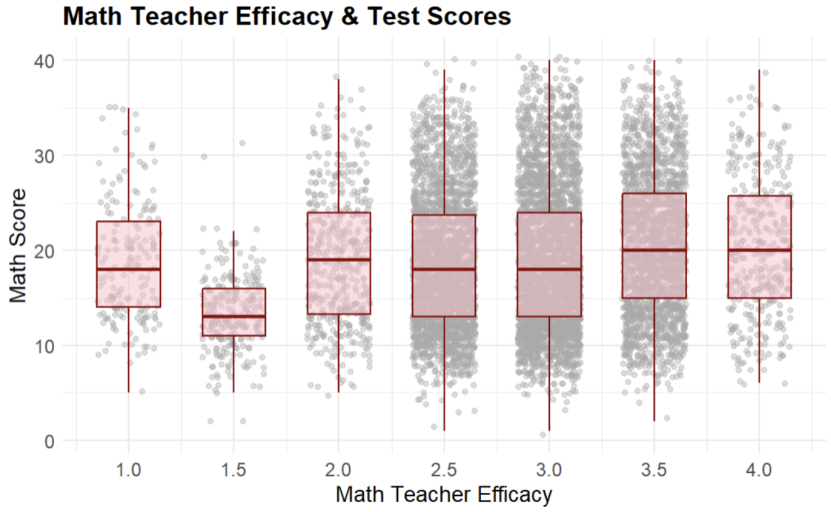


Figure 2: *Math Teacher Efficacy & Test Scores*



The final regression includes both controls and fixed effects. All of the regressions showed an increase in student test scores for a one-point increase in teacher efficacy. The basic regression showed an increase of 1.939 points per one-point increase in teacher efficacy. The regression including controls shows a 1.796 point increase in test scores per one-point increase in teacher efficacy. When using fixed effects, the coefficient drops to a 0.606 point increase, but when including both controls and fixed effects, there is a 1.199 point increase in test scores for a one-point increase in teacher efficacy. All of these results are significant at the 1% level of significance.

Table 3 shows regression models for math teacher efficacy. The basic regression and regression with controls both show an increase in student test score at 1.563 points and

Table 2: *English Teacher Efficacy Regression Results*

	(1)	(2)	(3)	(4)
VARIABLES	Basic	With Controls	FE	Controls + FE
Dependent variable: English Test Score				
English Teacher Efficacy	1.939***	1.796***	0.606***	1.199***
	(0.147)	(0.193)	(0.212)	(0.342)
Constant	14.801***	0.461	16.073***	9.027***
	(0.419)	(1.233)	(1.114)	(2.522)
Controls	No	Yes	No	Yes
Fixed Effects	No	No	Yes	Yes
Observations	9,557	6,542	9,557	6,542
R^2	0.018	0.273	0.383	0.434
Adjusted R^2	0.018	0.269	0.379	0.426
Residual Std. Error	7.120 (df = 9555)	6.256 (df = 6508)	5.661 (df = 9496)	5.547 (df = 6450)
F Statistic	174.334***	74.087***	98.249***	54.279***
	(df = 1; 9555)	(df = 33; 6508)	(df = 60; 9496)	(df = 91; 6450)

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 3: *Math Teacher Efficacy Regression Results*

	(1)	(2)	(3)	(4)
VARIABLES	Basic	With Controls	FE	Controls + FE
Dependent variable: Math Test Score				
Math Teacher Efficacy	1.563***	0.736***	−0.393*	−0.155
	(0.128)	(0.146)	(0.203)	(0.285)
Constant	14.740***	9.597***	11.312***	9.641***
	(0.376)	(1.277)	(1.079)	(1.909)
Controls	No	Yes	No	Yes
Fixed Effects	No	No	Yes	Yes
Observations	9,146	6,643	9,146	6,643
R^2	0.016	0.209	0.241	0.314
Adjusted R^2	0.016	0.204	0.236	0.304
Residual Std. Error	7.210 (df = 9144)	6.539 (df = 6600)	6.353 (df = 9084)	6.114 (df = 6543)
F Statistic	149.568***	41.510***	47.352***	30.317***
	(df = 1; 9144)	(df = 42; 6600)	(df = 61; 9084)	(df = 99; 6543)

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

0.736 points per one-point increase in teacher efficacy respectively. Both are significant at the 1% level. However, when including fixed effects, the relationship flips, with the third regression showing a 0.393 decrease in test scores for a one-point increase in teacher efficacy and the final regression showing a 0.155 decrease in test scores for a one-point increase in teacher efficacy. These negative coefficients are surprising, but the third regression is only significant at the 10% level and the final regression is statistically insignificant.

Various robustness checks were completed and are included in the Appendix. Figures 7 and 8 show residual plots for the regressions. Due to the nature of the teacher efficacy score, the residuals are siloed, though the plots do not appear to show any strong correlation or major issues with the fitted values. Tables 11 and 12 are clustered standard errors by classroom. The results of the robustness check for English teacher efficacy still show the first two regressions having statistical significance at the 1% level, but when including fixed

effects, the coefficients are not statistically significant. For math teacher efficacy, statistical significance disappears when adding controls. Tables 13 and 14 are a heteroskedasticity robustness check. These checks maintain the statistical significance of the original regressions across both English and math teacher efficacy and across the four kinds of regressions.

Tables 4 and 5 show regression results using the standardized z-scores for both teacher efficacy and student test scores. This means that the coefficients represent the standard deviation increase for a one-standard deviation increase in teacher efficacy. Across both the English and math regressions, statistical significance remains the same as in the raw regressions, as do the signs of the coefficients. These results imply that even when standardizing teacher efficacy and student test scores, the relationship between the two remains.

Table 4: *Z-Score Regression Results: English*

	(1)	(2)	(3)	(4)
VARIABLES	Basic	With Controls	With FE	Controls + FE
Dependent variable: English Test Score (Z)				
English Teacher Efficacy	0.137***	0.127***	0.043***	0.085***
	(0.010)	(0.014)	(0.015)	(0.024)
Constant	0.003	−1.107***	−0.340**	−0.316
	(0.010)	(0.175)	(0.151)	(0.357)
Controls	No	Yes	No	Yes
Fixed Effects	No	No	Yes	Yes
Observations	9,557	6,542	9,557	6,542
R^2	0.018	0.273	0.383	0.434
Adjusted R^2	0.018	0.269	0.379	0.426
Residual Std. Error	1.001 (df = 9555)	0.879 (df = 6508)	0.796 (df = 9496)	0.780 (df = 6450)
F Statistic	174.334***	74.087***	98.249***	54.279***
	(df = 1; 9555)	(df = 33; 6508)	(df = 60; 9496)	(df = 91; 6450)

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 5: *Z-Score Regression Results: Math*

	(1)	(2)	(3)	(4)
VARIABLES	Basic	With Controls	With FE	Controls + FE
Dependent variable: Math Test Score (Z)				
Math Teacher Efficacy	0.130***	0.061***	−0.033*	−0.013
	(0.011)	(0.012)	(0.017)	(0.024)
Constant	−0.004	−1.127***	−1.250***	−0.731***
	(0.010)	(0.167)	(0.135)	(0.258)
Controls	No	Yes	No	Yes
Fixed Effects	No	No	Yes	Yes
Observations	9,146	6,643	9,146	6,643
R^2	0.016	0.209	0.241	0.314
Adjusted R^2	0.016	0.204	0.236	0.304
Residual Std. Error	0.997 (df = 9144)	0.905 (df = 6600)	0.879 (df = 9084)	0.846 (df = 6543)
F Statistic	149.568***	41.510***	47.352***	30.317***
	(df = 1; 9144)	(df = 42; 6600)	(df = 61; 9084)	(df = 99; 6543)

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Tables 15, 16, 17, and 18 in the Appendix show robustness checks for the z-score regressions. Tables 15 and 16 show a clustered standard errors check by classroom. Here again, statistical significance is lost when including fixed effects for the English regression and is lost when including controls and fixed effects for the math regression. Tables 17 and 18 show a heteroskedasticity robustness check. The results maintain the same relationships as previous robustness checks, with the English coefficients maintaining the same coefficients and statistical significance as the z-score regression and the math coefficients also maintaining the same coefficients and statistical significance as the z-score regression.

Discussion

Though these results are notable, there are several limitations. The first are common in teacher efficacy research: ability to measure strength of beliefs and the possibility of reverse causality. Measuring teacher efficacy is subjective and there are many models to measure it. Additionally, because teachers rank their own responses to the survey statements, their subjectivity and feelings about their own teaching will impact their response. The possibility of reverse causality is the most concerning, because teachers that are less effective may reply to generalized "all teachers" questions more negatively. However, due to the structure of the survey questions, with one of the two questions that make up the teacher efficacy score being a "teachers" and one being an "I" statement, the possibility of reverse causality is somewhat minimized.

Another limitation comes from the structure of the Young Lives School Survey: there is a lack of socioeconomic data. Given that socioeconomic status has been shown in previous studies to impact student success, this is unfortunate. However, survey questions about students' families, their work for pay, and other student controls can substitute partially for socioeconomic data. Finally, there is also a risk of omitted variable bias. Because of the sheer scale of the surveys, choices had to be made about which control variables to include. As a result, there is a possibility that omitted variable bias plays a role in the results.

Conclusions

There are four main conclusions that can be drawn from these results. First, in keeping with theory and the literature, teacher efficacy generally appears to have a small but statistically significant effect on student test scores. Notably, this relationship holds whether using raw test score and teacher efficacy data or standardized z-scores. This study found a stronger correlation between the two than previous research in developed countries, which could be for a myriad of reasons. Second, there is a greater effect of teacher efficacy on English test scores than on math test scores. It is difficult to draw conclusions about why this is, but it could imply that other factors, particularly locality, region, or school fixed effects more greatly impact math scores, whereas teacher efficacy has a greater impact on English scores.

Third, these results underscore the need for more research into teacher efficacy in a developing country context. Given the lack of research on this topic in developing countries and the general lack of strong empirical evidence even from developed countries, it is difficult to make any strong generalizations, or to say that teacher efficacy does and always will have an impact on student achievement. Additionally, given that this study was only conducted

in one country, one cannot generalize these findings to all developing countries. Finally, these results could imply that pursuing and investing in policies that improve teacher efficacy would be valuable in improving the quality of education in developing countries, and perhaps in improving school progression and student achievement in the long run.

References

[1] Bandura, A. (1978). The self system in reciprocal determinism. *American Psychologist*, 33(4):344–358.

[2] Berman, P., MacLaughlin, M. W., and Bass, G. (1977). Factors affecting implementation and continuation. Technical report, Rand.

[3] Blazar, D. and Kraft, M. A. (2017). Teacher and teaching effects on students' attitudes and behaviors. *Educational Evaluation and Policy Analysis*, 39(1):146–170.

[4] Bold, T., Filmer, D., Molina, E., and Svensson, J. (2019). The lost human capital: Teacher knowledge and student achievement in africa. Technical report, World Bank, Washington, DC.

[5] Brewer, D. J., Hentschke, G. C., and Eide, E. R. (2010). Theoretical concepts in the economics of education. In *International Encyclopedia of Education*, pages 193–198. Elsevier.

[6] Cents-Boonstra, M., Lichtwarck-Aschoff, A., Denessen, E., Aelterman, N., and Haerens, L. (2021). Fostering student engagement with motivating teaching: An observation study of teacher and student behaviours. *Research Papers in Education*, 36(6):754–779.

[7] Coleman, J. S. (1968). *Equality of Educational Opportunity*. U.S. Department of Health, Education, and Welfare.

[8] Cueto, S., León, J., Sorto, M. A., and Miranda, A. (2017). Teachers' pedagogical content knowledge and mathematics achievement of students in Peru. *Educational Studies in Mathematics*, 94(3):329–345.

[9] Gibson, S. and Dembo, M. H. (1984). Teacher efficacy: A construct validation. *Journal of Educational Psychology*, 76(4):569–582.

[10] Goddard, R. D., Hoy, W. K., and Hoy, A. W. (2000). Collective teacher efficacy: Its meaning, measure, and impact on student achievement. *American Educational Research Journal*, 37(2):479–507.

[11] Hanushek, E. A. (1986). The economics of schooling: Production and efficiency in public schools. *Journal of Economic Literature*, 24(3):1141–1177.

[12] Hanushek, E. A. and Woessmann, L. (2020). Education, knowledge capital, and economic growth. In *The Economics of Education: A Comprehensive Overview*, pages 171–181. Elsevier.

[13] Harris, D. N. and Sass, T. R. (2011). Teacher training, teacher quality and student achievement. *Journal of Public Economics*, 95(7–8):798–812.

[14] Kim, K. R. and Seo, E. H. (2018). The relationship between teacher efficacy and students' academic achievement: A meta-analysis. *Social Behavior & Personality: An International Journal*, 46(4):529–540.

[15] Moore, W. P. and Esselman, M. E. (1992). Teacher efficacy, empowerment, and a focused instructional climate: Does student achievement benefit?

[16] Muijs, D. and Reynolds, D. (2015). Teachers' beliefs and behaviors: What really matters? *The Journal of Classroom Interaction*, 50(1):25–40.

[17] Paisun, Syarifah, and Masuwd, M. (2024). Investigating the relationship between teacher self-efficacy and student achievement. *Andragogi: Jurnal Pendidikan Dan Pembelajaran*, 4(2):98–104.

[18] Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review*, 94(2):247–252.

[19] Rotter, J. B. (1966). Generalized expectancies for internal versus external control of reinforcement.

[20] Sanders, W. L., Wright, S. P., and Horn, S. P. (1997). Teacher and classroom context effects on student achievement: Implications for teacher evaluation. *Journal of Personnel Evaluation in Education*, 11(1):57–67.

[21] Sirait, S. (2016). Does teacher quality affect student achievement? an empirical study in Indonesia.

[22] Thompson, A. G. (2006). Teachers' beliefs and conceptions: A synthesis of the research. In *Handbook of Research on Mathematics Teaching and Learning*, pages 127–146. The National Council of Mathematics Teachers.

[23] Tschannen-Moran, M. and Barr, M. (2004). Fostering student learning: The relationship of collective teacher efficacy and student achievement. *Leadership and Policy in Schools*, 3(3):189–209.

[24] Tschannen-Moran, M., Hoy, A. W., and Hoy, W. K. (1998). Teacher efficacy: Its meaning and measure. *Review of Educational Research*, 68(2):202–248.

[25] UK Data Service (2025). UK data service.

[UNICEF Ethiopia] UNICEF Ethiopia. Learning and development. `https://www.unicef.org/ethiopia/learning-and-development`. Retrieved November 13, 2025.

[United Nations Department of Economic and Social Affairs] United Nations Department of Economic and Social Affairs. Goal 4: Quality education. `https://sdgs.un.org/goals/goal4`. Retrieved November 13, 2025.

[28] Young Lives (2025). Home page: Young lives.

[29] Young Lives Ethiopia (2017). Young lives school surveys 2016–17: Ethiopia country report.

Appendix

Table 6: *Summary Statistics: Z-Scores*

Statistic	N	Mean	St. Dev.	Min	Max
Math Teacher Efficacy (Z)	8,126	0.000	1.000	−3.161	1.852
English Teacher Efficacy (Z)	8,126	−0.000	1.000	−3.057	2.549
Math Score (Z)	8,126	0.000	1.000	−2.500	2.810
English Score (Z)	8,126	−0.000	1.000	−2.385	2.673

Figure 3: *Kernel Density: English Test Score Z-Scores*

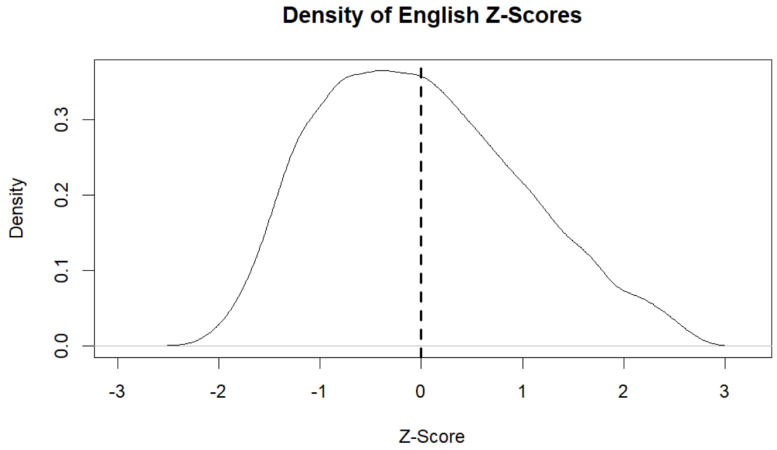


Figure 4: *Kernel Density: Math Test Score Z-Scores*

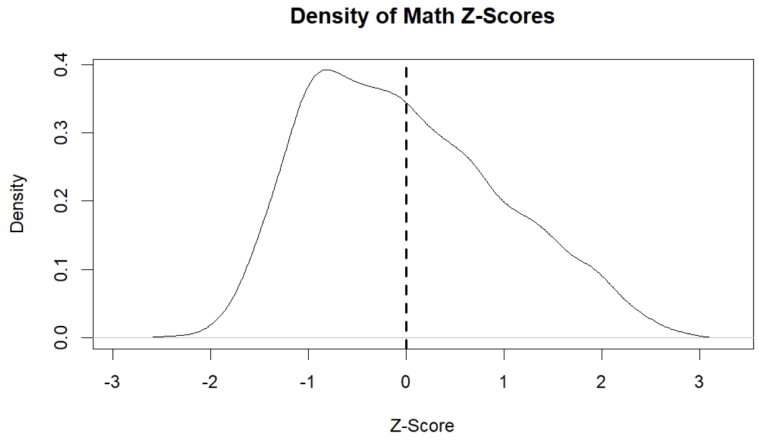


Figure 5: *Kernel Density: English Teacher Efficacy Z-Scores*

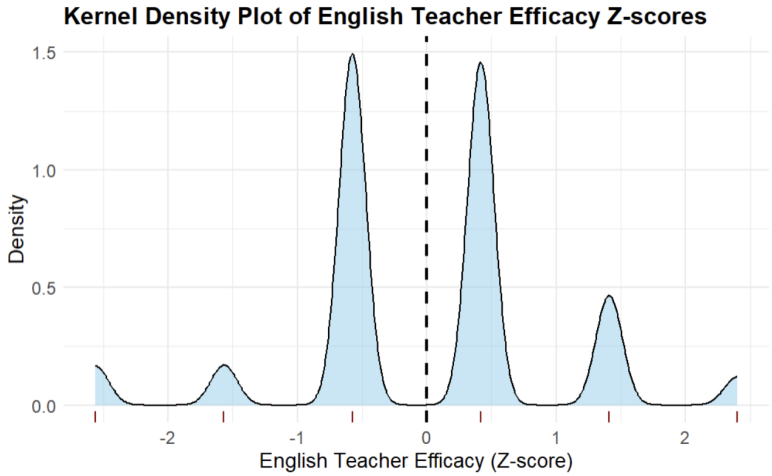


Figure 6: *Kernel Density: Math Teacher Efficacy Z-Scores*

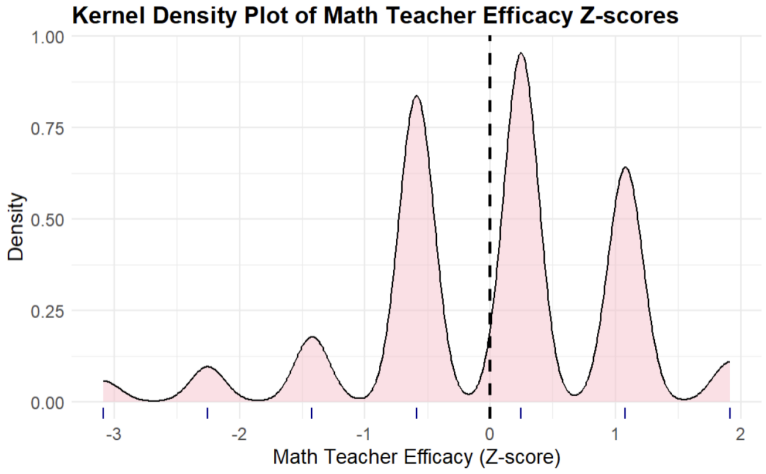


Table 7: *Summary Statistics by Region: English*

	English Teacher Efficacy					English Test Scores				
Region	N	Mean	SD	Min	Max	N	Mean	SD	Min	Max
Addis Ababa	2115	3.01	0.46	2.0	4.0	2005	23.73	6.91	5	40
Amhara	1185	2.79	0.35	2.5	4.0	999	17.50	5.43	4	39
Oromia	888	2.82	0.67	2.0	4.0	910	18.28	6.06	3	39
SNNP	2088	2.73	0.45	2.0	4.0	1997	21.11	7.10	5	39
Tigray	1654	2.98	0.35	2.5	3.5	1507	15.94	5.64	5	39
Somali	2571	2.58	0.59	1.5	3.5	1999	22.13	7.13	2	40
Afar	970	2.66	0.35	1.5	3.0	763	17.54	6.23	3	37

Table 8: *Summary Statistics by Region: Math*

	Math Teacher Efficacy					Math Test Scores				
Region	N	Mean	SD	Min	Max	N	Mean	SD	Min	Max
Addis Ababa	2302	3.09	0.51	2.0	4.0	2001	22.56	7.00	5	40
Amhara	1185	2.54	0.32	2.0	3.0	792	17.90	6.62	1	40
Oromia	847	3.15	0.30	2.5	3.5	892	19.77	7.43	1	40
SNNP	2322	2.93	0.53	2.0	4.0	1910	18.16	7.45	2	40
Tigray	1707	2.95	0.40	2.0	3.5	1420	17.92	6.55	4	38
Somali	2075	2.81	0.74	1.5	4.0	1942	18.83	7.10	2	40
Afar	970	2.16	0.66	1.0	3.0	776	17.34	6.27	3	36

Table 9: *Summary Statistics by Teaching Experience: English*

	English Teacher Efficacy					English Test Scores				
Experience	N	Mean	SD	Min	Max	N	Mean	SD	Min	Max
0–1	89	2.50	0.00	2.5	2.5	79	19.68	5.94	3	35
2–5	1427	2.80	0.46	1.5	4.0	1472	20.91	7.85	5	40
6–10	2668	2.93	0.39	2.0	4.0	2407	21.69	7.84	5	40
11–20	4289	2.85	0.37	2.0	3.5	3584	18.90	6.53	4	39
21+	2880	2.59	0.71	1.5	4.0	2548	20.41	6.36	2	39
NA	118	2.50	0.00	2.5	2.5	90	13.40	4.44	7	31

Table 10: *Summary Statistics by Teaching Experience: Math*

	Math Teacher Efficacy					Math Test Scores				
Experience	N	Mean	SD	Min	Max	N	Mean	SD	Min	Max
0–1	198	2.64	0.23	2.5	3.0	181	19.34	7.55	7	35
2–5	2330	3.02	0.47	1.5	3.5	1957	19.02	7.32	1	40
6–10	3752	2.85	0.68	1.0	4.0	3102	20.11	7.40	3	40
11–20	3507	2.69	0.61	1.5	4.0	3014	17.86	6.60	2	40
21+	1538	3.00	0.49	2.5	4.0	1409	20.56	7.48	1	40
NA	83	3.00	0.00	3.0	3.0	70	17.74	6.47	6	34

Figure 7: *Residual Plot: English Basic Regression*

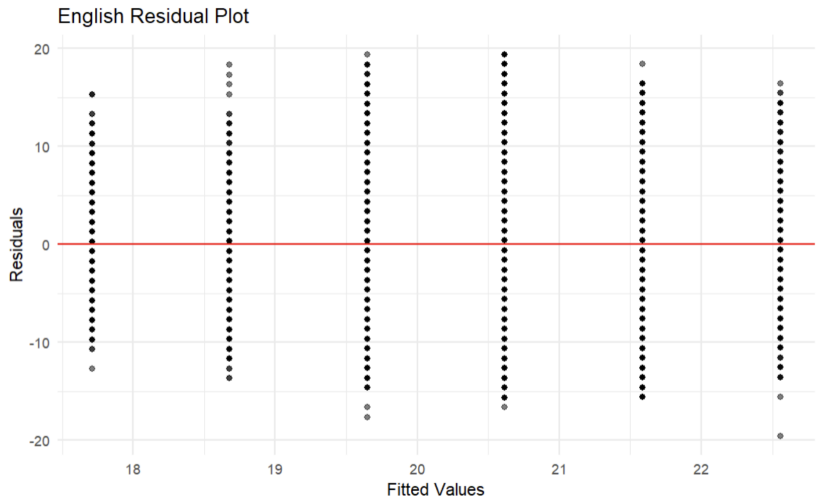


Figure 8: *Residual Plot: Math Basic Regression*

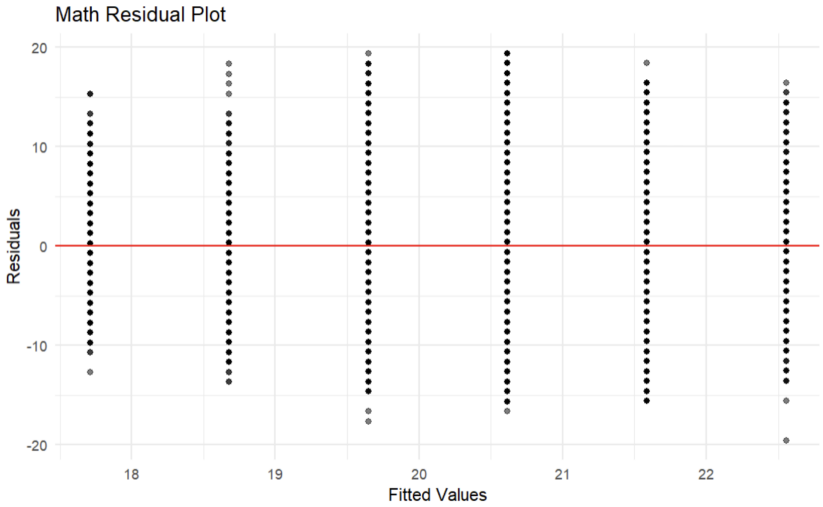


Table 11: *Clustered Standard Errors by School ID: English (Raw)*

	(1)	(2)	(3)	(4)
VARIABLES	Basic	Controls	FE	Controls + FE
Dependent variable: English Test Score				
English Teacher Efficacy	1.939***	1.796***	0.606	1.199
	(0.529)	(0.537)	(0.487)	(1.119)
Constant	14.801***	0.461	16.073***	9.027
	(1.415)	(2.951)	(1.968)	(7.169)
Controls	No	Yes	No	Yes
Fixed Effects	No	No	Yes	Yes
Observations	9,557	6,542	9,557	6,542
R^2	0.018	0.273	0.383	0.434
Adjusted R^2	0.018	0.269	0.379	0.426
Residual Std. Error	7.120 (df = 9555)	6.256 (df = 6508)	5.661 (df = 9496)	5.547 (df = 6450)
F Statistic	174.334***	74.087***	98.249***	54.279***
	(df = 1; 9555)	(df = 33; 6508)	(df = 60; 9496)	(df = 91; 6450)

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 12: *Clustered Standard Errors by School ID: Math (Raw)*

	(1)	(2)	(3)	(4)
VARIABLES	Basic	Controls	FE	Controls + FE
Dependent variable: Math Test Score				
Math Teacher Efficacy	1.563***	0.736	−0.393	−0.155
	(0.419)			
Constant	14.740***	9.597***	11.312***	9.641
	(1.198)	(2.951)	(1.968)	(7.169)
Controls	No	Yes	No	Yes
Fixed Effects	No	No	Yes	Yes
Observations	9,146	6,643	9,146	6,643
R^2	0.016	0.209	0.241	0.314
Adjusted R^2	0.016	0.204	0.236	0.304
Residual Std. Error	7.210 (df = 9144)	6.539 (df = 6600)	6.353 (df = 9084)	6.114 (df = 6543)
F Statistic	149.568***	41.510***	47.352***	30.317***
	(df = 1; 9144)	(df = 42; 6600)	(df = 61; 9084)	(df = 99; 6543)

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 13: *Heteroskedasticity Robustness Check: English (Raw)*

	(1)	(2)	(3)	(4)
VARIABLES	Basic	With Controls	FE	Controls + FE
Dependent variable: English Test Score				
English Teacher Efficacy	1.939***	1.796***	0.606***	1.199***
	(0.136)	(0.186)	(0.220)	(0.402)
Constant	14.801***	0.461	16.073***	9.027***
	(0.378)	(1.215)	(1.007)	(2.852)
Controls	No	Yes	No	Yes
Fixed Effects	No	No	Yes	Yes
Observations	9,557	6,542	9,557	6,542
R^2	0.018	0.273	0.383	0.434
Adjusted R^2	0.018	0.269	0.379	0.426
Residual Std. Error	7.120 (df = 9555)	6.256 (df = 6508)	5.661 (df = 9496)	5.547 (df = 6450)
F Statistic	174.334***	74.087***	98.249***	54.279***
	(df = 1; 9555)	(df = 33; 6508)	(df = 60; 9496)	(df = 91; 6450)

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 14: *Heteroskedasticity Robustness Check: Math (Raw)*

	(1)	(2)	(3)	(4)
VARIABLES	Basic	With Controls	FE	Controls + FE
Dependent variable: Math Test Score				
Math Teacher Efficacy	1.563***	0.736***	−0.393*	−0.155
	(0.121)	(0.143)	(0.209)	(0.300)
Constant	14.740***	9.597***	11.312***	9.641***
	(0.353)	(1.277)	(1.055)	(1.817)
Controls	No	Yes	No	Yes
Fixed Effects	No	No	Yes	Yes
Observations	9,146	6,643	9,146	6,643
R^2	0.016	0.209	0.241	0.314
Adjusted R^2	0.016	0.204	0.236	0.304
Residual Std. Error	7.210 (df = 9144)	6.539 (df = 6600)	6.353 (df = 9084)	6.114 (df = 6543)
F Statistic	149.568***	41.510***	47.352***	30.317***
	(df = 1; 9144)	(df = 42; 6600)	(df = 61; 9084)	(df = 99; 6543)

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 15: *Clustered Standard Errors by Classroom: English (Z-Score)*

	(1)	(2)	(3)	(4)
VARIABLES	Basic	Controls	FE	Controls + FE
Dependent variable: English Test Score (Z)				
English Teacher Efficacy	0.137***	0.127***	0.043	0.085
	(0.037)	(0.038)	(0.034)	(0.079)
Constant	0.003	−1.107***	−0.340	−0.316
	(0.043)	(0.325)	(0.253)	(0.824)
Controls	No	Yes	No	Yes
Fixed Effects	No	No	Yes	Yes
Observations	9,557	6,542	9,557	6,542
R^2	0.018	0.273	0.383	0.434
Adjusted R^2	0.018	0.269	0.379	0.426
Residual Std. Error	1.001 (df = 9555)	0.879 (df = 6508)	0.796 (df = 9496)	0.780 (df = 6450)
F Statistic	174.334***	74.087***	98.249***	54.279***
	(df = 1; 9555)	(df = 33; 6508)	(df = 60; 9496)	(df = 91; 6450)

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 16: *Clustered Standard Errors by Classroom: Math (Z-Score)*

	(1)	(2)	(3)	(4)
VARIABLES	Basic	Controls	FE	Controls + FE
Dependent variable: Math Test Score (Z)				
Math Teacher Efficacy	0.130***	0.061	−0.033	−0.013
	(0.035)			
Constant	−0.004	−1.127***	−1.250***	−0.731
	(0.037)	(0.325)	(0.253)	(0.824)
Controls	No	Yes	No	Yes
Fixed Effects	No	No	Yes	Yes
Observations	9,146	6,643	9,146	6,643
R^2	0.016	0.209	0.241	0.314
Adjusted R^2	0.016	0.204	0.236	0.304
Residual Std. Error	0.997 (df = 9144)	0.905 (df = 6600)	0.879 (df = 9084)	0.846 (df = 6543)
F Statistic	149.568***	41.510***	47.352***	30.317***
	(df = 1; 9144)	(df = 42; 6600)	(df = 61; 9084)	(df = 99; 6543)

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 17: *Heteroskedasticity Robustness Check: English (Z-Score)*

	(1)	(2)	(3)	(4)
VARIABLES	Basic	With Controls	FE	Controls + FE
Dependent variable: English Test Score (Z)				
English Teacher Efficacy	0.137***	0.127***	0.043***	0.085***
	(0.010)	(0.013)	(0.016)	(0.028)
Constant	0.003	−1.107***	−0.340**	−0.316
	(0.010)	(0.166)	(0.136)	(0.382)
Controls	No	Yes	No	Yes
Fixed Effects	No	No	Yes	Yes
Observations	9,557	6,542	9,557	6,542
R^2	0.018	0.273	0.383	0.434
Adjusted R^2	0.018	0.269	0.379	0.426
Residual Std. Error	1.001 (df = 9555)	0.879 (df = 6508)	0.796 (df = 9496)	0.780 (df = 6450)
F Statistic	174.334***	74.087***	98.249***	54.279***
	(df = 1; 9555)	(df = 33; 6508)	(df = 60; 9496)	(df = 91; 6450)

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 18: *Heteroskedasticity Robustness Check: Math (Z-Score)*

	(1)	(2)	(3)	(4)
VARIABLES	Basic	With Controls	FE	Controls + FE
Dependent variable: Math Test Score (Z)				
Math Teacher Efficacy	0.130***	0.061***	−0.033*	−0.013
	(0.010)	(0.012)	(0.017)	(0.025)
Constant	−0.004	−1.127***	−1.250***	−0.731***
	(0.010)	(0.165)	(0.130)	(0.246)
Controls	No	Yes	No	Yes
Fixed Effects	No	No	Yes	Yes
Observations	9,146	6,643	9,146	6,643
R^2	0.016	0.209	0.241	0.314
Adjusted R^2	0.016	0.204	0.236	0.304
Residual Std. Error	0.997 (df = 9144)	0.905 (df = 6600)	0.879 (df = 9084)	0.846 (df = 6543)
F Statistic	149.568***	41.510***	47.352***	30.317***
	(df = 1; 9144)	(df = 42; 6600)	(df = 61; 9084)	(df = 99; 6543)

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Income and Mortgage Choice in the Conforming and Jumbo Markets

Daniel Wiltse

ECON 381: Introduction to Econometrics (Advisor: Amy Damon)

Introduction

Prospective home buyers in the United States can choose between two general mortgage structures: fixed-rate and adjustable-rate mortgages. Standard fixed-rate mortgages (FRMs) have a constant nominal payment plan designed to pay off the principal balance and interest over the term of the loan. Adjustable-rate mortgages (ARMs) require the borrower to pay a floating nominal interest rate indexed to the return on short-term securities, most commonly the 30-day average secured overnight funds rate, as well as additional payments to amortize the principal balance (11). ARM terms typically stipulate an initial rate period, during which rates cannot be adjusted, as well as adjustment intervals that specify how frequently rates can be adjusted. A 5/6 ARM, for example, has an initial period of five years and an adjustment period of 6 months.

Historically, ARMs were a staple of subprime mortgage markets where credit-constrained borrowers utilized lower initial ARM rates to qualify for larger loans. This makes sense for households that expect stable or declining interest rates or anticipate an increase in the value of their home, allowing them to refinance to a FRM if rates rise. The 2008 financial crisis disrupted this decision-making by shattering households' expectations of consistent home value appreciation and housing market stability. What role do ARMs play in the mortgage market since the Financial Crisis ended? Recent Home Mortgage Disclosure Act (HMDA) data suggest that the clientele of ARM loans has shifted substantially. Whereas previously the vast majority of ARMs originated from subprime and prime conforming markets, now a third of ARMs originate from eleven major metropolitan areas are concentrated in the relatively small jumbo loan market. This change suggests that wealthier households now make up a higher share of ARM borrowers and are using ARMs to leverage and hedge their investment portfolios. Section I provides a commentary on the literature that informed this project, section II presents a simple theoretical model, section III outlines my empirical model, section IV describes my dataset and section V analyses my results. I conduct robustness testing in section VI, discuss limitations in section VII and summarize my thoughts in section VIII

Literature Review

FRMs transfer interest rate and inflation risk to lenders, who charge a risk premium. FRM borrowers benefit from the predictability of stable nominal payments in return for paying the risk premium (4). ARMs shift interest rate risk to borrowers, whose payments are

[3]Daniel Wiltse is a Junior majoring in Economics. Correspondence concerning this article should be addressed to dwiltse@macalester.edu.

indexed to short-term rates. This exposes borrowers to payment shocks but eliminates the risk premium, resulting in lower initial rates. If the national monetary authority adheres to the Taylor principle, interest rates should mirror inflation, mitigating ARM borrowers' real gain from inflation (8). The risk dispersion of ARMs makes them attractive to several different categories of borrowers. Borrowers who plan to move in the near future can benefit from the lower initial rates and then move before the first adjustment period. Credit-constrained households can also benefit from ARMs (7). The lower initial rates on ARMs result in a lower debt-to- income ratio for an equivalently sized loan, allowing constrained borrowers to qualify for larger loans than if they financed with a FRM (9). Lower initial rates may also appeal to borrowers who are currently payment-constrained but expect their income to increase substantially in the future (5). All of these circumstances are prevalent among young borrowers with relatively low wealth and income. ARMs allow these constrained borrowers to push their credit further. There is, however, the risk that simultaneous decreases to income and interest rate hikes may stress ARM borrowers' budgets, necessitating decreased consumption, or in extreme cases, delinquency or default.

ARMs also appeal to borrowers who aren't at high risk of binding payment constraints, especially those with large capital or bond holdings (3). These households don't have to worry about rate increases because the return on their stocks and bonds will also likely increase in an inflationary environment. If not, the borrower can easily sell capital holdings and either buy higher return bonds or delever by paying down their mortgage. In this case, ARM mortgages act as cheap leverage rather than a means to push constraints (2). Andersen and his coauthors describe the ARMs as having a "dual clientele that includes both borrowers with very low income and high bank debt, who are plausibly constrained, and borrowers with high financial assets who may be using ARMs to leverage financial investments." This is an important consideration when examining the national ARM share, as it aggregates two distinct groups of ARM borrowers that respond differently to economic shocks.

Another important macro-level driver of mortgage choice is the rate spread between ARMs and FRMs. As previously discussed, ARMs usually have lower initial rates than FRMs because they don't include the risk premium. In practical terms, the rate spread tracks the spread between long-term rates, which FRMs are generally indexed to, and the short-term rates to which ARMs are indexed. If the spread between FRM and ARM rates is relatively large, all categories of borrowers will be much more inclined to finance with ARMs (12; 11; 10).

A large pre-2008 financial crisis literature studies mortgage choice through the lens of interest-rate and credit-constraints. Campbell and Cocco (5) build a rich theoretical model which shows ARMs can be optimal for households with optimistic expectations about their future income. Coulibaly and Li (7) use a logit model to study how the rate spread between ARM and FRM rates effects mortgage choice and find that a one standard deviation in the rate spread leads homeowners to be 42 percent more likely to finance with an ARM. They also examine measures or credit constraint and find that households that spent more than they earned were 27 percent more likely to finance with an ARM, and that households with larger mortgage balance to income ratios were more likely to have ARMs. Johnson and Li (9) use a probit model and find that ARM borrowers are 6 percent more likely to utilize a high portion of their credit card limits and around 10 percent more likely to hardly ever pay

off their credit card balances each month. These three studies generally agree that ARM borrowers in the pre-2008 housing market tended to be young, risk-neutral, and somewhat credit constrained.

The dual clientele that Andersen and his colleagues describe is present in pre-crisis data, but the effect of high-income ARM borrowers is muted by the sheer volume of ARMs and the high proportion of low-income borrowers. For example, Moench et al. (12) used Lender Processing Services data to compute the ARM share in certain mortgage sub-markets in the U.S. between 2005-2010. In 2005, the year with the highest total ARM share, 9.5% of government mortgages, 52.5% of nonprime mortgages, 30.3% of prime conforming mortgages, and 68.7% of prime jumbo mortgages were adjustable-rate, for a total of 38.5% of the total US market. After the 2008 crash, the share of adjustable-rate mortgages plummeted. In 2009, ARMs made up only 2.3% of national originations, with nonprime loans having the lowest concentration of only 0.8%. Prime jumbo ARMs maintained their share much better, only falling to 43.8%. Only 2.2% of prime conforming loans and 1.9% of government loans were ARM, indicating that prime jumbo loans constituted a much higher share of total ARM originations in 2009 compared to 2005. Moench, Vickery, and Aragon's statistical findings suggest a major shift within the ARM market. The market for sub- prime ARMs appears to have collapsed after the 2008 financial crisis, while the market for jumbo ARMs stayed relatively stable. This distribution has continued long after the financial crisis abated. In a 2024 study using 2019 SCF data, Chiang and Dueholm found that younger borrowers with high incomes tended to finance with ARMs (6). This paper contributes to the established mortgage choice literature by reexamining the U.S mortgage market in the post- crisis period, arguing that the 2008 financial crisis fundamentally altered the composition of ARM borrowers, such that income is now positively associated with ARM adoption.

Theoretical Model

In this section, I present a basic two-period model that supports my hypothesized relationship between income and ARM choice. In the first period, the household chooses one of the following mortgage structures to maximize its utility, subject to its intertemporal budget constraint.

$$\mathbb{E}[U] = \ln(c_0) + \beta\mathbb{E}[\ln(c_1)] \tag{1}$$

$$\text{FRM} : M_0^F = M_1^F \tag{2}$$

$$\text{ARM} : M_0 = M_0^A, \quad M_1 = M_0^A + \gamma \tag{3}$$

$$c_0 = Y - M_0 + B_0 - Q_0 B_1 \tag{4}$$

$$c_1 = Y - M_1 + B_1 \tag{5}$$

$$\mathbb{E}[M_1^A] = M_1^F + \eta_0 \tag{6}$$

In the budget constraints, Y denotes the household's income, which is held constant for simplicity, and M_t denotes the mortgage payment the household must pay in each period. The FRM payment is held constant while the ARM payment varies via γ, a random, normally distributed shock centered on mean 0. B_0 and B_1 are real bonds that the household can use to smooth their consumption, Q_0 is the price households can exchange bonds for in period 1. Equation (6) represents the risk discount associated with ARMs where η_0 is the time 0 risk premium. The household expects the ARM payment to be slightly lower than the FRM payment, but recognizes the risk that the ARM payment could be much higher if interest rates rise. Households with lower income may opt for an FRM so that they can maintain a stable level of consumption in both periods, while a household with a higher income may prefer ARMs because they offer a lower expected payment, and the household's higher income guarantees that their consumption will not decline significantly.

Empirical Model

The theory that ARMs now predominantly appeal to wealthier households can be tested empirically using linear probability and logit models. Coulibaly and Li and Johnson and Li exclusively use logit and probit models because ARMs were relatively common in their pre-2008 data. A linear probability model (LPM) will give a more accurate estimate in the modern conforming markets where ARMs are very rare. I supplement the linear regression on the jumbo market with a logit model which will be properly specified due to the higher share of ARMs.

$$P(\text{ARM}_i) = \alpha_i + b_1 \ln(\text{income}) + b_2(\text{constraint score}) + b_3(\text{age}) + \alpha_{\text{race}} + \alpha_{\text{metro}} + \alpha_{\text{time}} \quad (7)$$

$$P(\text{ARM}_i) = \frac{1}{1 + e^{\alpha_i + b_1 \ln(\text{income}) + b_2(\text{constraint score}) + b_3(\text{age}) + \alpha_{\text{race}} + \alpha_{\text{metro}} + \alpha_{\text{time}}}} \quad (8)$$

Income is the primary explanatory variable, which directly measures the household's income, but should also be a reasonable proxy for its capital holdings and wealth more broadly. Households with a higher annual income can afford to invest more of their money in stocks and bonds. To identify credit-constrained borrowers, I created a variable called "Constraint Score." This score ranges from one to three: one denotes unconstrained borrowers with debt-to-income ratios below 36%, two denotes somewhat constrained borrowers with debt-to-income ratios between 36% and 45%, and three denotes seriously constrained borrowers with debt-to-income ratios above 45%. These classifications reflect the Freddie Mac seller guidelines. Freddie Mac requires an explanation for loans with debt-to-income levels above 36%, and does not purchase loans with debt-to-income levels above 45%. The debt-to-income levels are calculated counterfactually, using average prime offer rates to remove reverse causality between the borrower's debt-to-income level and whether they financed with an ARM. High-income borrowers are likely to be less credit constrained than lower-income borrowers; therefore, how constrained a borrower is, likely correlates with their income and affects mortgage choice. Controlling for credit constraints removes this omitted variable bias. Age serves a similar purpose. Previous studies posit that younger

borrowers may prefer ARMs because they tend to be more risk-neutral than older borrowers (9). Adding age into the regression controls for this relationship, as well as the possibility that this relationship has flipped, and older, more financially stable borrowers maybe less risk-averse than younger borrowers who grew up during the 2008 crisis. Metro and time fixed effects control for the effects of local and national economic shocks, such as fund rate changes or local housing affordability on ARM originations, including racial fixed effects controls for how different communities perceive mortgage risk. In the 20th century, black Americans were systematically excluded from government subsidized mortgages and targeted by predatory lenders (13; 1). This historic marginalization likely influences how black borrowers internalize mortgage risk and could bias my income estimator.

Data Description

The core of the data set is from the Home Mortgage Disclosure Act (HMDA) database and includes every first-lien, owner-occupied, approved mortgage application between 2018 and 2024 from the Austin, Minneapolis, Cincinnati, Baltimore, Denver, Kansas City, Phoenix, Pittsburgh, Portland, Riverside, and San Diego metropolitan areas. The decision to limit the sample to these metros is due to technical limitations on my ability to process the full national-level dataset. Focusing on these metropolitan areas ensures wealth, racial, and regional heterogeneity, which allows for trend identification. Conversely, limiting the data to major metropolitan areas will likely overestimate the jumbo and conforming components of the national ARM share, because there is a higher concentration of wealth in major cities than in rural areas. To mitigate this bias, I excluded wealthier metros such as New York, Los Angeles, and San Francisco. Nonetheless, it is still necessary to recognize that these results apply to large metropolitan areas and do not necessarily hold for the nation as a whole.

Another important caveat is the data set's classification of loans as ARMs or FRMs. HMDA data does not include an indicator variable as to whether a loan is fixed or adjustable-rate; instead, it reports the initial rate period for each loan. Any loan with an initial rate period should be an adjustable-rate loan but many loans were listed with erroneous initial-rate periods. Including these loans results in an ARM share over 30%, which doesn't align with Federal Housing Finance Ad- ministration data, which reports the quarterly ARM origination share fluctuated between 1% and 8% between 2018 and 2024. Restricting my ARM classier to only include loans with adjustment periods between 6 and 120 months yielded a total ARM share of 3.6%. This classifier may fail to identify some unconventional ARM products, such as those with extremely aggressive teaser rates or very long initial periods; however, the indicators' summary statistics mirror reports from aggregated government data well enough that I am confident in its accuracy.

Breaking the sample into different mortgage markets allows me to examine ARM and FRM borrower heterogeneity within the conforming and jumbo markets. In both the conforming and jumbo markets, ARM borrowers tend to be older, have higher incomes, lower debt-to-income levels, take out larger loans, and more often opt for longer-term mortgages. ARM loans appear to have higher interest rates than FRMs, but this is misleading. Borrowers are more likely to finance with an ARM when current rates are high because borrowers believe that rates will go down in the future. The rate spread (APR-APOR) indicates that ARM borrowers tend to pay lower interest rates than they otherwise would

Table 1: *Summary Statistics: Mean Values*

	(1) ARM	(2) FRM	(3) Conf ARM	(4) Conf FRM	(5) Jumbo ARM	(6) Jumbo FRM
Total	159,414	4,275,818	99,070	4,077,039	60,344	198,779
Income	241,526	121,747	150,911	113,280	390,294	295,423
Loan Amount	627,528	344,218	372,209	319,236	1,046,700	856,630
Loan Term	356.93	338.06	354.94	337.35	360.19	352.61
Age	46.78	43.62	46.57	43.57	47.12	44.64
Debt-to-Income	34.54	35.52	34.67	35.58	34.32	34.37
Interest Rate	4.27	4.02	4.41	4.03	4.04	3.95
APR-APOR	−0.02	0.30	−0.01	0.31	−0.05	0.09

if they financed with an FRM. The average spread is 0.32 percentage points between conforming ARM and FRM borrowers, but is only 0.16 percentage points in the jumbo market. This difference could be the result of two phenomena: either conforming borrowers are more sensitive to changes in the rate-spread, and only choose ARMs when the spread is high, or lenders do not offer as high a risk premium to jumbo lenders. The most surprising statistic is the vast difference in income between ARM and FRM borrowers in both the conforming and jumbo markets. In the conforming market ARM borrowers have an average income of $150,911 while FRM borrowers have an average income of $113,280. In the jumbo market ARM borrowers make an average of $390,294 a year compared to $295,423 for FRM borrowers. When examining the income variable, it is important to keep in mind that this is the income used to underwrite the loan, which, in the case of a married couple, includes both partners, giving context to surprisingly high values.

Figure 1 provides more context for the mean income levels. ARM share is slightly elevated for the poorest 1% of borrowers, but is very low among the poorest 30% of borrowers. Around the 30th percentile, the share increases linearly until approximately the 80th percentile, at which point it begins increasing exponentially. The gradual increase of ARM share within the middle percentiles suggests that middle-class conforming borrowers see some utility in them. This makes sense as some of these borrowers are likely to own some investments, which give them a moderate buffer to insulate them from rate shocks. However, the increase is slow and linear, which suggests that most of the benefits from using ARMs as low-cost leverage aren't realized outside of the top 10% percent of incomes in the sample.

Results

This section presents the empirical results of the linear probability and logit models. I begin by analyzing the results from the conforming market, followed by the results in the jumbo market. The models suggest there is a strong positive relationship between income and ARM choice in both markets, and that this relationship is amplified in the jumbo market. The models also suggest that age, race, and credit constraints are strongly correlated with ARM choice.

In the conforming univariate regression, the natural log of income has a coefficient of 0.0211, which suggests that a one-percent increase in a borrower's income is associated with a 0.0211 percentage point increase in the probability they finance with an ARM. To

Figure 1

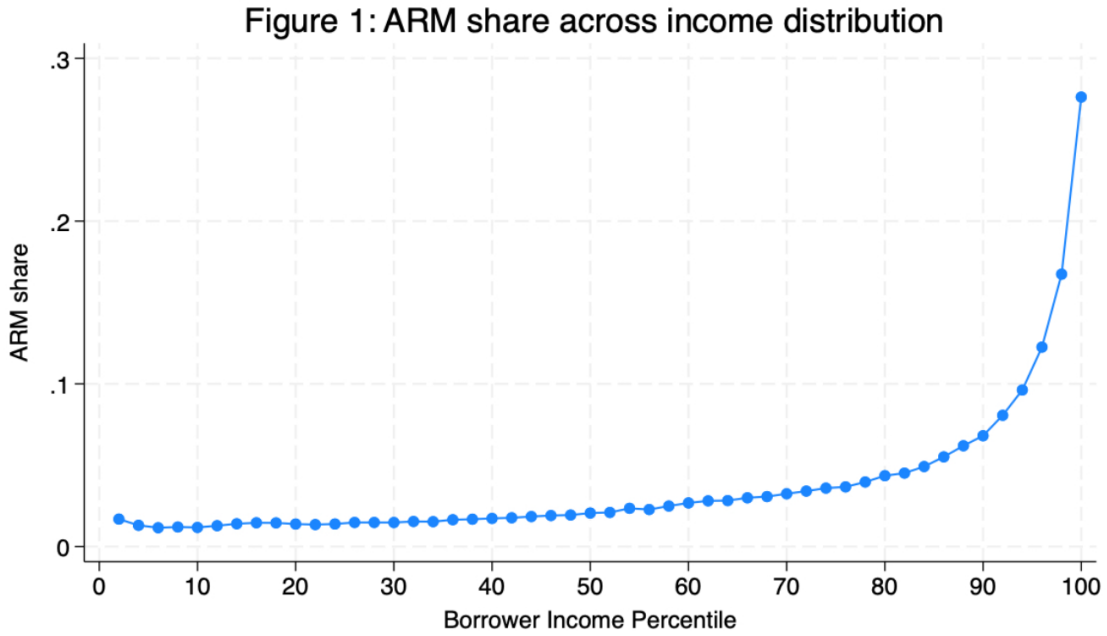


Table 2: *Mortgage Choice Models*

	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	Conf: Univar	Conf: Multivar	Conf: FE	Jumbo: Multivar	Jumbo: FE	Jumbo: Logit
ln(Income)	0.0211***	0.0216***	0.0189***	0.174***	0.146***	0.835***
	(0.000139)	(0.000144)	(0.000153)	(0.00164)	(0.00178)	(0.0105)
Age		0.000395***	0.000527***	0.00252***	0.00299***	0.0184***
		(5.44e-06)	(5.46e-06)	(7.35e-05)	(7.36e-05)	(0.000445)
Constraint Score		0.00141***	−0.00125***	0.0474***	0.0360***	0.199***
		(0.000102)	(0.000106)	(0.00127)	(0.00130)	(0.00795)
Black		−0.00948***		−0.0766***		
		(0.000341)		(0.00489)		
Constant	−0.219***	−0.244***	−0.214***	−2.130***	−1.827***	−13.24***
	(0.00160)	(0.00173)	(0.00205)	(0.0211)	(0.0259)	(0.164)
Observations	4,176,109	4,176,109	4,175,751	259,123	259,103	259,095
R-squared	0.005	0.007	0.023	0.051	0.083	

Standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

better illustrate this relationship, consider moving from the 25th income percentile to the 75th percentile. A bor- rower earning $69,000 has a 1.61% chance of financing with an ARM, while a borrower earning $140,000 has a 3.10% chance of financing with an ARM. This represents a 1.7 percentage point increase, or a 92% percent increase relative to the lower-income baseline. Adding controls and fixed effects decreases the magnitude of this effect slightly but does not significantly alter it, suggesting that this relationship is robust. This aligns with the intuition of the theoretical model, which suggests that a higher income reduces the marginal utility cost of unexpected rate hikes, making ARMs an attractive financial instrument despite their risk.

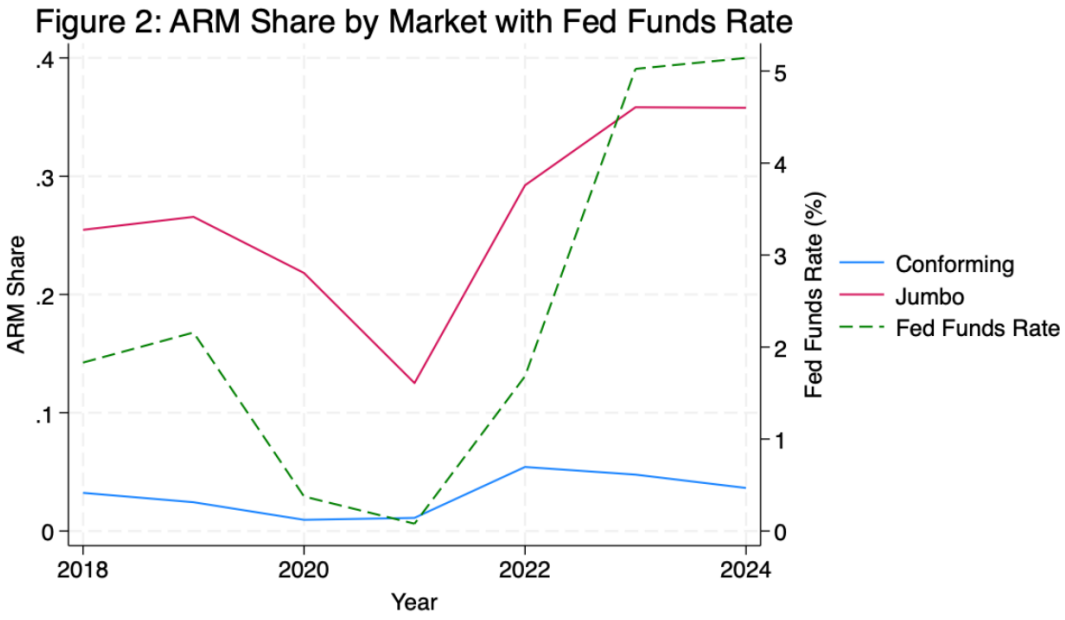
The coefficient associated with age is significant in the control and fixed-effect models. In the fixed-effect model, the coefficient is equal to 0.000527, which suggests that a one-year increase in a borrower's age is associated with a 0.0527 percentage point increase in the probability they finance with an ARM. Holding all else constant, a 50-year-old borrower is 1.054 percentage points more likely to finance with an ARM than a 30-year-old borrower. This is a surprising result because in many studies using pre-2008 data, younger borrowers were perceived as being less risk- averse and therefore more likely to finance with an ARM (9). This reversal may reflect changes in how different generations perceive risk. Younger borrowers may remember their parents struggling during the 2008 Financial Crisis, which may cause them to be more risk- averse when borrowing. Young borrowers may also be generally anxious about their economic future because of rising national debt, growing consciousness about the effects of global warming, increased automation, and generative artificial intelligence. All of these factors lower young borrowers' expectations of their future net income, which, according to my theoretical model and the established literature, should lower their probability of financing with an ARM.

The only coefficient that differs substantially between models is the one associated with the borrower's credit constraint score. The coefficient has a value of 0.00141 in the control regression and -0.00125 in the fixed effect regression. This indicates that a time-varying, metro-varying, or race-varying variable has a strong positive relationship with ARM choice and is strongly correlated with an applicant's credit-constraint score. One explanation is that ARM popularity increased during the post-COVID 2021-2023 rate hike cycle. While the economy rebounded quite quickly after COVID, it is plausible that the lingering pandemic-related financial stress and the contraction caused by the rate hikes increased the number of credit-constrained borrowers while ARMs grew in popularity. Another explanation is that different metros may have different average debt burdens, and that heavily indebted metros may coincidentally have a higher share of ARM originations. Yet another explanation is that some racial groups are more credit-constrained than others, and that more-constrained races also prefer ARMs. I believe that the COVID rate hike explanation is most likely true (see figure 2). Once I add fixed effects, the credit constraint coefficient reflects the theoretical model. More constrained borrowers are less likely to finance with an ARM because an unexpected rate shock is likely to decrease their future consumption. It is hard to distinguish the differing effects of payment and credit constraints, because my credit constraint score is calculated using the debt-to-income level, which reflects both payment and credit constraints. However, the fact that this coefficient is negative suggests that in the conforming market, payment and credit- constrained households are more likely to be cautious and opt for an FRM than try to use an ARM to push their credit constraints.

The binary indicator for whether the applicant is black has a coefficient of -0.00948, which suggests that black borrowers are 0.948 percentage points less likely to finance with an ARM compared to comparable applicants of other races. This supports the theory that the historical subprime targeting of black Americans influences their perception of mortgage risk. It could also indicate that black Americans are substantially more pessimistic about their future income growth or hold fewer capital assets, which would help insulate them from future rate hikes. Any combination of these factors suggests that black borrowers feel more exposed to economic fluctuations.

My results suggest that there is a statistically and economically significant positive

Figure 2



relation between income and ARM choice in the conforming market. The magnitudes of the coefficients are small because the ARM share in the conforming market is quite low. These results suggest that in the conforming market, wealthier, older, less constrained borrowers are more likely to finance with an ARM.

While the conforming market mostly encompasses middle and upper-middle-class households, the jumbo market is almost exclusively composed of upper-class households. Borrowers in the conforming market have a median income of $98,000, while those in the jumbo market have a median income of $260,000. Jumbo borrowers' higher income allows them to accumulate capital assets, which they can effectively leverage using ARMs. The returns on these assets often move with benchmark interest rates, effectively hedging against ARM rate shocks. The combination of low-cost leverage and effective hedging makes ARMs particularly attractive to wealthy jumbo borrowers.

In the jumbo fixed effects model, the natural log of income has a coefficient of 0.146, suggesting that a one percent increase in income is associated with a 0.146 percentage point increase in the probability that the borrower will finance with an ARM. This is a much larger effect than in the conforming market, but ARMs make up 23.3% of jumbo originations compared to 2.4conforming originations. Using my fixed effects model, I can compare the 25th and 75th income percentiles. Holding all else constant, a household making $189,000 dollars a year has a 17.95percent probability of financing with an ARM, while a household making $369,000 dollars a year has a 27.77% percent chance of financing with an ARM. This constitutes a 9.82 percentage point increase, or a 54.7% percent increase relative to the lower-income baseline. While the magnitude of the difference in the jumbo market is much larger than in the conforming market, the percent increase is smaller. This makes

sense because the economic conditions of two households making $189,000 and $369,000 are more similar than the economic conditions of households making $69,000 and $140,000 a year. In the jumbo market, the vast majority of households have sufficient wealth that they do not face significant default risk and are only differentiated by their ability to realize the wealth benefits of ARMs. In the conforming market, there is a greater diversity of economic circumstances. In the lower-income range, there are households with no capital assets, pessimistic expectations of future income growth, and a minuscule income buffer. At the upper end, households are relatively wealthy and have substantial capital assets, optimistic expectations of future income growth, and the flexibility to absorb rate shocks.

Another important difference between the models is that they give very different estimates for comparable households financing in the jumbo and conforming markets. My conforming fixed- effects model predicts that a household making $189,000 a year has a 3.6% probability of financing with an ARM, while the jumbo fixed-effects model predicts a 17.95% probability. This could be the result of the linear probability models struggling to accurately predict the extreme ranges of the income distributions within their respective markets, a weakness I address during robustness testing. Another explanation is that there are differences in credit availability in the jumbo and conforming markets. Jumbo loans cannot be bought by Freddie Mac and Fannie Mae, and lenders may be hesitant to supply large fixed-rate loans, which they may be unable to sell. I partially address this supply-side heterogeneity in my robustness tests by using lender fixed-effects, which will control for heterogeneity between lenders but would not effectively control for a market-wide trend.

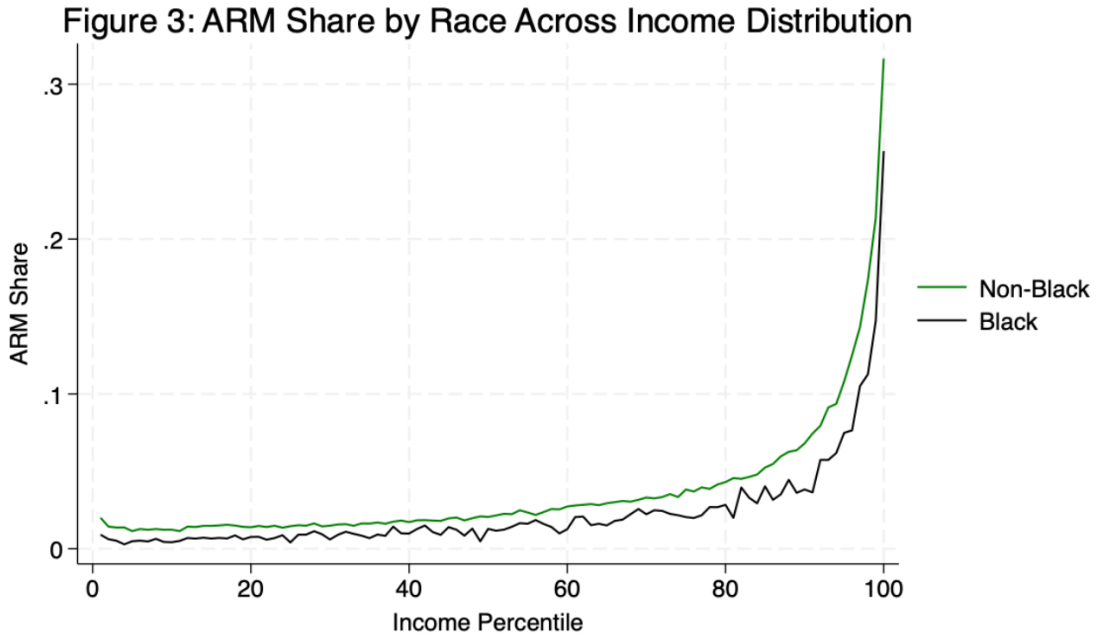
Age is significant and positively correlated with ARM choice in the jumbo market. The coefficient of 0.00299 suggests that a one-year increase in a borrower's age is associated with a 0.299 percentage point increase in the probability that the borrower will select an ARM. Estimating the difference between the 25th and 75th age percentiles (jumbo borrowers tend to be older than conforming borrowers) using the fixed-effects model returns a probability of 21.73% for 40-year- old borrowers and 24.71% for 50-year-old borrowers. In the jumbo market, age may proxy less for risk-aversion and more for accumulated wealth beyond income. Jumbo borrowers don't need to worry about rate hikes as much as conforming borrowers, so the generational differences in risk aversion shouldn't play as large of a role in ARM choice. Instead, older jumbo borrowers likely have more accumulated wealth, which isn't represented in their current income. Older borrowers may also be planning for retirement and be more interested in the sort of leveraged investment strategies that ARMs facilitate.

The constraint score variable has a positive coefficient in the jumbo models, suggesting that there is a positive relationship between constraints and ARM choice in the jumbo market. The coefficient of 0.0360 suggests that a one-point increase in a borrower's constraint score is associated with a 3.6 percentage point increase in the probability they will finance with an ARM. I see two possible explanations. The first is that a jumbo borrower's debt-to-income ratio is not an accurate prediction of whether they are constrained, because jumbo borrowers likely make a higher percentage of their income in capital gains. Borrowers who have relatively high capital asset holdings relative to their income can finance an expensive house with a high debt-to-income ratio while not being actually credit or payment constrained. These households are exactly the ones my theory predicts should benefit the most from ARMs. The other plausible explanation is that jumbo households

utilize ARMs to push credit constraints in the same way low-income households did in the 2000s. Consider a wealthy jumbo household trying to finance a very expensive home. Even though this household is wealthy, they may be credit-constrained if they want to buy an expensive home. Financing with an ARM may allow them to qualify for this larger loan, and their capital holdings likely make them better able to absorb rate hikes than a low-income household trying to finance a house that is similarly beyond their means.

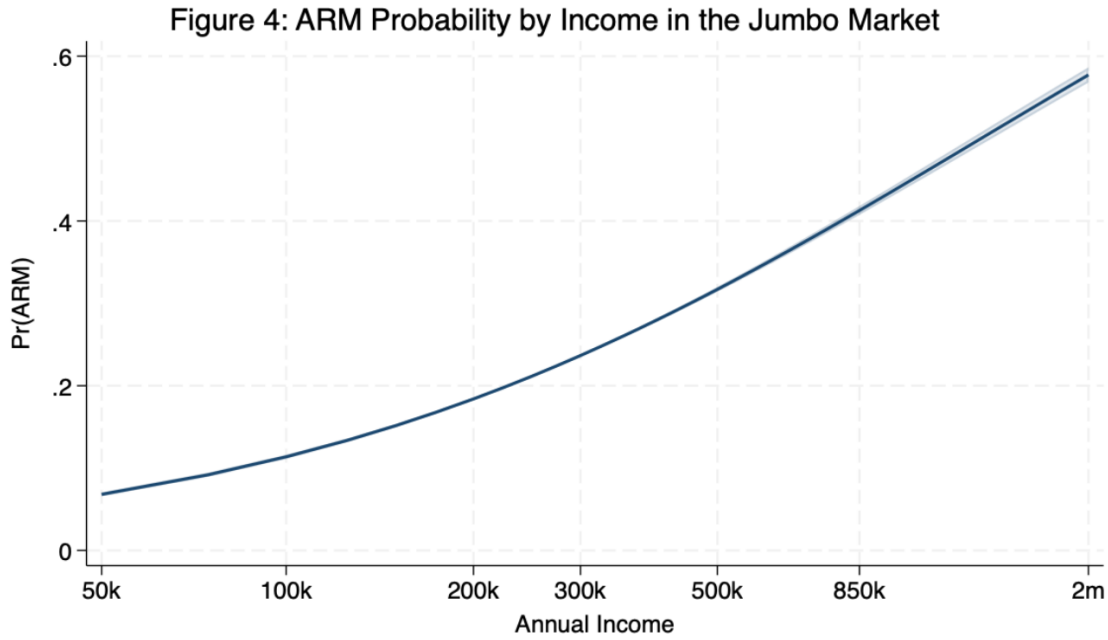
Black borrowers in the jumbo market are similarly wary of ARMs. The coefficient of -0.0766 suggests that black jumbo borrowers are 7.66 percentage points less likely to finance with an ARM compared to other races in the sample. This suggests that wealthy black borrowers internalize mortgage risk similarly to poorer black borrowers, and that this difference is not due to current financial circumstances but rather from the legacy of past discrimination (see figure 3). One explanation is that wealthy black borrowers distrust mortgage lenders due to historical sub- prime targeting. Another is that wealthy black borrowers may not have inherited as much financial knowledge from their parents as wealthy white borrowers, due to historical discrimination which prevented black households from accumulating wealth. This may lead black borrowers to feel more comfortable with the relatively simple structure of an FRM. The difference in ARM rate may also be the result of lenders who are overly cautious about suggesting riskier loans to black borrowers because they are conscious of historical predatory lending. These racial differences across income levels are fascinating and deserve further research.

Figure 3



The larger ARM share in the jumbo market allowed me to run a logit model. Logit models give bounded predictions, which more accurately model relationships at the extreme ranges of the independent variable. The logit model failed to converge in the conforming

Figure 4



model, but it successfully ran in the jumbo model. Figure 4 provides an interpretable visualization of the logit model. Holding all else constant, a household with an annual income of $200,000 has an 18.40% probability of financing with an ARM. A household making $500,000 a year has a 31.70% probability of financing with an ARM, and a household making $850,000 has a 41.27% chance of financing with an ARM. The tight band around the estimated line indicates that these predictions are quite precise. These results corroborate the findings of the LPM models and, in turn, support the hypothesis that wealthier households are more likely to select ARMs.

Robustness Testing

To verify the robustness of my results, I conducted three robustness checks: using level income rather than logged income, estimating a probit model, and adding lender fixed-effects to control for supply-side factors. For all of these tests, I used the full sample to test whether the income-ARM relationship holds across both mortgage markets. The one exception is the lender fixed-effects model, which only utilizes mortgages from Austin, Texas, due to technical limitations.

My main empirical models all use the natural log of income based on the assumption that wealth effects are proportional. To test whether this assumption drives my results, I ran an LPM model using income in levels (expressed in $10,000s). The coefficient of 0.00433 indicates that a $10,000 increase in a borrower's income is associated with a 0.43 percentage point increase in the probability that they finance with an ARM, which is broadly consistent with my log models.

The Probit model is an alternative to the LPM and logit models. The coefficient

Table 3: *Robustness Checks*

VARIABLES	(1) LPM (Level income)	(2) Probit	(3) LPM (Lender FE)
Income ($10,000s)	0.00420***		
	(9.20e-06)		
Age	0.000639***	0.00793***	7.34×10^{-5}***
	(6.40e-06)	(8.94e-05)	(2.53e-05)
Constraint Score	0.00465***	0.0835***	0.00796***
	(0.000121)	(0.00167)	(0.000433)
ln(Income)		0.690***	0.0536***
		(0.00199)	(0.000611)
Black		−0.248***	−0.0115***
		(0.00708)	(0.00158)
Constant	−0.0505***	−10.41***	−0.608***
	(0.00115)	(0.0250)	(0.00748)
Observations	4,434,854	4,435,232	318,631
R-squared	0.067	0.104	0.204

Standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

on ln(Income) is positive and highly significant. Holding all else equal, a household making $100,000 a year has a predicted probability of 2.44%, while a household making $200,000 a year has a predicted probability of 6.786%. The probit model constrains predictions between 0 and 1, addressing a potential concern with the LPM models. The quantitative conclusions of the probit model are very similar to my LPM specifications, indicating that the LPM models are robust.

Adding lender fixed effects controls for lender-specific ARM marketing strategies and underwriting standards. The income coefficient of 0.0536 is substantially larger but characteristically similar to the conforming fixed-effects model (0.0189), indicating that lender heterogeneity does not seriously confound the income and ARM choice relationship. Remarkably, the model has an R2 value of 0.204, a massive increase from 0.023 in the conforming fixed-effects model. This suggests that supply-side factors such as ARM promotion and underwriting standards explain substantial variation in ARM choice. While this does not invalidate my demand-side findings, it highlights that both borrower preferences and lender behavior shape mortgage choice patterns. The robustness checks confirm that the positive income-ARM relationship is stable across model specifications. The core finding that wealthier households prefer ARMs appears robust.

Limitations

The results are statistically and economically significant, yet there are some limitations to my model. The primary issue is that income does not equal wealth. While income is an important component of wealth, it is not a perfect indicator, especially among jumbo borrowers. Many young professionals have high incomes but haven't had the time to accumulate substantial wealth, and many retirees have almost no income but substantial accumulated wealth. This bias should attenuate my results towards zero by introducing

unexplained heterogeneity. The true effect of wealth on ARM choice is likely larger than my model predicts using income. Only including mortgages from eleven metro areas means my sample likely overstates the national ARM share and the ARM/income relationship. Rural areas have fewer jumbo loans, lower wealth concentration, and less competitive credit markets. These differences may change the magnitude of the relationship between income and ARM choice. There are also unobserved omitted variables which could be biasing my estimator. My dataset does not observe net assets, plans to move, or a metric of risk aversion, which could all plausibly be correlated with income and are know to influence ARM choice. Because of these potential biases, my results cannot show a definitive causal relationship between wealth and ARM choice, but they cleanly suggest a strong correlation.

Conclusion

The empirical results affirmed by successful robustness testing support the theoretical model's conclusion that wealthier, higher-income households are more likely to finance with ARMs than poorer, low-income households. The fixed effects LPM model indicated that a one percent increase in borrower income is associated with a 0.0189 percentage point increase in the probability they will finance with an ARM in the conforming market. In the jumbo market the same model suggests that a one percent increase in borrower income is associated with a 0.146 percentage point increase in the probability the borrower will finance with an ARM. The relationship has a greater magnitude in the jumbo market, where the ARM share is much higher, but has a larger relative effect in the conforming market. This makes sense because there is more economic diversity between households in the conforming market. Between the 25th and 75th income percentiles in the conforming market ARMs go from a high risk bargain on a poor household's most valuable asset to a low risk way for upper-middle class households to leverage and hedge their middling investment portfolios. In the jumbo market as household wealth increases the gains from ARMs increase and the risks decrease, but the way households utilize ARMs doesn't fundamentally change. While further research with more detailed wealth data is necessary to establish a definitive causal relationship, these results are qualitatively robust and carry significant policy implications.

Class and racial heterogeneity in ARM adoption implies that monetary policy pass-through has an uneven effect on mortgage rates across socioeconomic groups. Recently confirmed Fed Chair Kevin Warsh has long argued that the expansion of the Fed's balance sheet via quantitative easing has disproportionately benefited wealthy asset owners (14). The heterogeneity among ARM borrowers this paper outlines supports Warsh's argument by presenting a complementary channel: if wealthier households hold more ARMs, interest-rate cuts will reduce their payments, increasing their real wealth relative to poorer FRM holding households who don't receive the same payment relief. Warsh's quantitative easing argument and the potential for mortgage heterogeneity to affect interest rate pass through both suggest that the distributional effects of monetary policy may be broader than previously recognized.

References

[1] Aalbers, M. B. (2009). The sociology and geography of mortgage markets: Reflections on the financial crisis. *International Journal of Urban and Regional Research*, 33(2):281–290.

[2] Andersen, S., Campbell, J. Y., Cocco, J., Hansman, C., and Ramadorai, T. (2023). Heterogeneous mortgage choice: Evidence from Denmark. *American Economic Review*, 113(6):1–44.

[3] Berkovec, J. A., Kogut, D. J., and Nothaft, F. E. (2001). Determinants of the ARM share of FHA and conventional lending. *The Journal of Real Estate Finance and Economics*, 22(1):23–41.

[4] Campbell, J. Y. (2012). Mortgage market design. Working Paper 18339, National Bureau of Economic Research.

[5] Campbell, J. Y. and Cocco, J. F. (2003). Household risk management and optimal mortgage choice. *The Quarterly Journal of Economics*, 118(4):1449–1494.

[6] Chiang, Y. and Dueholm, M. (2024). Which households prefer ARMs vs. fixed-rate mortgages? Federal Reserve Bank of St. Louis, On the Economy Blog.

[7] Coulibaly, B. and Li, G. (2009). Choice of mortgage contracts: Evidence from the survey of consumer finances. *Real Estate Economics*, 37(4):659–673.

[8] Garriga, C., Kydland, F. E., and Šustek, R. (2013). Mortgages and monetary policy. *The Review of Financial Studies*, 26(5):1337–1375.

[9] Johnson, K. W. and Li, G. (2014). Are adjustable-rate mortgage borrowers borrowing constrained? *Real Estate Economics*, 42(2):457–471.

[10] Koijen, R. S. J., Van Hemert, O., and Van Nieuwerburgh, S. (2009). Mortgage timing. *Journal of Financial Economics*, 93(2):292–324.

[11] McDonald, D. and Thornton, D. L. (2008). A primer on the mortgage market and mortgage finance. *Federal Reserve Bank of St. Louis Review*, 90(1):31–45.

[12] Moench, E., Vickery, J., and Aragon, D. (2010). Why is the market share of adjustable-rate mortgages so low? Staff Report 509, Federal Reserve Bank of New York.

[13] Rothstein, R. (2017). *The Color of Law: A Forgotten History of How Our Government Segregated America*. Liveright Publishing Corporation.

[14] Wessel, D. (2015). Kevin warsh, don kohn on QE and inequality. Brookings Institution.

Medicaid Expansion and Private Health Insurance: Evidence on Crowd-Out and Discontinuities at the 138% FPL Eligibility Threshold

Dong Nguyen

ECON 381: Introduction to Econometrics (Advisor: Amy Damon)

Introduction

Human health can be conceptualized as a capital stock that depreciates over time, and individuals invest in it through medical care and lifestyle choices (14). Health insurance plays a central role in this investment by lowering the effective price of medical care and reducing financial risk, but access to coverage varies substantially with income, education, and geography (8; 9; 11). Unequal access to coverage may help explain why the United States performs poorly on life expectancy and preventable disease compared to other high-income countries, despite spending more per person on healthcare than any other nation (15).

The U.S. system combines three main forms of coverage: employer-sponsored insurance (ESI), individual-market plans, and public programs such as Medicaid and Medicare. The Affordable Care Act (ACA) reshaped this landscape in two ways that are directly relevant to this paper: it expanded Medicaid eligibility to adults with incomes up to 138 percent of the Federal Poverty Level (FPL) in adopting states, and it created subsidized Marketplace plans for those between 100 and 400 percent FPL (7; 4). A central question in the literature is whether expanding public insurance reduces private coverage rather than simply filling in gaps for the uninsured. Recent work confirms that Medicaid expansion increases public enrollment while partially crowding out private insurance, especially ESI (12).

Other work shows that eligibility rules change what type of coverage people hold, not just whether they have any. Kim et al. (10) find that the ACA individual mandate and income jointly shaped state-level private insurance enrollment, with mandate penalties making coverage decisions more responsive to income. Their evidence implies that ACA financial incentives changed which people held private insurance, not only whether the uninsured gained coverage.

The 138% FPL threshold marks the income cutoff above which people become ineligible for Medicaid in expansion states and below which they do not qualify for Marketplace subsidies, and it is the key policy boundary in this paper. Li (13) finds that individuals strategically report income near this threshold, which complicates empirical designs that exploit it as a discontinuity. Barcellos et al. (1) use the same cutoff and find a clear shift in coverage type right at the boundary: Medicaid below and subsidized private plans above, with relatively little change in overall coverage rates.

[3]Dong Nguyen is a Sophomore majoring in Economics and Statistics. Correspondence concerning this article should be addressed to tnguye14@macalester.edu.

My paper connects these two strands: the crowd-out literature and the literature on coverage responses at the 138% FPL threshold. I ask whether crossing the threshold shifts private insurance enrollment and whether that effect differs between states that expanded Medicaid and those that did not. To do this, I combine a regression discontinuity design with a difference-in-differences framework, separating the local threshold effect from the broader state-level expansion effect. What this paper adds is tracing crowd-out separately across the ESI and direct-purchase channels using all three designs on the same data. I hypothesize that Medicaid expansion crowds out private insurance primarily through the ESI channel, and that the 138% FPL threshold generates a detectable shift toward direct-purchase coverage in expansion states but not in non-expansion states.

Data

This study uses data from the Integrated Public Use Microdata Series (IPUMS) USA for the years 2000, 2011, 2016, and 2021–2024. These are repeated cross-sections, meaning each year's sample is drawn independently rather than following the same people over time. The original dataset has 22,541,822 observations; I draw a random sample of 100,000 using seed 381 to reduce computation time.

The three outcomes of interest are all binary indicators: direct-purchase private insurance, employer-sponsored insurance, and any private health insurance. Each equals 1 if coverage is reported and 0 otherwise, following standard IPUMS coding recoded from the original 1/2 format. The income distributions of those with and without direct-purchase coverage overlap considerably (Appendix Figure A1), which is one reason raw income alone does not cleanly separate the two groups.

Income enters regressions in two forms. Log income is the natural logarithm of total personal income. I use the log transformation because it compresses the long right tail of raw income and makes the income–coverage relationship more linear in the regression. The transformed distribution is unimodal but visibly left-skewed in this sample, as shown in Appendix Figure A2, which is worth keeping in mind when interpreting income coefficients. Standardized income subtracts the sample mean and divides by the standard deviation. I include this as an alternative specification because it does not require dropping people with zero income and allows comparison of effect sizes across variables measured in different units. Individuals with zero reported income are missing from log-income regressions; there are 7,656 such cases in the post-2014 sample.

The poverty ratio measures personal income as a percentage of the applicable FPL for the respondent's family size and year, and serves as the running variable in the RDD. Using the FPL ratio rather than raw income is important because the policy threshold is defined in terms of the FPL, and the FPL adjusts for family size and year. Appendix Figure A3 plots the distribution of the poverty ratio in the RDD sample.

The treatment structure follows a staggered DiD design. A treatment indicator equals one if the state ever adopted Medicaid expansion during the sample period and zero otherwise. Ten states never expanded: Alabama, Florida, Georgia, Kansas, Mississippi, South Carolina, Tennessee, Texas, Wisconsin, and Wyoming. The main treatment variable equals one when state s had implemented Medicaid expansion by year t and zero otherwise. Late-expanding states are assigned zero until their actual expansion year: Louisiana (2016),

Idaho and Nebraska (2020), Missouri and Oklahoma (2021), and North Carolina and South Dakota (2023).

All specifications control for age, sex, marital status, education group, and employment status. Marital status is recoded as a binary indicator equal to one if married and zero if not. I exclude widowed individuals because their insurance situation is often shaped by survivor benefits or a transition onto Medicare, neither of which fits cleanly into a married/unmarried split. Education is grouped into five categories: less than high school, high school, some college, bachelor's degree, and graduate or professional degree. The DiD adds state and year fixed effects, while the probit specifications include only year fixed effects. Income is excluded from the DiD because Medicaid expansion may itself shift labor supply and reported income for newly eligible individuals (7), which would make income a downstream outcome rather than an exogenous control. Including a regressor that is itself affected by the treatment introduces bad control bias and would absorb part of the policy effect we are trying to measure.

Summary Statistics

Table 1 presents summary statistics. Direct-purchase insurance coverage is 16.6 percent on average, compared to 54.1 percent for ESI and 68.8 percent for any private insurance. The average respondent is about 49 years old with log income of 10.19 (roughly $26,500). The sample is 51.3 percent female and 55.5 percent married.

Table 1: *Summary Statistics: Full Sample*

Panel A: Continuous variables				
Variable	Mean	SD	Min	Max
Log Total Income	10.188	1.290	0.0	14.1
Age (years)	48.524	19.891	15.0	97.0
Panel B: Binary variables				
Variable	% = 1			*N*
Any Private Insurance	68.8%			73,583
Direct-Purchase Insurance	16.6%			73,583
Employer-Sponsored Insurance	54.1%			73,583
Medicaid	16.0%			73,583
Medicare	27.2%			73,583
Female (= 1)	51.3%			83,349
Married (= 1)	55.5%			77,717

Notes: Sample is a 100,000-observation random draw from IPUMS USA, years 2000, 2011, 2016, and 2021–2024. Panel A reports summary statistics for continuous variables only. Panel B reports the share equal to 1 for binary variables; standard deviation, min, and max are omitted because they are determined mechanically by the proportion for any dummy variable.

Table 2 breaks down means by expansion status and time period, with binary variable means expressed as percentage points. In the pre-ACA period, expansion and non-expansion states had nearly identical direct-purchase insurance rates (15.3 pp vs. 15.0 pp).

This similarity is important because it supports the assumption that the two groups would have followed similar trends absent the policy change. ESI rates were somewhat higher in expansion states before the ACA (57.1 pp vs. 49.7 pp), likely reflecting differences in state labor markets rather than any pre-existing policy difference. After expansion, ESI fell slightly in expansion states while direct-purchase coverage rose in both groups, which suggests that broader market trends drove at least part of the post-ACA increase in individual market enrollment.

Table 2: *Means by Expansion Status and Period (Binary Variables in Percentage Points)*

	Non-Expansion States		Expansion States	
Variable	Pre-2014	Post-2014	Pre-Expansion	Post-Expansion
Any Private Insurance	63.6%	67.6%	69.3%	69.5%
Direct-Purchase Insurance	15.0%	17.9%	15.3%	16.4%
Employer-Sponsored Insurance	49.7%	50.8%	57.1%	55.1%
Medicaid	12.8%	12.6%	13.8%	17.8%
Medicare	25.4%	28.4%	22.9%	27.8%
Log Total Income	9.794	10.245	9.875	10.324
Age (years)	46.2	49.6	46.1	49.3
Female (= 1)	50.3%	51.5%	51.6%	51.2%
Married (= 1)	57.9%	55.3%	57.7%	54.5%
Education (1–5 groups)	2.619	2.862	2.704	2.941

Notes: Binary variable means expressed as percentage points. Pre-expansion covers years 2000 and 2011 for expansion states and all years before 2014 for non-expansion states. Post-expansion uses each state's actual expansion year.

Empirical Strategy

The three designs share a common economic logic. When Medicaid becomes available at no cost to newly eligible individuals, basic demand theory predicts some will switch from private insurance to Medicaid. This is the crowd-out effect. The 138% FPL threshold creates a sharp change in who qualifies for Medicaid in expansion states, so if individuals respond to eligibility, we should see a jump in coverage choices right at that cutoff. Each of the three designs uses a different source of variation to test for these effects.

Probit Models

I estimate probit models to describe how income relates to the probability of holding direct-purchase insurance. The model is:

$$\Pr(Y_{it} = 1 \mid \mathbf{X}) = \Phi(\alpha + \beta_1 \log(\text{inc}_{it}) + \beta_2\, expanded_{st} + \beta_3\, treat_s + \mathbf{X}'_{it}\boldsymbol{\gamma}) \tag{1}$$

where $\Phi(\cdot)$ is the standard normal CDF, i indexes individuals, t indexes survey years, and $\mathbf{X}$ includes age, sex, marital status, education, employment status, other insurance indicators, and year fixed effects. Standard errors are clustered at the state level. Because treatment (Medicaid expansion) is assigned at the state level, individuals within the same

state share the same policy environment and their outcomes are therefore correlated. Clustering at the state level accounts for this by allowing error terms to be correlated within states while treating states as independent.

Because Φ is a nonlinear function, the raw probit coefficients do not directly tell us how much a one-unit change in income changes the probability of coverage. I therefore report average marginal effects (AMEs) for interpretation:

$$\text{AME} = \frac{1}{n}\sum_{i=1}^{n} \phi\Big(\hat{\alpha} + \mathbf{X}_i'\hat{\boldsymbol{\beta}}\Big)\,\hat{\beta}_1, \tag{2}$$

where $\phi = \Phi'$ is the standard normal density. The AME averages the slope of the probability curve over all individuals at their actual covariate values. I also estimate an interaction specification that adds $\beta_4 \log(\text{inc}_{it}) \times \textit{expanded}_{st}$ to equation (1), testing whether the income–coverage relationship differs in expansion states. Because Φ is nonlinear, the interaction AME is computed as the difference in income AMEs between expansion and non-expansion states, not simply as the coefficient on the interaction term.

I estimate two sets of models: a simple version with no additional controls (Results Tables 4 and 5) and a full version with the complete set of controls (Results Table 6).

Difference-in-Differences

To estimate the causal effect of expansion on private insurance, I use a staggered difference-in-differences (DiD) with a linear probability model:

$$Y_{ist} = \alpha + \delta\, \textit{expanded}_{st} + \mathbf{X}_{it}'\boldsymbol{\gamma} + \mu_s + \lambda_t + \varepsilon_{ist}, \tag{3}$$

where μ_s are state fixed effects and λ_t are year fixed effects. The state fixed effects remove any permanent differences across states, such as persistently higher insurance rates or different labor market conditions. The year fixed effects remove time trends that affect all states at once. After accounting for both, what remains is how insurance coverage changed within states that expanded, relative to states that did not. That within-state change is the DiD estimate δ. Because *treat* is constant within states over time, it is absorbed by μ_s and excluded from the model. Standard errors are clustered at the state level for the same reason as in the probit.

I use the linear probability model (LPM) instead of probit for three reasons. First, adding state fixed effects to a probit creates the incidental parameters problem: probit requires estimating a separate intercept for each state, and unlike in linear models, these extra parameters cannot be removed by demeaning. Their imprecision introduces bias in all other estimates. The LPM does not have this problem. Second, LPM coefficients are directly interpretable as percentage point changes in the probability of coverage. Third, with average outcome rates well above 0 and below 1, the approximation error from using a linear model is small.

Under staggered adoption, the two-way fixed-effects estimator may be biased if the treatment effect size differs across states that adopted in different years (2; 6; 5). As a robustness check, I drop late-expanding states and compare only 2014 expanders against never-expanders.

The key identifying assumption is that insurance rates in expansion and non-expansion states would have changed at the same rate over time if no state had ever expanded Medicaid. Table 3 shows pre-period balance tests to assess whether this is plausible. For direct-purchase insurance, the main outcome in the probit analysis, the two groups look nearly identical before the ACA (difference = −0.4 pp, $p = 0.629$). ESI and any-private rates differ between groups in the pre-period, but those differences are permanent features of each state that the state fixed effects absorb. The parallel trends assumption asks only whether the trend over time would have been similar, not whether the starting levels were.

Table 3: *Pre-Period Balance Tests*

Variable	Non-Expansion	Expansion	Difference	p-value
Any Private Insurance	63.6%	69.3%	−5.7 pp	0.000
Direct-Purchase Insurance	15.0%	15.3%	−0.4 pp	0.629
Employer-Sponsored Insurance	49.7%	57.1%	−7.4 pp	0.000
Log Total Income	9.794	9.875	−0.080	0.000
Age (years)	46.2	46.1	+0.1	0.683
Married (= 1)	57.9%	57.7%	−0.2 pp	0.833
Education (1–5 groups)	2.619	2.704	−0.084	0.000

Notes: Two-sample t-tests comparing pre-ACA means (years 2000 and 2011) between expansion and non-expansion states. $p > 0.05$ means no statistically significant pre-period difference. Balance on direct-purchase insurance ($p = 0.629$) is the most relevant check for the parallel trends assumption in the DiD design.

Regression Discontinuity Design

I use a regression discontinuity design (RDD) that exploits the 138% FPL Medicaid eligibility cutoff. The sample is restricted to working-age adults (ages 19–64) in states that had already expanded Medicaid, observed in years after their expansion took effect, with poverty ratios between 50% and 500% (Appendix Figure A3). I restrict to post-expansion observations in expansion states because that is the only context where the 138% FPL threshold actually determines Medicaid eligibility; in non-expansion states or before expansion took effect, crossing the threshold has no policy consequence. The lower bound of 50% excludes individuals with implausibly low reported income, while the upper bound of 500% drops high-income households for whom the threshold is economically irrelevant; neither restriction meaningfully affects the local estimates near 138%. The estimating equation is:

$$Y_i = \alpha + \tau \cdot \mathbf{1}[poverty_i \geq 138] + f(poverty_i) + \varepsilon_i, \tag{4}$$

where τ is the RD estimate and $f(\cdot)$ is estimated separately on each side of the cutoff. Under the assumption that individuals cannot precisely control which side of the threshold they fall on, the estimate τ can be interpreted as the causal effect of crossing the threshold. I use a local linear polynomial and a triangular kernel, which gives more weight to observations close to the cutoff. The bandwidth h is chosen using the MSE-optimal selector from Calonico et al. (3), which balances the tradeoff between bias (using observations far from the cutoff)

and variance (using too few observations). Inference is based on bias-corrected confidence intervals (3).

Validity checks. I run five checks to assess whether the RDD estimates can be interpreted causally. First, a McCrary density test checks whether the density of the running variable is smooth at the threshold. If people are bunching just above 138% FPL, the assumption that assignment is as good as random is violated. Second, I repeat the RDD in non-expansion states, where the threshold should have no effect; a significant estimate there would suggest the design is picking up something other than the Medicaid policy. Third, I test placebo cutoffs at 100% and 200% FPL, where no policy change occurs. Fourth, I check whether the estimates change substantially under narrower ($h = 20$) and wider ($h = 60$) bandwidths. Fifth, the 138% FPL boundary is not a clean Medicaid cutoff alone. Individuals just above the threshold simultaneously become ineligible for Medicaid and become eligible for ACA Marketplace subsidies, since premium tax credits cover people between 100% and 400% FPL while Medicaid covers those below 138% FPL. This means the RD estimate τ captures the combined effect of losing Medicaid access and gaining subsidized Marketplace coverage, not the effect of Medicaid eligibility in isolation. This is an important limitation of the RDD regardless of the other validity checks (1).

The 138 percent FPL cutoff is also not the only program boundary in this income range. Supplemental Nutrition Assistance Program (SNAP) eligibility extends to roughly 130 percent of FPL on a gross-income basis, Special Supplemental Nutrition Program for Women, Infants, and Children (WIC) to 185 percent, and many states use 200 percent or higher for Children's Health Insurance Program (CHIP). None of these programs targets adult insurance directly, but each is correlated with household financial conditions and could in principle shift coverage decisions for households near these thresholds. To the extent that any such effect is concentrated exactly at 138 percent rather than spread across the income distribution, it would be absorbed into τ. The placebo cutoffs at 100 and 200 percent FPL in Panel B of Table 10 are insignificant, which is reassuring, but I cannot rule out that the estimate reflects several program-driven effects rather than Medicaid eligibility alone.

Results

Probit Results

Table 4 reports results using standardized income. Without controls, neither income nor expansion status is a significant predictor of direct-purchase insurance, and the interaction is also insignificant ($p = 0.75$). This is expected, since without controlling for other factors, unobserved differences across people and states dominate the estimates. The similarity of log-income distributions for those with and without direct-purchase coverage (Appendix Figure A1) is consistent with this null.

Table 5 uses log income instead. In the no-interaction model (Column 1), log income enters negatively and significantly ($\hat{\beta} = -0.019$, $p < 0.01$). Higher-income individuals tend to have access to employer-sponsored plans, so they are less likely to purchase insurance directly. The formal interaction test is not significant ($p = 0.37$), but the marginal effects in the note show a directionally consistent pattern: the effect of income is more negative in expansion states (AME $= -0.54$ pp) than in non-expansion or pre-expansion states (AME

Table 4: *Simple Probit Results: Standardized Income*

	(1) No Interaction	(2) With Interaction
Standardized Income	−0.0021	0.0011
	(0.0087)	(0.0085)
Expansion State (*treat*)	−0.0499	−0.0496
	(0.0625)	(0.0627)
Post-Expansion (*expanded*)	0.0089	
	(0.0222)	
Post-Expansion [base = Pre-Expansion]		0.0085
		(0.0223)
Post-Expansion × Std. Income		−0.0046
		(0.0142)
Constant	−1.0088***	−1.0087***
	(0.0563)	(0.0563)
Observations	73,583	73,583
Log-likelihood	−3,009,408.7	−3,009,401.9
Pseudo R^2	0.0002	0.0002
Controls	No	No

Notes: Raw probit coefficients. Robust standard errors clustered by state in parentheses. Interaction test: $\chi^2(1) = 0.11$, $p = 0.75$. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

= −0.31 pp). This is consistent with Medicaid drawing lower-income individuals out of the direct-purchase market in states that expanded.

Table 6 adds the full set of controls. Log income switches sign and becomes positive but insignificant (Column 1: $\hat{\beta} = 0.010$, $p = 0.137$), while standardized income is positive and significant (Column 2: $\hat{\beta} = 0.024$, $p < 0.01$). The sign change for log income happens because the controls now absorb the substitution between ESI and direct-purchase coverage. Once ESI status is held constant, higher income is associated with more direct-purchase enrollment; richer households without access to employer coverage seek out private plans on their own. The main finding from Table 6 is the negative income–expansion interaction in columns (3) and (4). The raw interaction coefficient on Post-Expansion × Log Income is −0.026 ($p < 0.05$), meaning the effect of income on direct-purchase coverage is significantly weaker in states where Medicaid is available. In AME terms, log income raises direct-purchase coverage by 0.19 pp on average ($p < 0.10$), but this positive relationship is smaller in expansion states, consistent with Medicaid absorbing some lower-income individuals who would otherwise have purchased directly. The standardized income interaction tells the same story at the 10 percent significance level. Among controls, ESI is the largest predictor (AME = −21.2 pp in Column 3), which makes sense because having employer coverage largely removes any reason to purchase insurance on your own. A graduate degree raises direct-purchase probability by 11.6 pp compared to less than high school, and unemployment reduces it by 3.7 pp.

Table 5: *Simple Probit Results: Log Income*

	(1) No Interaction	(2) With Interaction
Log Total Income	-0.0188^{***}	-0.0128^{*}
	(0.0054)	(0.0066)
Expansion State (*treat*)	-0.0640	-0.0629
	(0.0642)	(0.0642)
Post-Expansion (*expanded*)	0.0278	
	(0.0241)	
Post-Expansion [base = Pre-Expansion]		0.1301
		(0.1175)
Post-Expansion × Log Income		-0.0101
		(0.0112)
Constant	-0.7951^{***}	-0.8564^{***}
	(0.0800)	(0.0894)
Observations	64,435	64,435
Log-likelihood	$-2{,}675{,}436.0$	$-2{,}675{,}382.8$
Pseudo R^2	0.0005	0.0005
Controls	No	No

Notes: Raw probit coefficients. Robust standard errors clustered by state in parentheses. Interaction test: $\chi^2(1) = 0.82$, $p = 0.37$. Average marginal effects from Column (2) via `margins, dydx(log_inctot) over(expanded)`: income AME = -0.31 pp per log-point in pre-expansion states and -0.54 pp in post-expansion states. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

Difference-in-Differences

Table 7 shows the DiD results from equation (3). For any private insurance (Column 1), Medicaid expansion is associated with a statistically significant decline of 3.5 pp ($p < 0.01$). Column 3 shows a nearly identical decline for ESI ($p < 0.01$), suggesting that most of the crowd-out runs through the employer-sponsored channel rather than direct-purchase. For direct-purchase insurance (Column 2), the estimate is negative (-1.2 pp) but not statistically significant ($p = 0.30$). This null result likely reflects two offsetting forces: ACA exchange subsidies draw people into the individual market, while Medicaid expansion pulls some of those same people into public coverage instead.

Table 8 restricts to the 2014 cohort. The crowd-out estimates are slightly larger (-4.7 pp for any private, -4.6 pp for ESI), and direct-purchase remains insignificant (-1.8 pp, $p = 0.15$). The main results do not appear to be driven by staggered adoption bias.

Regression Discontinuity Design Results

Table 9 shows that the RDD estimate for direct-purchase insurance at the 138% FPL threshold is positive (3.2 pp) but statistically insignificant ($p = 0.45$, 95% CI: $[-0.063, 0.142]$). The estimates for any private insurance and ESI are also insignificant.

The wide confidence intervals are partly a result of the small number of effective observations near the threshold (only 844 to the left and 995 to the right for direct-purchase), combined with the discrete construction of the IPUMS poverty ratio, which produces gaps in the running-variable density near the cutoff (see Table 10, Panel E). The visual flatness of the running-variable conditional mean at the threshold is also evident in Appendix Figure A4.

Table 10 reports the validity checks. The McCrary density test rejects the null of no manipulation ($T = 2.25$, $p = 0.025$). There are zero observations just below 138% FPL and 69 just above it, suggesting people are either misreporting income or the data contain rounding artifacts near the threshold. When the running variable is manipulated, the assumption that assignment is as good as random is violated and $\hat{\tau}$ does not have a causal interpretation. The placebo checks in Panels A–D are all insignificant, which is reassuring, but the density failure alone is sufficient to question the design.

Even setting aside the manipulation concern, the compound treatment problem described in Section 4.3 means $\hat{\tau}$ cannot be attributed to Medicaid eligibility alone. Individuals just above 138% FPL simultaneously lose Medicaid and gain access to heavily subsidized Marketplace coverage. The near-zero positive estimate may reflect these two effects roughly canceling out rather than a true null effect of Medicaid eligibility.

Conclusion

This paper uses three empirical designs to examine how Medicaid expansion affects private health insurance enrollment. The evidence consistently points to crowd-out operating through the employer-sponsored channel, which is broadly consistent with the first hypothesis and with the findings of Lennon (12). The second hypothesis, that the 138% FPL threshold generates a detectable shift toward direct-purchase coverage, is not supported, though the RDD evidence should be interpreted with caution given the density failure.

The probit results show that the effect of income on direct-purchase insurance is more negative in expansion states, which is consistent with Medicaid drawing lower-income individuals away from the direct-purchase market. The DiD estimates are more direct: Medicaid expansion reduced both ESI and any private coverage by 3.5 pp ($p < 0.01$), while leaving direct-purchase enrollment unchanged ($p = 0.30$). The null result for direct-purchase likely reflects ACA exchange subsidies offsetting Medicaid crowd-out in the individual market. The 2014-cohort robustness check produces similar or slightly larger estimates, so the results are probably not driven by staggered adoption bias. The concentration of crowd-out in the ESI channel means that projections of ACA coverage gains should account for erosion of job-based coverage, not just changes in the uninsured rate (12).

Four limitations deserve mention. First, this analysis uses a 100,000-observation sample, and running the full dataset would likely improve precision, especially for the RDD where effective sample sizes near the threshold are small. Second, the staggered DiD has not been validated using a formal cohort-based estimator like Callaway and Sant'Anna (2). Third, income may itself be affected by Medicaid expansion through labor supply responses or changes in income reporting behavior, which would make it an inappropriate control in the probit. The standard fix is an instrumental variable for income, but that is left for future work; the probit results should therefore be treated as descriptive rather than causal.

Fourth, the RDD is not credible for causal inference in this setting for the reasons described above.

Still, the results add to what we know about crowd-out (12) and about how coverage decisions shift near the FPL threshold (1). Future work using administrative income data could sidestep the running-variable manipulation concern and, if paired with a design that separates Medicaid eligibility from Marketplace subsidy eligibility at the 138% cutoff, could separately identify the two policy effects that this paper cannot disentangle.

References

[1] Barcellos, S. H., Jacobson, M., and Levy, H. G. (2023). The impact of eligibility for Medicaid versus subsidized private health insurance on medical spending, self-reported health, and public program participation. *American Journal of Health Economics*, 9(2):262–295.

[2] Callaway, B. and Sant'Anna, P. H. C. (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics*, 225(2):200–230.

[3] Calonico, S., Cattaneo, M. D., and Titiunik, R. (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica*, 82(6):2295–2326.

[4] Eibner, C. and Saltzman, E. (2015). Assessing alternative modifications to the Affordable Care Act: Impact on individual market premiums and insurance coverage. *RAND Health Quarterly*, 4(4):4.

[5] Goin, D. E. and Riddell, C. A. (2023). Comparing two-way fixed effects and new estimators for differences-in-differences: A simulation study and empirical example. *Epidemiology*, 34(4):535–543.

[6] Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, 225(2):254–277.

[7] Gruber, J. and Sommers, B. D. (2019). The Affordable Care Act's effects on patients, providers, and the economy: What we've learned so far. *Journal of Policy Analysis and Management*, 38(4):1028–1052.

[8] Hosseini Shokouh, S. M., Arab, M., Emamgholipour, S., Rashidian, A., Montazeri, A., and Zaboli, R. (2017). Conceptual models of social determinants of health: A narrative review. *Iranian Journal of Public Health*, 46(4):435–446.

[9] Kaestner, R. and Lubotsky, D. (2016). Health insurance and income inequality. *Journal of Economic Perspectives*, 30(2):53–78.

[10] Kim, H., Cha, J., Jang, J., Han, A., Oh, D., and Lee, K.-H. (2025). The impact of individual mandate and income on private health insurance enrollment: A state-level analysis on individual behavior change. *Risk Management and Healthcare Policy*, 18:1229–1239.

[11] Lawanson, O., Abu-Halimeh, A., and Ajiferuke, O. (2025). Socioeconomic determinants of health outcomes and healthcare access in the United States: A review approach. *International Journal of Innovative Science and Research Technology*, pages 2393–2405.

[12] Lennon, C. (2025). Did the Affordable Care Act's Medicaid eligibility expansions crowd out private health insurance coverage? *Journal of Policy Analysis and Management*, 44(1):208–235.

[13] Li, M. (2025). Does Medicaid expansion lead to income adjustment? Evidence from the Survey of Income Program Participation. arXiv.

[14] Muurinen, J.-M. (1982). Demand for health: A generalised Grossman model. *Journal of Health Economics*, 1(1):5–28.

[15] Smith, S., Newhouse, J. P., and Freeland, M. S. (2009). Income, insurance, and technology: Why does health spending outpace economic growth? *Health Affairs*, 28(5):1276–1284.

Appendix

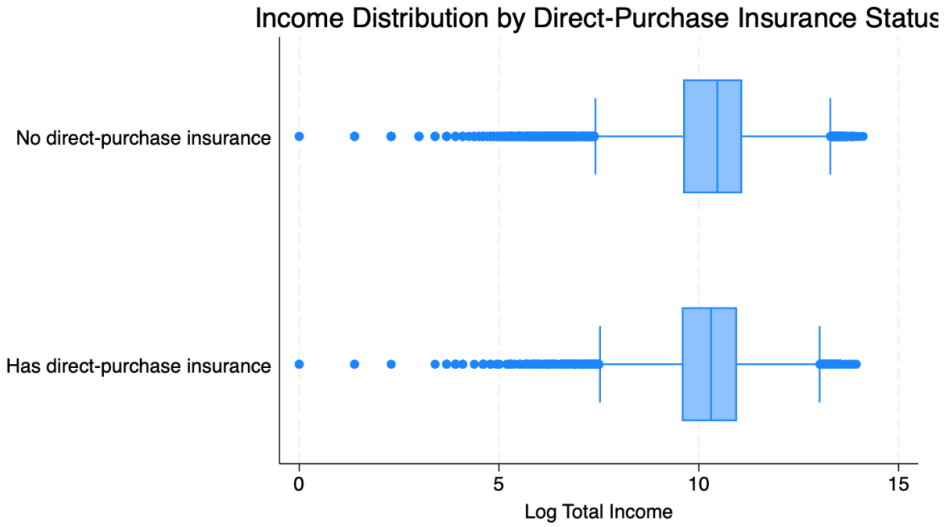


Figure 1: *Log Income Distribution by Direct-Purchase Insurance Status*

Figure 1 plots the log income distribution separately for respondents with and without direct-purchase private insurance. The two groups look strikingly similar: medians and interquartile ranges largely overlap, and both distributions show a long left tail of low-income observations. This overlap reflects the heterogeneous population that holds direct-purchase coverage as it includes self-employed and higher-income individuals who lack employer access, as well as lower-income individuals who do not qualify for public programs. The similarity also explains why the simple (no-controls) probit in Table 4 finds no significant effect of income: raw income alone does not cleanly separate direct-purchase holders from non-holders.

Figure 2 shows the kernel density of log total income for the subsample with positive income (Epanechnikov kernel, bandwidth = 0.104). The distribution has a primary peak around log(income) $\approx$ 10.5 (roughly \$36,000) and a secondary hump in the 7–9 range, reflecting the mix of full-time and part-time or lower-wage earners in the sample. The log transformation compresses the long right tail visible in raw income and makes the income–insurance relationship more linear, but the resulting distribution is left-skewed rather than symmetric, which is worth keeping in mind when interpreting the income coefficients in Tables 5 and 6.

Figure 3 plots the density of the poverty ratio for the RDD sample: working-age adults ages 19–64 in expansion states, with ratios between 50% and 500%. The dashed vertical line marks the 138% FPL Medicaid eligibility cutoff. The histogram looks roughly uniform across the range, which might seem to suggest no bunching at the threshold. However, the McCrary density test in Table 10 detects a statistically significant asymmetry: there are zero observations immediately below 138% FPL and 69 at or just above it. This asymmetry is not visually prominent because the IPUMS poverty variable is discretely constructed and the histogram bins are wide relative to the affected range, but it is sufficient

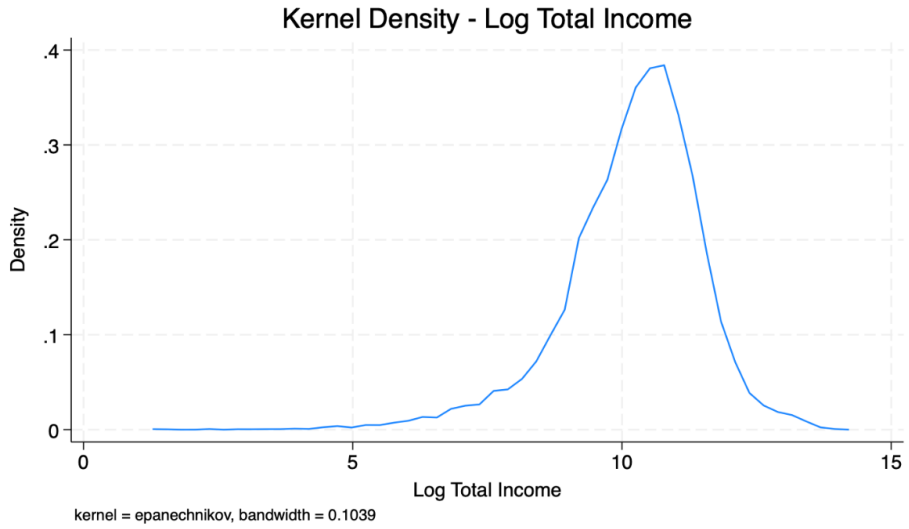


Figure 2: *Kernel Density of Log Total Income*

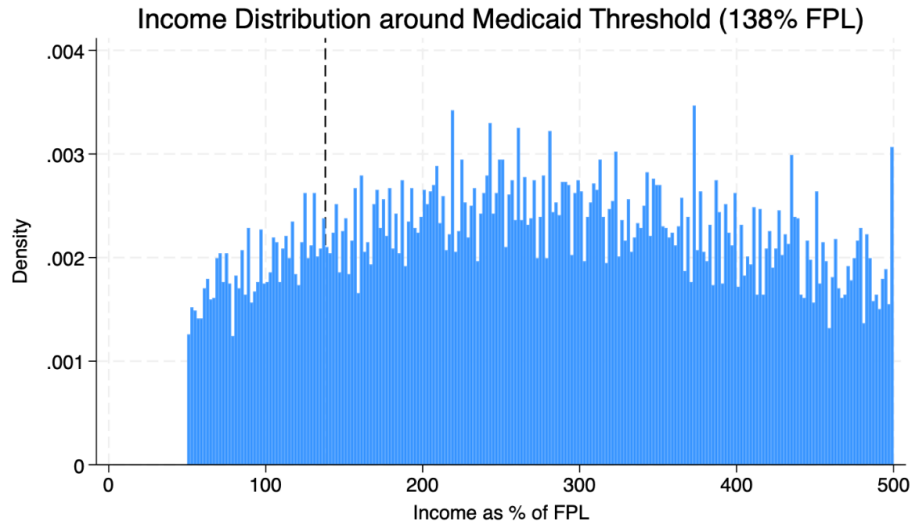


Figure 3: *Distribution of Income Relative to the 138% FPL Threshold*

to raise serious concerns about the validity of the RDD identifying assumption.

Figure 4 is the RD plot for direct-purchase insurance at the 138% FPL cutoff. The x-axis is income as a percentage of the FPL and the y-axis is the probability of direct-purchase coverage. Gray dots are sample averages within income bins; the solid curve is a fourth-order global polynomial fit included for visual smoothing only (the main estimates in Table 9 use a local linear fit). The fitted line passes smoothly through the 138% cutoff with no discernible jump, which is consistent with the statistically insignificant RD estimate of 3.2 pp ($p = 0.45$). The visual flatness also reflects the compound treatment problem discussed in Section 4.3: individuals just above the threshold lose Medicaid access while simultaneously gaining subsidized Marketplace coverage, and the two effects may roughly

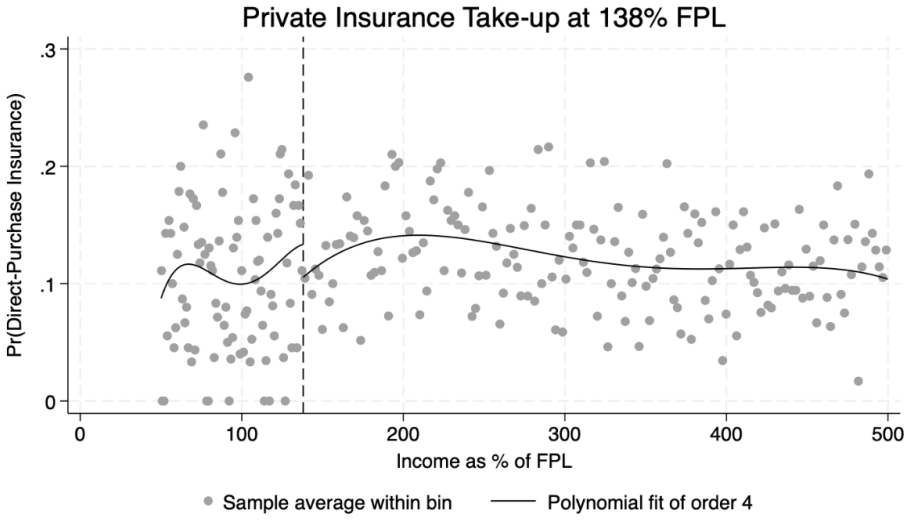


Figure 4: *RD Plot: Direct-Purchase Insurance at the 138% FPL Threshold*

offset one another even if each is individually non-trivial.

Table 6: *Multivariate Probit Results: Direct-Purchase Insurance*

	(1) Log Income (Coeff.)	(2) Std. Income (Coeff.)	(3) Log × Expansion (AME)	(4) Std. × Expansion (AME)
Income				
Log Total Income	0.0096		0.0019*	
	(0.0064)		(0.0011)	
Standardized Income		0.0243***		0.0052***
		(0.0073)		(0.0016)
Treatment				
Expansion State (*treat*)	0.0524	0.0843	0.0099	0.0160
	(0.0566)	(0.0597)	(0.0108)	(0.0113)
Post-Expansion (*expanded*)	−0.0382	−0.0618	−0.0071	−0.0118
	(0.0505)	(0.0511)	(0.0097)	(0.0098)
Demographics				
Age	0.0057***	0.0050***	0.0011***	0.0009***
	(0.0011)	(0.0011)	(0.0002)	(0.0002)
Female	0.0633***	0.0695***	0.0122***	0.0133***
	(0.0152)	(0.0136)	(0.0029)	(0.0025)
Married	0.0663**	0.0646**	0.0126**	0.0121**
	(0.0283)	(0.0264)	(0.0055)	(0.0051)
Other Insurance (ref.: none)				
Employer-Sponsored (ESI)	−1.1085***	−1.0938***	−0.2123***	−0.2072***
	(0.0305)	(0.0287)	(0.0073)	(0.0071)
Medicaid	−0.8757***	−0.9114***	−0.1681***	−0.1728***
	(0.0604)	(0.0598)	(0.0118)	(0.0117)
Medicare	0.1116**	0.1837***	0.0213**	0.0346***
	(0.0520)	(0.0496)	(0.0100)	(0.0094)
Education (ref.: less than high school)				
High School	0.5147***	0.4755***	0.0716***	0.0665***
	(0.0515)	(0.0447)	(0.0057)	(0.0051)
Some College	0.6393***	0.6095***	0.0953***	0.0917***
	(0.0533)	(0.0449)	(0.0068)	(0.0061)
Bachelor's Degree	0.6847***	0.6601***	0.1046***	0.1020***
	(0.0590)	(0.0547)	(0.0075)	(0.0074)
Graduate/Professional	0.7360***	0.6979***	0.1156***	0.1098***
	(0.0573)	(0.0497)	(0.0076)	(0.0070)
Employment Status (ref.: employed)				
Unemployed	−0.2226***	−0.1848***	−0.0374***	−0.0317***
	(0.0590)	(0.0495)	(0.0093)	(0.0081)
Not in Labor Force	0.0812**	0.0309	0.0160**	0.0062
	(0.0348)	(0.0318)	(0.0069)	(0.0062)
Observations	59,545	67,539	59,545	67,539
Log-likelihood	−2,018,904.3	−2,271,386.7	−2,018,643.6	−2,271,197.4
Pseudo R^2	0.1664	0.1638	0.1665	0.1639
Year FE	Yes	Yes	Yes	Yes

Notes: Cols. (1)–(2): raw probit coefficients from eq. (1). Cols. (3)–(4): average marginal effects from interaction models via `margins, dydx(*) post`. The interaction term does not appear directly in cols. (3)–(4) because the AME averages the income slope across all observations; raw interaction results: Post-Expansion × Log Income = −0.026** (0.013); Post-Expansion × Std. Income = −0.027* (0.015). Year FE AMEs omitted. Robust SEs clustered by state. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

Table 7: *DiD Linear Probability Model: Effect of Medicaid Expansion on Insurance Coverage*

	(1) Any Private	(2) Direct-Purchase	(3) Employer-Sponsored
Medicaid Expansion (*expanded*)	-0.0347^{***}	-0.0115	-0.0346^{***}
	(0.0116)	(0.0111)	(0.0121)
Observations	67,539	67,539	67,539
R^2	0.184	0.036	0.188
Year FE	Yes	Yes	Yes
State FE	Yes	Yes	Yes
Controls	Yes	Yes	Yes

Notes: OLS estimates of equation (3). *expanded* equals 1 if the state had implemented Medicaid expansion by that year. State FEs absorb *treat*. Controls: age, sex, marital status, education, employment status. Income excluded because it may be affected by expansion. Robust SEs clustered by state. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

Table 8: *DiD Robustness: 2014 Expansion Cohort Only*

	(1) Any Private	(2) Direct-Purchase	(3) Employer-Sponsored
Medicaid Expansion (*expanded*)	-0.0468^{***}	-0.0183	-0.0462^{***}
	(0.0145)	(0.0125)	(0.0158)
Observations	61,494	61,494	61,494
R^2	0.183	0.035	0.187
Year FE	Yes	Yes	Yes
State FE	Yes	Yes	Yes
Controls	Yes	Yes	Yes

Notes: Late-expanding states (ID, LA, MO, NE, NC, OK, SD) dropped to compare only the 2014 cohort against never-expanders. Addresses potential staggered adoption bias (2). Robust SEs clustered by state. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

Table 9: *RDD Estimates at the 138% FPL Threshold*

	(1) Direct-Purchase	(2) Any Private	(3) Employer-Sponsored
RD Estimate	0.032	0.032	−0.002
Robust Std. Err.	(0.038)	(0.046)	(0.040)
Robust p-value	0.451	0.577	0.979
95% CI	[−0.063, 0.142]	[−0.101, 0.182]	[−0.123, 0.126]
Observations (total)	15,508	15,508	15,508
Eff. obs. left of cutoff	844	1,065	1,089
Eff. obs. right of cutoff	995	1,242	1,274
MSE-optimal bandwidth h	29.95	36.16	37.92
Polynomial order	1	1	1
Kernel	Triangular	Triangular	Triangular

Notes: Sharp RDD estimates from equation (4) using `rdrobust` with triangular kernel and MSE-optimal bandwidth. Sample: working-age adults 19–64 in expansion states, poverty ratios 50–500%. Person weights applied. Inference uses bias-corrected robust confidence intervals (3). Mass-point adjustments applied automatically. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

Table 10: *RDD Validity Checks*

Check	z-statistic	p-value
Panel A: Placebo group (direct-purchase, cutoff: 138% FPL)		
Non-expansion states, RD estimate = 0.053	0.980	0.327
Panel B: Placebo cutoffs (expansion states, direct-purchase)		
Cutoff at 100% FPL, RD estimate = −0.017	−0.591	0.555
Cutoff at 200% FPL, RD estimate = −0.044	−1.531	0.126
Panel C: Placebo outcome (expansion states, cutoff: 138% FPL)		
Medicare coverage, RD estimate = 0.023	0.368	0.713
Panel D: Bandwidth sensitivity (expansion states, direct-purchase)		
Narrow bandwidth: $h = 20$, RD estimate = 0.065	0.467	0.641
Wide bandwidth: $h = 60$, RD estimate = 0.016	0.701	0.483
Panel E: Running-variable density test (McCrary)		
Robust T-statistic	2.247	0.025**

Notes: All checks use `rdrobust` with triangular kernel and MSE-optimal bandwidth unless otherwise noted. McCrary test via `rddensity`. The density test finds zero observations immediately below 138% FPL and 69 at or just above. Whether this reflects income misreporting or a construction artifact in the IPUMS poverty ratio is not separately identified in these data. All placebo checks in Panels A–D are insignificant. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

The Impact of Marijuana Legalization on Adolescent Drug Use and Educational Outcomes in the U.S.

Annie Hsu

ECON 381: Introduction to Econometrics (Advisor: Amy Damon)

Introduction

American high school students are frequently exposed to controlled substances, most notably alcohol, nicotine, and marijuana. Of all American high school 12th graders, 26% reported that they used marijuana at least once in 2025 (7). The topic of youth drug consumption is of interest to labor economists as it can impact long-term human capital. Additionally, the (de)criminalization of marijuana has been a heated debate within American politics for decades now. This intersection raises a critical question: Does the specific legal status of marijuana have a significant effect on adolescent drug usage and educational outcomes in US states?

There is both behavioral and economic relevance in examining marijuana policy and adolescent outcomes. Cannabis is the most commonly used illicit drug amongst US youth, and usage has been associated with impaired cognition, poorer academic performance, and weaker job outcomes, suggesting potential long-run effects on human capital formation and labor-market productivity (8). Criminalization has resulted in youth incarceration and criminal records that restrict access to higher education, housing, and employment, creating a direct link between policy and lifetime economic opportunity (8). This shows how variation in marijuana criminalization may influence youth behavior and downstream labor-market outcomes. In a 2017 study from the Pediatrics Journal, researchers find that trend-wise, marijuana prevalence did not increase overall during the past decade. Despite a declining perceived risk of marijuana use, overall prevalence did not increase over the past decade, largely because a reduction in cigarette smoking offset rising usage within certain subgroups (11). Patterns of usage remain associated with beliefs about harm, implying that policy effects likely operate through expectations and social norms rather than direct price or access channels alone (11). However, this research relies on analysis of precedent drug/alcohol usage, not directly on environmental factors (legalization). Taken together, this literature suggests that marijuana criminalization may influence youth usage rates through behavioral and institutional mechanisms with implications for human capital accumulation.

The existing literature regarding how these policies impact adolescent marijuana usage presents a varied and often contrasting set of outcomes. Miech et al. (11) find that the prevalence of marijuana use among adolescent groups has increased from 2005 to 2016. In a 2020 study conducted by the American Journal of Preventive Medicine, multilevel modeling of Seattle youth shows that nonmedical marijuana legalization showed a higher likelihood of self-reported marijuana usage in 2019 (3). Additionally, a 2023 study on

[3]Insert

middle and high school students in Oregon, from the Journal of Health Economics, asserts that states with legal recreational marijuana have higher rates of adolescent use, as well as higher rates of substance abuse offences in school (5). They only find an increase in usage of 0.14 per 100 students in Oregon middle schools, and find a correlation between offenses and distance to a marijuana outlet (5). Regarding medical marijuana, Smart and Doremus (15) find that youth marijuana usage responds to medical marijuana market size. They find that if a state where marijuana is illegal were to reach the median medical marijuana registration rate, 12–17 year old marijuana usage would increase by 8% in the past month. The researchers suggest this increase in youth usage is due to decreases in price and social markets. However, they acknowledge that their study finds null effects on youth outcomes (15). These varying findings highlight the incongruent outcomes of current research on adolescent marijuana usage.

However, conflicting evidence suggests legalization may not substantially increase youth marijuana use. In a 2023 study conducted by researchers at the Journal of Economic Literature, Anderson and Rees (2) find that researchers have produced little credible evidence to suggest that medical marijuana legalization promotes use amongst teenagers, suggesting delineations between the type of marijuana being legalized. To show a more specific breakdown of youth usage by legal status, Zvonarev et al. (16) find that in states where nonmedical marijuana is legal, adolescent usage has remained the same or declined, showing that legalization has little to no effect on youth usage rates. This finding is consistent with a 2022 study from the University of Minnesota, where a survey of Minnesota high school students revealed that fewer than one in five high school students ever used cannabis, and that there has been an overall decline over time in usage (12).

Even if adolescent usage may be decreasing, there is still concern about its potential effects, especially on educational outcomes. Homel et al. (9) find that, compared to abstainers of marijuana usage, frequent users have the lowest high school grades and are least likely to enroll in post-secondary education, indicating a negative impact on both education and educational outcomes. Similarly, Huwit (10) finds that the risk of early marijuana exposure is correlated with decreased distance to a marijuana dispensary as well as lower AP and SAT scores on average. The study also asserts that all subjects were significantly negatively impacted by higher marijuana exposure, not just subjects who use cognitive-based abilities. On the other hand, Blake et al. (4) examine how a permissive public policy change for recreational marijuana influences college applications and find a 54% increase. Additionally, the first and third quartiles of admitted students do not see a decrease in SAT scores. Dave et al. (6) also show a positive relationship between recreational marijuana laws and labor market outcomes, finding that there is little evidence that recreational marijuana laws negatively affect labor market outcomes amongst working-age individuals. These studies outline the varying research that exists on the possible effect that marijuana legalization has on adolescent educational outcomes.

Criminalization is not necessarily the better alternative for youth. From a 2015 study from the Pediatrics Journal, evidence at the time was against the punishment of youth who use marijuana, and it was suggested that the decriminalization of simple possession for minors and adults may prove to be a better alternative, especially if it is coupled with drug education (1). On the other hand, Salehin and Pillai (13) not only find that states with restrictive/stronger marijuana laws had lower prevalence of adolescent usage,

but also find an inverse relationship with marijuana usage and youth activities, suggesting a strategy to decrease marijuana usage. Therefore, this shows that there is evidence to support alternatives in criminalizing the usage of marijuana for adolescents.

While prior literature has examined adolescent marijuana usage trends and educational outcomes, findings remain mixed regarding the causal role of marijuana legal status. However, existing research often focuses on localized policies or specific subpopulations, leaving a gap in comprehensive, state-level comparative analyses. Consequently, the broad causal role of varying state legal statuses remains unclear. Additionally, although research has linked marijuana use to educational outcomes, fewer studies examine how state-level legalization policies influence both adolescent usage rates and educational outcomes simultaneously. My study will aim to contribute to the literature by examining how differences in marijuana legal status across U.S. states are associated with variation in adolescent marijuana usage and educational outcomes. By utilizing a robust two-way fixed effects framework, this paper demonstrates that the apparent positive relationship between legalization and youth drug use is primarily driven by baseline state-level differences rather than a true causal policy effect.

Data Source

To investigate the effect of state recreational/medical marijuana legalization policies on adolescent marijuana usage, I used data from the National Survey of Drug Usage and Health (NSDUH) from the Substance Abuse and Mental Health Services Administration (SAMHSA) (14), as well as compiled data for both the legalization of medical marijuana and the legalization of recreational marijuana from each state. The NSDUH compiles data for a limited number of substance use and mental health measures. They are generally based on two years of combined data and cover each state and the District of Columbia. Most state estimates are based on a small area estimation (SAE) methodology, in which state-level NSDUH data are combined with county and sub-county level census data from the state.

Within the NSDUH adolescent marijuana usage data, there is a data gap from 2020 to 2021, due to the COVID-19 pandemic. Therefore, this causes an unbalanced panel and creates the possibility that national trends during COVID may have affected both usage and legalization patterns. Additionally, this data is measured in thousands (of adolescent marijuana users), not percentages, meaning larger states will mechanically have more users.

Data Description

The dependent variable I am observing is adolescent usage rates, or the number of 12–17 year-olds who use marijuana, measured in thousands. The independent variable is the level of legality, represented by a fully legal variable and a medically legal variable. The null hypothesis is that there is no statistically significant evidence that the criminalization of marijuana affects adolescent usage rates or educational outcomes. The alternative hypothesis is that there is statistically significant evidence that the legality level affects the adolescent marijuana usage rate and educational outcomes.

In total, there are 400 observations across 50 states, with Washington D.C., covering the time period 2015 to 2024 without 2020–2021 due to lack of data during the pandemic.

Table 1: *Summary Statistics*

VARIABLES	Description	N	Mean	Std. Dev.	Min	Max
12–17 Year Old Usage	Adolescent marijuana usage (ages 12–17), measured in thousands per state	400	60.42	68.34	—	—
Legalized Recreation	Recreational marijuana legalization status (0 = not legalized, 1 = legalized)	400	—	—	0	1
Legalized Medical	Medical marijuana legalization status (0 = not legalized, 1 = legalized)	400	—	—	0	1
High School Degree or Higher	Percentage of individuals with a high school diploma or higher	400	89.60	3.00	—	—
Year	Observation year (2015–2024, excluding 2020–2021)	400	—	—	2015	2024

There are 50 observations per year, and per state there are 8 observations. The variables are as follows: *state* covers all 50 American states plus Washington, D.C. The *year* variable spans 2015 to 2024 as available from the NSDUH, omitting 2020 and 2021. Adolescent marijuana usage is the amount of marijuana usage for 12 to 17-year-olds, measured in thousands, per US state. Legalized Recreational Marijuana is the year in which recreational marijuana was legalized within a given state: 0 means not legalized and 1 means legalized. For simplicity, legalization means the year in which the possession of cannabis (not just CBD oil/THC adjacent products) is legalized, so the year a legalization measure goes into effect is the year it is marked 1. Legalized Medical Marijuana acts the same as the previous variable but measures when a state has legalized medical marijuana. High School Degree or Higher is the percent of individuals within a given state who have achieved a high school diploma or higher from 2015 to 2024.

I conducted a t-test to compare marijuana usage in states with versus without recreational legalization. States with recreational legalization had a mean of 73.7 thousand adolescent users compared to 55.6 thousand in non-legal states, a difference of 18.1 thousand ($t = -2.34$, $p = 0.020$). The standard deviations are very large (57 and 92, respectively), meaning there is enormous variation within each group. This suggests state population size is driving a lot of the difference; larger states will naturally have more users in raw numbers regardless of legal status. However, the null hypothesis (legality does not affect adolescent usage) can be rejected.

To account for unobserved heterogeneity across states and over time, my analysis incorporates both state and year fixed effects into the regression framework. State fixed effects control for all time-invariant characteristics that differ across states but do not change within a state over the study period, such as political culture, geographic factors, baseline attitudes toward drug use, and long-standing demographic differences.

Without state fixed effects, the estimated relationship between legalization and adolescent marijuana use can simply reflect the fact that states that chose to legalize marijuana were already systematically different from states that did not—states with a more permissive culture, are more urban, or have higher baseline usage rates—rather than capturing any true causal effect of the policy change itself. Year fixed effects control for national trends that affect all states simultaneously in a given year, such as shifts in federal drug policy attitudes, national media coverage of marijuana, or broader macroeconomic conditions that may influence adolescent behavior across the board. Together, these two sets of fixed effects mean that identification of the legalization effect comes exclusively from within-state variation over time, comparing a state's adolescent marijuana use before and after it legalized, relative to states that did not change their legal status during the same period. This is a considerably more conservative empirical strategy than a simple cross-sectional comparison and directly addresses the concern that baseline OLS estimates may overstate the effect of legalization by confounding policy changes with pre-existing state-level differences.

Empirical Analysis

To measure the relationship, I estimate the following regression:

$$\text{Use}_{st} = \beta_0 + \beta_1 \text{Legalized Recreational}_{st} + \beta_2 \text{Legalized Medical}_{st} + \beta_3 \text{Education}_{st} + \alpha_s + \gamma_t + u_{st} \quad (1)$$

The subscript s indexes states, and the subscript t indexes years. Use_{st} is the estimated number of adolescents (measured in thousands) aged 12–17 who used marijuana in the past year t per state s. This is the primary outcome of interest for the overall prevalence of adolescent marijuana usage at the state level. $\text{Legalized Recreational}_{st}$ is the primary variable of interest. $\text{Legalized Medical}_{st}$ is a variable for an intermediate level of legal access to marijuana and may indicate reduced stigma amongst non-patients. Education_{st} is the percentage of adults in the state within a year with a high school diploma or higher, loosely representative of the human capital level of the state and controlling for the possibility that more educated states have both lower drug use and different legalization patterns. α_s represents the state fixed effects, capturing unobserved confounders across states so that identification comes from within-state variation over time. γ_t is the year fixed effect, representing confounding or unobservable variables such as shifting federal attitudes or macroeconomic conditions. u_{st} is the error term.

The key parameter of interest is β_1, which measures the effect of recreational legalization on adolescent marijuana use. I assume that adolescent marijuana use rates do not themselves cause states to legalize marijuana.

The tests I run are as follows: a baseline OLS with no fixed effect to see the most basic correlation, a state fixed effect, a year fixed effect, and then a full model with both state and year fixed effects. The baseline results may show pre-existing state differences rather than a true legalization effect.

Standard economic theory predicts that consumption of any good is a decreasing function of its full price, where full price includes both the monetary cost and the non-monetary costs of obtaining a good. Before legalization, social stigma and search costs that adolescents may face in trying to obtain marijuana may lessen post-legalization. Therefore, I expect that legalization will be associated with increased marijuana use, including amongst adolescents, as the full price of consumption falls. Additionally, the inclusion of educational attainment is a control variable motivated by human capital theory. Higher educational attainment is associated with greater awareness of health risks and higher opportunity costs of risky behavior, all of which can be predictors of lower substance use.

Regression Results

All of the following coefficients are measured in thousands. Table 2 shows all four model specifications side by side: moving from left to right, each column progressively adds controls for state and year fixed effects. The OLS shows there is a positive effect on both recreational marijuana (26.97) and medical (19.63) legalization on adolescent usage, both $p < 0.01$. Additionally, the negative effect on the percentage of high school degree acquisition is -12.24 and is also statistically significant, which indicates that a one percentage point increase in high school graduation rates is associated with approximately 12.2 thousand fewer adolescent users, holding legalization constant.

Table 2: *Regression Results*

VARIABLES	(1) Baseline OLS	(2) State FE	(3) Year FE	(4) State + Year FE
Legalized Recreational	26.970***	−1.494	23.928***	0.678
	(7.370)	(1.896)	(7.644)	(1.796)
Legalized Medical	19.625***	2.521	18.917***	1.987
	(6.774)	(2.383)	(6.858)	(1.684)
High school degree or higher	−12.236***	−1.748***	−12.847***	2.501**
	(1.016)	(0.529)	(1.071)	(1.166)
Constant	1,136.484***	187.956***	1,185.523***	−157.785
	(90.911)	(45.232)	(94.961)	(102.683)
Observations	400	400	400	400
R-squared	0.292	0.984	0.299	0.107
State FE	No	Yes	No	Yes
Year FE	No	No	Yes	Yes
Number of states		50		50

Standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

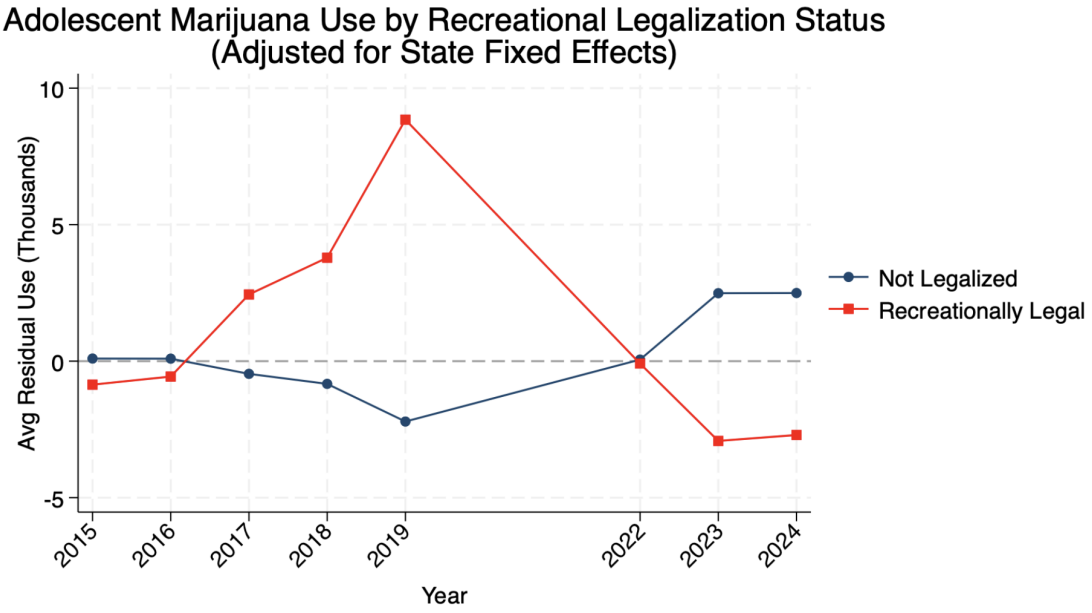
Column two introduces state fixed effects, changing the results. The coefficient of recreational legalization decreases to −1.494 and loses all statistical significance, while medical legalization also shrinks to 2.521 and becomes insignificant. The reduction indicates that once time-invariant state-level characteristics are controlled for, there is no statistically significant evidence that legalization increases adolescent marijuana use within states over time. The R-squared increased to 0.984, showing that the state FE absorbs a large share of variation in adolescent usage.

Column three adds year fixed effects without state fixed effects, showing results similar to baseline (recreational legality is 23.93 and medical legality is 18.92, both statistically significant). This shows that controlling for national time trends alone is insufficient to remove the bias in the legalization variables, as they do not address the problem that legalizing states may be fundamentally different from non-legalizing states to begin with.

Column four presents both state and year fixed effects, the most conservative estimate of the effect of legalization. Recreational legalization has a coefficient of 0.678 and medical has a coefficient of 1.987, both statistically insignificant. Education becomes positive at 2.5, suggesting multicollinearity between educational trends and year FE. Overall, the strong positive correlation between legalization and adolescent marijuana use may be entirely due to pre-existing differences between states, and once those differences are controlled for, legalization shows no statistically significant causal effect on adolescent consumption.

Figure 1 confirms the above findings. It plots the average entity-demeaned residuals from the two-way fixed effects model separately for states with and without recreational marijuana legalization from 2015 to 2024. The two lines both trace closely around zero across the period, which is exactly what is to be expected if recreational legalization has no systematic within-state effect on adolescent use. The spike in the red line in 2019 and its subsequent collapse in 2022–2024 is noteworthy, as the recreationally legal states actually show higher residual use in 2019 before dropping below zero in 2023 and 2024, suggesting a slight decline in within-state use in more recent years among legalizing states.

Figure 1: *Average entity-demeaned residuals for the two-way fixed effects model*



Meanwhile, the non-legalized states are slightly positive in 2023–2024. The fact that neither line shows a clear, sustained divergence from zero or from each other across the study period visually corroborates the regression finding that recreational legalization has no statistically significant causal effect on adolescent marijuana consumption once state-level heterogeneity is removed. The within-state residual trends for medical legalization follow a similar pattern, with both legal and non-legal states tracing closely near zero without any noticeable divergence, further supporting the regression finding that medical legalization has no significant causal effect on adolescent use once state fixed effects are applied (see Appendix Figure 2).

Robustness Checks

Table 7 shows the robustness checks I employed to determine if my findings hold under alternative specifications. The first column (1) represents results when states that legalized early are excluded. These states are Colorado, Washington, Alaska, and Oregon. The second column (2) is the same as the first, except with the inclusion of fixed effects. The third column (3) and the fourth (4) address the fact that the adolescent usage data is measured in thousands, not percentages, so it allows for coefficients to be interpreted as approximate percentage changes rather than changes in thousands of users.

Column one was included to check if early-legalizing states may be driving the results, as they may be outliers in terms of usage levels or their long history of legal marijuana. The exclusion of these states slightly increases the recreational legalization coefficient to 31.29 thousand, compared to the previous baseline of 26.97 thousand, suggesting that early legalizing states were not inflating the baseline estimate. Column two matches the main

Table 3: *Robustness Checks*

	(1)	(2)	(3)	(4)
VARIABLES	Excl. Early States	Excl. Early States FE	Log Outcome OLS	Log Outcome FE
Legalized Recreational	31.287***	0.487	0.381***	0.042
	(11.757)	(1.953)	(0.110)	(0.027)
Legalized Medical	19.668***	2.024	0.151	0.003
	(5.729)	(1.762)	(0.101)	(0.026)
High school degree or higher	−12.052***	2.448**	−0.181***	0.090***
	(1.813)	(1.199)	(0.015)	(0.023)
Constant	1,120.005***	−151.680	19.609***	−4.280**
	(162.575)	(105.359)	(1.360)	(2.033)
Observations	368	368	400	400
R-squared	0.294	0.091	0.273	0.279
State FE	No	Yes	No	Yes
Year FE	No	Yes	No	Yes
Number of states	46	46	50	50

Robust standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

results, as once state and year fixed effects are added, the recreational legalization falls to 0.487 and becomes completely insignificant, as does medical legalization with its decrease to 2.024. Column three shows that recreational legalization is associated with a 46.4% higher adolescent marijuana use (from the log coefficient of 0.381, converted via $\exp(0.381) - 1$ for interpretation), larger than the medical effect of 15%. The educational coefficient shows an 18% reduction in adolescent marijuana use, consistent with the main results. Column four then applies the log within the full state and year fixed effect. Recreational and medical lose all significance, but education remains significant at the one percent level.

The four robustness checks reinforce the conclusions drawn from the main analysis. Regardless of whether early legalizing states are excluded, the pattern remains the same: significant legalization effects disappear once state fixed effects are introduced.

Comparison to Existing Literature

These findings are broadly consistent with the large body of literature that doubts the narrative that marijuana legalization causally increases adolescent use. Most notably, Anderson and Rees (2), writing in the Journal of Economic Literature, conclude that researchers have produced little credible evidence to suggest that medical marijuana legalization promotes use amongst teenagers, a conclusion that the present study extends to recreational marijuana as well. Similarly, Zvonarev et al. (16) find that in states where nonmedical marijuana is legal, adolescent usage has remained the same or declined, suggesting that legalization has little to no direct effect on youth usage rates. The full fixed effects model, which shows recreational and medical legalization coefficients of 0.678 and 1.987, both statistically insignificant, aligns with these conclusions. At the same time, these results are at odds with those of Bailey et al. (3) and Cil et al. (5), both of which find positive associations between recreational legalization and adolescent use in specific state-level contexts such as Oregon and California. However, those studies rely on more localized samples and may be capturing state-level dynamics rather than a generalizable national pattern. Taken together, the weight of evidence—including the present study—suggests that the relationship between legalization and adolescent use is far more nuanced than a

simple positive causal story, and that pre-existing state characteristics play a role in driving the observed cross-sectional correlation.

Limitations

Several limitations exist within this analysis. First, the dependent variable is measured in raw thousands of users rather than as a percentage of the adolescent population, meaning that larger states have higher usage counts regardless of their legalization status or cultural attitudes toward marijuana. A population-adjusted percentage measure would allow for cleaner cross-state comparisons and more precise coefficient interpretation.

Second, the absence of NSDUH data for 2020 and 2021 due to COVID-19 creates a gap in the panel that may obscure important trends, as the pandemic coincided with significant shifts in adolescent behavior, school attendance, and substance use patterns nationally.

Third, this analysis is subject to potential omitted variable bias. While state and year fixed effects absorb a large share of unobserved heterogeneity, they cannot account for time-varying confounders that differ across states within the study period. Variables such as dispensary proximity, state-level enforcement intensity, public health campaign expenditures targeting youth drug use, changes in the price of marijuana in both legal and illegal markets, and broader socioeconomic shifts at the state level are all plausibly correlated with both legalization timing and adolescent consumption but are not included in the model. To the extent that these omitted variables are correlated with the legalization indicators, the estimated coefficients on recreational and medical legalization may still reflect confounding influences rather than clean causal effects, even in the preferred two-way fixed effects specification.

Fourth, the functional form of the regression may be misspecified. This analysis assumes a linear relationship between legalization status and adolescent marijuana use, but the true relationship may be nonlinear. For example, the effect of legalization on adolescent use may be concentrated in the first few years following policy change and diminish thereafter as novelty wears off, or it may grow over time as retail markets mature and dispensary access expands. The use of binary indicators for recreational and medical legalization also collapses substantial variation in implementation—how quickly retail markets opened, how aggressively age verification was enforced—into a single zero-one variable, potentially obscuring meaningful heterogeneity in the treatment itself.

Fifth, there is a possibility of reverse causality introducing endogeneity into the legalization estimates. This analysis assumes that adolescent marijuana use rates do not themselves cause states to legalize, but this assumption may not be fully defensible. States with higher baseline rates of marijuana use among all age groups, including adolescents, may face stronger political pressure from advocacy groups to legalize, or conversely, may face stronger opposition from public health advocates concerned about youth exposure. If pre-existing usage rates influence the timing of legalization, the direction of causality runs in both directions simultaneously, making it difficult to isolate the effect of the policy from the underlying usage environment that shaped the political decision to legalize in the first place. While the two-way fixed effects framework mitigates this concern by exploiting within-state variation over time, it cannot fully eliminate the possibility of reverse causality in the

absence of a credible instrumental variable that predicts legalization timing independently of adolescent consumption trends.

Sixth, the analysis is subject to potential measurement error in both the dependent and independent variables. The NSDUH data is self-reported survey-based data, and respondents, particularly adolescents, may not be fully honest in their responses due to social desirability bias—the tendency to underreport behaviors perceived as socially unacceptable or illegal. In the context of marijuana use, which remains illegal for minors in all states regardless of adult legalization status, adolescents may be reluctant to disclose their consumption even under confidentiality assurances. This underreporting problem may not be constant across states or over time, as states where marijuana has been legalized for adult recreational use may show reduced social stigma, making adolescents more willing to report use honestly, while adolescents in fully illegal states may face stronger pressure to conceal their behavior. If this differential reporting exists, observed differences in usage rates between legalizing and non-legalizing states may partially reflect changes in reporting honesty rather than true changes in consumption behavior, introducing a systematic bias into the estimated legalization coefficients that cannot be corrected without external validation data. Additionally, the legalization timing variable captures the year a law went into effect but does not account for variation in implementation across states, such as how quickly retail markets opened or how aggressively age verification was enforced, which may matter for adolescent access and actual consumption independent of the legal status itself.

Conclusion

This paper examines whether state-level marijuana legalization, for both recreational and medical use, has a statistically significant effect on adolescent marijuana use among 12- to 17-year-olds across all 50 states, from 2015 to 2024. Using data from the National Survey on Drug Use and Health and a two-way fixed effects framework, the analysis finds that the strong positive relationship between legalization and adolescent use observed in the baseline OLS disappears almost entirely once state fixed effects are introduced. Recreational legalization falls from a coefficient of 26.97 in the baseline to 0.678 in the full state and year fixed effects model. Neither recreational nor medical legalization remains statistically significant in the preferred specification. These results are robust to excluding early-legalizing states and to log-transforming the outcome variable, which shows that the findings are not driven by outliers or the raw thousands measurement of the dependent variable. However, the most consistent finding across all specifications is that higher state-level educational attainment is negatively associated with adolescent marijuana use in the baseline and state fixed effects specifications, suggesting that promoting increases in education may serve as a more reliable policy lever for reducing youth marijuana consumption than the legal status of marijuana itself. These findings align with Anderson and Rees (2) and Zvonarev et al. (16) in suggesting that legalization does not causally increase adolescent marijuana use and that decriminalization does not come at the cost of increased youth consumption. Future research would benefit from population-adjusted usage measures and longer pre-treatment windows to more rigorously test the parallel trends assumption underlying the fixed effects design.

Appendix

Figure 2: *Average entity-demeaned residuals for medical marijuana legalization*

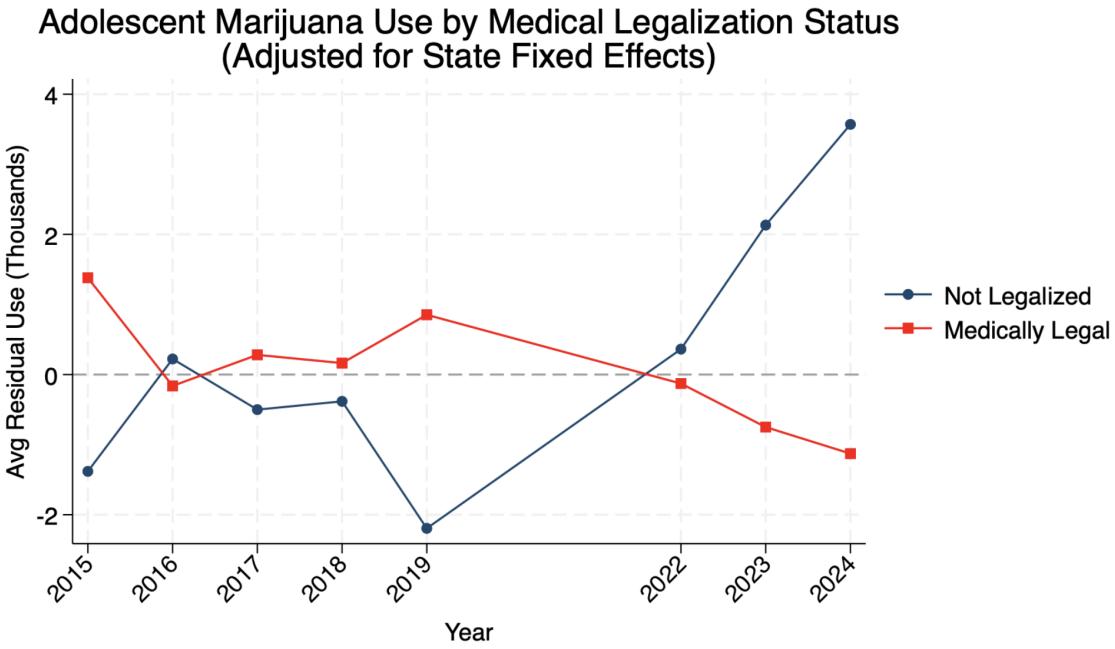


Figure 2 plots the average entity-demeaned residuals from the two-way fixed effects model separately for states with and without medical marijuana legalization from 2015 to 2024. Values on the vertical axis represent deviations from each state's mean adolescent marijuana use after removing state and year fixed effects, such that the graph isolates within-state variation over time rather than cross-sectional differences between states. This pattern is consistent with the regression results presented in Table 2, which show that medical legalization coefficients collapse to small and statistically insignificant values once state fixed effects are introduced.

References

[1] Ammerman, S., Ryan, S., and Adelman, W. P. (2015). The impact of marijuana policies on youth: clinical, research, and legal update. *Pediatrics*, 135(3):769–785.

[2] Anderson, D. M. and Rees, D. I. (2023). The public health effects of legalizing marijuana. *Journal of Economic Literature*, 61(1):86–143.

[3] Bailey, J. A., Epstein, M., Roscoe, J. N., Oesterle, S., Kosterman, R., and Hill, K. G. (2020). Marijuana legalization and youth marijuana, alcohol, and cigarette use and norms. *American Journal of Preventive Medicine*, 59(3):309–316.

[4] Blake, C. D., Thomas, D. K., and Hess, J. (2023). Higher education: The impact of recreational marijuana on college applications. *Contemporary Economic Policy*, 42(2):259–277.

[5] Cil, G., Winters, K. C., Austin, S. C., Kittelman, A., Smolkowski, K., Westling, E., and Seeley, J. R. (2023). Legalization and retail availability of recreational marijuana and adolescent use in schools. *Health Economics*, 33(1):107–120.

[6] Dave, D. M., Liang, Y., Muratoni, C., and Sabia, J. J. (2025). The effects of recreational marijuana legalization on employment and earnings. *Journal of Population Economics*, 38(47).

[7] Drug Abuse Statistics (2025). Marijuana addiction statistics [2026]: Usage & abuse rates. `https://drugabusestatistics.org/marijuana-addiction/`. Retrieved February 26, 2026.

[8] Hammond, C. J., Chaney, A., Hendrickson, B., and Sharma, P. (2020). Cannabis use among U.S. adolescents in the era of marijuana legalization: a review of changing use patterns, comorbidity, and health correlates. *International Review of Psychiatry*, 32(3):221–234.

[9] Homel, J., Thompson, K., and Leadbeater, B. (2014). Trajectories of marijuana use in youth ages 15–25: Implications for postsecondary education experiences. *Journal of Studies on Alcohol and Drugs*, 75(4):674–683.

[10] Huwit, N. (2023). *Three Essays on the Effects of Marijuana Legalization on Health, Education, and Social Security Disability Insurance*. PhD thesis, American University.

[11] Miech, R., Johnston, L., and O'Malley, P. M. (2017). Prevalence and attitudes regarding marijuana use among adolescents over the past decade. *Pediatrics*, 140(6).

[12] Planalp, C., Hest, R., and Stewart, A. (2024). Cannabis use and perceptions among minnesota high school students. Technical report, SHADAC.

[13] Salehin, M. and Pillai, V. K. (2025). Influence of socio-ecological and economic correlates on marijuana legalization policy across the states of america. *International Journal of Environmental Research and Public Health*, 22(6):823.

[14] SAMHSA (2025). NSDUH national releases. `https://www.samhsa.gov/data/data-we-collect/nsduh-national-survey-drug-use-and-health/national-releases`. Retrieved February 26, 2026.

[15] Smart, R. and Doremus, J. (2023). The kids aren't alright: The effects of medical marijuana market size on adolescents. *Journal of Health Economics*, 87.

[16] Zvonarev, V., Fatuki, T. A., and Tregubenko, P. (2019). The public health concerns of marijuana legalization: An overview of current trends. *Cureus Journal of Medical Science*, 11(9).

How Do Severe Winter Cold Shocks Impact Midwest Electricity Systems: First Through Varied Demand Responses Across Income Levels, and Then Through Supply-Side Rebalancing via Wind and Other Electricity Generation Sources in Minnesota?

Jerry Zheng

ECON 381: Introduction to Econometrics (Advisor: Amy Damon)

Introduction

Imagine a colony on Mars: people move through enclosed spaces, the outdoors is hostile, and survival depends on reliable energy. This is the reality for Minnesotans in the winter, when on some days, being outside with exposed skin meant severe frostbite. In Minnesota, the electrical grid, which is responsible for enabling heating across homes, is a matter of survival. This paper studies the effects of winter cold shocks on the electricity system through two perspectives: first, how does the electricity demand response vary by household income; and second, how does wind energy availability change during these severe-cold days, and how does that affect electricity supply when all other generation sources are taken into consideration. During winter cold shocks, demand for electricity rises. After examining the details of rising demand, it is crucial to ask whether supply can keep up. If wind availability increases during cold shocks, the grid may exhibit a self-correcting mechanism. If wind energy generation decreases during cold shocks, the problem is more serious, especially because Minnesota is heavily reliant on wind energy. This paper contributes to the literature by examining both the demand and supply sides on a overlapping set of severe-cold days in a heavily wind-reliant grid, which lets us directly assess whether the Minnesota system absorbs cold shocks through self-correction or through stress on conventional generation.

Literature Review

Studies have shown that electricity systems are affected by weather through both the demand and the supply side, although the two are often studied separately (14; 20). On the demand side, many agree that electricity does not respond linearly to temperature, and this non-linear relationship becomes especially important in extreme conditions (4; 13; 15**?**). Instead of a single linear slope, Bessec and Fouquau (4) used a threshold panel approach, while Lee and Chiu (13) used a panel smooth transition regression to estimate a U-shaped relationship with a threshold around 53°F. More recent work also suggests that demand may be more sensitive to cold-wave conditions than to ordinary variations in temperature in winter (15), reinforcing the idea that cold shocks should be modeled separately. Cold

[3]Jerry Zheng is a Sophomore majoring in Economics. Correspondence concerning this article should be addressed to jzheng@macalester.edu.

weather also raises electricity use more sharply in places where households rely on electric heating, with Berkouwer (3) finding a strong response below the heating threshold and tying it directly to the prevalence of electric heating.

More specifically, household demand responses are not uniform, because income, appliance hold- ings, housing structure, and heating technology all shape how weather affects electricity use (1; 6; 7; 19; 24). Doremus et al. (6) emphasize that extreme weather creates unequal energy burdens across income groups rather than a single common household response. Ohler et al. (19) further explains these differences through the composition of households, with appliance stock, water heating, and space heating reliance all varying systematically with income. Related work on residential energy demand emphasizes the importance of technology choice, as energy use depends on the heating systems and the appliances chosen by the household (1; 7; 24). This literature also implies that the differences in area income should not be treated as the sole X variable; instead, income is best understood as a proxy for underlying differences in heating technology, appliance stock, and household adjustment margins during severe cold.

Looking at the supply side, we have to ask whether severe winter conditions alter wind generation and force systems to rebalance among conventional generation methods and trade margins (14; 20). Studies of winter weather regimes show that high demand cold conditions can correlate with weaker wind availability, but the relationship is not mechanically causal in one direction because different weather patterns produce different wind outcomes (14; 20). Cold conditions may also affect wind generation more directly through performance losses such as reductions in turbine efficiency due to icing (10). Related literature shows that wind intermittency and other operating constraints shape how much wind displaces coal and gas, and these substitution patterns vary across markets and fuel price conditions (5; 8; 12). Together, these papers imply that the supply consequences of severe cold days are likely to be state dependent, operating through meteorology, turbine performance, dispatch, transmission, and trade simultaneously (2; 9).

Although the literature provides strong evidence on nonlinear temperature demand relationships, household heterogeneity based on income, and wind related system adjustment, these strands are rarely integrated in a single cold region framework. Papers have not examined how severe cold days affect the demand and supply side simultaneously, with the central mechanism being wind power, which accounts for 25% of Minnesota's electricity generation in 2023 (22). Demand papers tend to focus on either average temperature effects or household level energy outcomes, while supply papers focus on wind displacement, intermittency, or trade without centering on severe cold shocks as the underlying disturbance (2; 6; 12; 19). This leaves an opening for my paper, which asks whether severe cold in a Midwest setting simultaneously changes electricity demand, changes the slope of the temperature-demand relationship, and alters the contribution of wind relative to gas, coal, and imports.

To address these gaps, this paper tests four hypotheses. The first two are on the demand side: whether severe cold significantly raises hourly electricity demand (H1), and whether that response varies systematically with median household income (H2). The next two are on the supply side: whether wind availability changes during severe cold events (H3), and whether higher wind generation reduces natural gas, coal, and net imports (H4). The paper finds that severe cold significantly raises hourly demand and that the demand

response grows monotonically with income; on the supply side, wind generation actually falls during the coldest hours, while each MWh of wind displaces roughly 0.49 MWh of gas, 0.32 MWh of coal, and 0.13 MWh of net imports. The full statements with dependent and independent variables are given in the Empirical Strategy section.

Theory

Demand Mechanism

Households produce warmth using electricity, gas, and insulation. When temperature drops below a heating threshold (∼55–60°F), marginal heating demand jumps, creating the non-linear demand curve seen in Bessec and Fouquau (4) and Lee and Chiu (13). Severe cold of lower than 0°F should push households across a second threshold where backup electric heating, space heaters, and plug load start to activate even in gas-heated homes. This explains why cold-wave elasticity exceeds ordinary winter elasticity (15).

Higher-income households have larger square footage, more appliances, and remain home in extreme cold instead of reducing comfort. This leads to their demand response being mechanically larger per °F. On the contrary, lower-income households face a more limiting budget constraint and a tradeoff of comfort against cost, so the observed demand response should be lower than the true heating need—important for interpreting $\beta_2 < 0$.

Supply Mechanism

The Midcontinent Independent System Operator, responsible for energy generation across a dozen states in the Midwest, dispatches following a merit order: wind (∼$0 marginal cost) clears first, then nuclear, coal, and gas, with imports filling the remainder. One MWh of new wind displaces the most expensive marginal unit, which is gas in MISO-North in winter. Total generation must equal load every hour, so coefficients on wind across (gas + coal + imports + other) must sum to close to −1. There are also two contradicting mechanisms for wind in subzero conditions. While subzero temperatures can theoretically increase air density and produce stronger winds, blade icing reduces aerodynamic efficiency and can trigger automatic shutdown, cutting output 10–30% even when wind speed is adequate (10).

If wind generation persists during cold shocks, the grid absorbs the shock through wind itself rather than through conventional ramp-up; if wind collapses instead, the grid has to lean on gas, coal, and imports.

Empirical Strategy

Part A: Demand Response to Severe-Cold Days

Estimate how severe-cold exposure changes electricity demand and whether that response differs by area income.

H1: Severe cold significantly increases hourly electricity demand.

- Y: Hourly electricity demand (18)
- X: Hourly temperature (11)

H2: The severe-cold demand response varies systematically with median income.

- Y: Hourly electricity demand (18)
- X: Hourly temperature (11); median income (National Institute on Minority Health and Health Disparities)

To find out how mean income impacts responses to extreme cold days, we implement a panel regression that accounts for differences between the nine climate divisions of Minnesota. The threshold for an extreme cold day, in a given climate division, is a day where at least one hour in that division recorded a temperature below 0°F. The full regression is:

$$Y_{it} = \alpha_i + \eta_h + \tau_d + \beta_1 T_{it} + \beta_2 (T_{it} \times \text{Income}_i) + \beta_3 W_{it} + \beta_4 (T_{it} \times \text{HeatMix}_i) + \varepsilon_{it} \quad (1)$$

where the variables are defined as follows:

- i: Climate division index for Minnesota, $i = 1, \ldots, 9$.
- t: Time index by the hour.
- Y_{it}: Total electricity usage (MWh) for climate division i at time t.
- α_i: Division fixed effect. Captures all time-invariant characteristics of division i, such as baseline demand level and average infrastructure differences.
- η_h: Hour-of-day fixed effect. Captures systematic demand patterns within the day, such as differences in usage during daytime versus evening hours.
- τ_d: Date fixed effect, measured daily. Captures day-specific shocks such as holidays.
- T_{it}: Average temperature (°F) for division i at time t.
- $T_{it} \times \text{Income}_i$: A continuous-by-continuous interaction term measuring whether the effect of temperature on electricity usage differs by mean income level (in thousands of dollars) across divisions. This is the main coefficient of interest.
- W_{it}: Average wind speed (mph) for division i at time t.
- HeatMix_i: Heating mix—the share of households in division i that use electricity as their primary heating fuel.
- $T_{it} \times \text{HeatMix}_i$: Interaction term measuring how the division-level heating mix moderates the average electricity response to a cold shock.
- ε_{it}: Error term.

The regression aims to isolate the causal effect of income on the demand response to rapid temperature decreases. Income and heating mix enter only as interaction terms because their direct effects on electricity usage are absorbed by the division fixed effect. Wind speed and heating mix are included as controls because they vary systematically

across divisions and are the variables most likely to confound the temperature–demand relationship. The remaining fixed effects account for time-varying common shocks.

For the panel regression to deliver a causal estimate of the temperature–demand relationship, three assumptions must hold. First, temperature variation conditional on the fixed effects must be exogenous to electricity demand shocks—a reasonable assumption because weather is not chosen by households or by the grid. Second, the temperature–demand relationship must be approximately linear within the extreme-cold sample, which the histogram in Figure 2 supports given that the daily minimums deciding sample selection cluster in a relatively narrow range below the 0°F cutoff. Reverse causality is impossible, as aggregate electricity demand cannot affect outdoor temperature. Even so, electricity usage per division depends on many factors not captured by the fixed effects and controls included here, so the result is best interpreted as the causal effect under these assumptions rather than an unconditional claim.

The six columns of Table 3 walk from a misleading cross-sectional estimate to the preferred panel specification, and the sign of the temperature coefficient tells most of the story. Column 1 is a pooled OLS on temperature, income, and their interaction; the temperature coefficient is 0.186 MWh/°F and positive, which would imply that warmer hours have slightly higher demand. This is the wrong sign because the pooled regression conflates the cold demand response with structural differences between divisions: more urban, higher-income divisions have both higher baseline demand and steeper cold responses, and pooled OLS picks up the cross-section. Column 2 adds wind speed but does not fix the identification problem. The key shift is in Column 3, where division fixed effects absorb time-constant divisional differences and the temperature coefficient flips to -1.186 MWh/°F. Column 4 layers in hour-of-day fixed effects to absorb the daily demand cycle, which would otherwise be confounded with cold hours that tend to fall during the morning and evening peaks. Column 5 adds date fixed effects to absorb day-specific shocks such as holidays, market events, and storms unrelated to temperature, and is the preferred specification carried through the rest of the paper. Column 6 layers a $T_{it} \times \text{HeatMix}_i$ interaction onto Column 5 as a robustness check; the interaction is not separately significant and the $T_{it} \times \text{Income}_i$ coefficient barely moves, suggesting that the electric-heating share is already being absorbed through the income measure.

Part B: Wind Displacement During Extreme-Cold Days

Estimate how wind availability affects gas generation, electricity imports, and coal usage in winter.

H3: Wind availability increases during severe cold events.

Y: Hourly MISO wind generation (23)

X: Hourly Minnesota temperature (11), division-averaged

H4: Higher wind energy generation reduces natural gas generation, net imports, or coal usage.

Y: MISO gas, coal, and net imports generation (23)

X: MISO wind generation (23)

Given that Part A shows electricity demand spikes during severe cold days in Minnesota, the next question is whether wind is actually displacing conventional generation on those same days. The generation data in this part comes from MISO, not just Minnesota. MISO covers 15 states, so the dependent variables (gas, coal, net imports) are MISO-wide hourly generation, while the shock variable remains Minnesota temperature. This remains defensible for three reasons. First, Minnesota is one of the largest wind producers in MISO, so a swing in Minnesota wind appears clearly in the MISO aggregate. Second, extreme cold events in Minnesota are not isolated—the weather patterns that produce below-zero temperatures in Minnesota typically cover the whole upper Midwest, making Minnesota temperature a reasonable proxy for regional cold. Third, MISO operates as a single market, so any displacement by wind occurs at the MISO level regardless of which state the wind is generated in.

The main specification is:

$$y_{ht} = \alpha_h + \delta_{\text{dow}} + \gamma_m + \theta_y + \beta_1 \,\text{wind}_{ht} + \beta_2 \,\text{load}_{ht} + \beta_3 \,\text{temp}_{ht} + u_{ht} \tag{2}$$

where the variables are defined as follows:

y_{ht}: MISO hourly generation (MWh) from source y (gas, coal, or net imports), estimated separately for each source.

wind_{ht}: MISO hourly wind generation (MWh). Main variable of interest.

load_{ht}: MISO hourly load (MWh). Included as a control; since generation must equal load every hour, omitting load biases the wind coefficient toward zero.

temp_{ht}: Minnesota temperature (°F) at hour h of date t. Only extreme-cold days are considered.

α_h, δ_{dow}, γ_m, θ_y: Hour-of-day, day-of-week, month, and year fixed effects. Absorb regular intraday and seasonal patterns.

u_{ht}: Error term, clustered at the daily level.

I also estimate a version of Equation 2 that includes a $\text{wind}_{ht} \times \text{temp}_{ht}$ interaction term, in order to test whether the marginal displacement effect of wind varies with temperature within the extreme-cold sample. In all regression tables for Part B, temperature is reported in Fahrenheit and demeaned at the sample mean for easier interpretation; this shifts the intercept but does not affect the temperature slope.

One limitation is that the fixed effects are coarse and do not absorb day-level shocks such as gas prices, pipeline congestion, or transmission constraints. Daily fixed effects were considered but would absorb too much of the variation needed to identify the displacement effect.

Data Description

This paper aligns hourly Minnesota electricity demand, weather, area income, and heating mix at the NOAA/NCEI climate division level, which has nine divisions in the state (NOAA National Centers for Environmental Information). The Part A panel uses Minnesota-specific demand, weather, income, and heating mix data assembled at the climate division and hourly level. The Part B panel uses MISO-wide hourly generation merged with the same Minnesota temperature series.

Electricity Demand

County-level hourly demand comes from the Hourly Electricity Demand Profiles for the Contiguous United States dataset, distributed through the Open Energy Data Initiative (18). The source contains hourly demand in megawatts for every county in the lower 48 from 2016 through 2023. I extract the Minnesota counties using Python and aggregate to climate divisions using a county-to-division crosswalk taken directly from NOAA documentation, giving 70,128 hourly observations per county over the full sample.

Temperature and Wind Speed

Station-level hourly temperature in Fahrenheit and wind speed in miles per hour come from the Iowa Environmental Mesonet's Minnesota ASOS network (11). Using QGIS, I spatially join each station to one of the nine NOAA climate divisions based on its latitude and longitude, then average the station-level series into hourly division-level temperature and wind speed. The spatial join is shown in Figure 1, which plots the climate division boundaries together with the ASOS station locations within each. The result is roughly 70,000 hourly observations per division over 2016–2023.

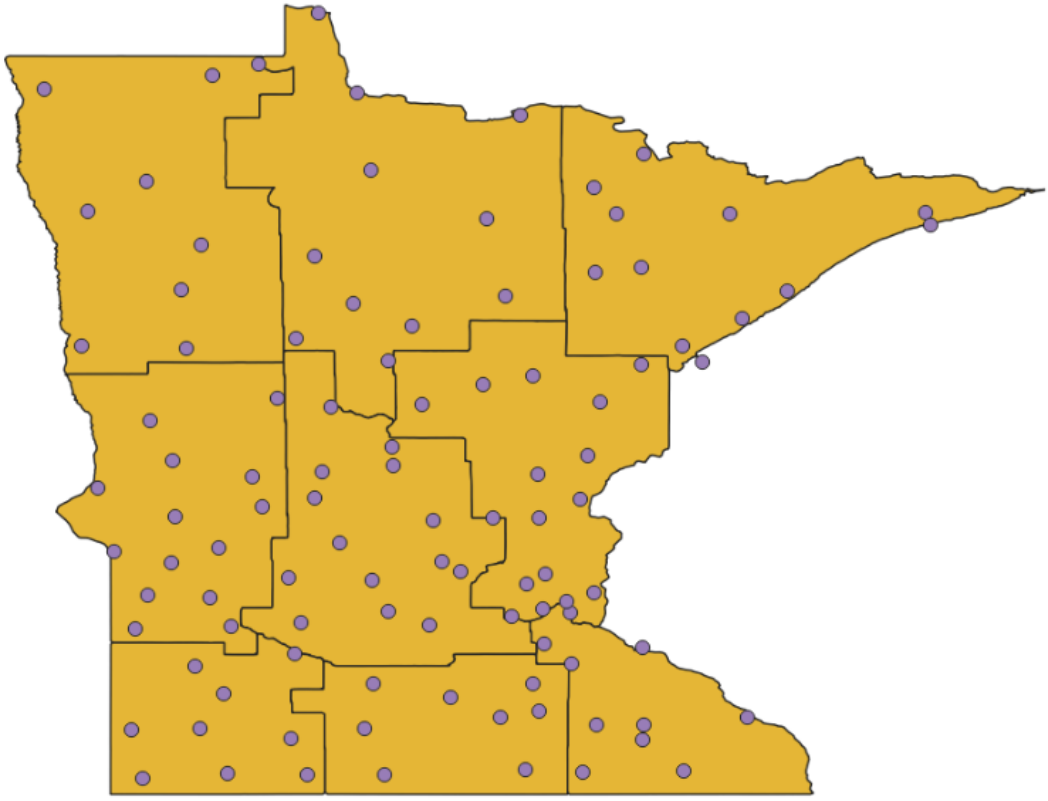
Income and Heating Mix

County-level median household income comes from the NIMHD HDPulse Data Portal as a 2019– 2023 five-year average (National Institute on Minority Health and Health Disparities). The electric-heating share comes from the U.S. Census Bureau's American Community Survey 5-year estimates, Table B25040, House heating fuel (21), defined as the share of occupied housing units reporting electricity as their primary heating fuel. Both are aggregated from county to climate division using the same NOAA crosswalk and treated as time invariant in the panel since each is an average from multiple years.

MISO Generation Data

The Part B regression uses hourly generation data for the MISO region from the EIA Real-Time Operating Grid, Midwest region (REG-MIDW), via the EIA Grid Monitor (23). For each hour, I extract gas, coal, wind, total load, and net imports in MWh. The temperature variable for Part B is the Part A division-level series collapsed to one Minnesota-wide hourly temperature. The Grid Monitor series begins later than the NREL demand data, so Part B covers a shorter window than Part A but still spans multiple full winters.

Figure 1: *Minnesota's nine NOAA climate divisions and the ASOS station locations used to construct the division-level temperature and wind speed series. Stations are spatially joined to divisions using QGIS, and the station-level hourly observations are then averaged within each division.*



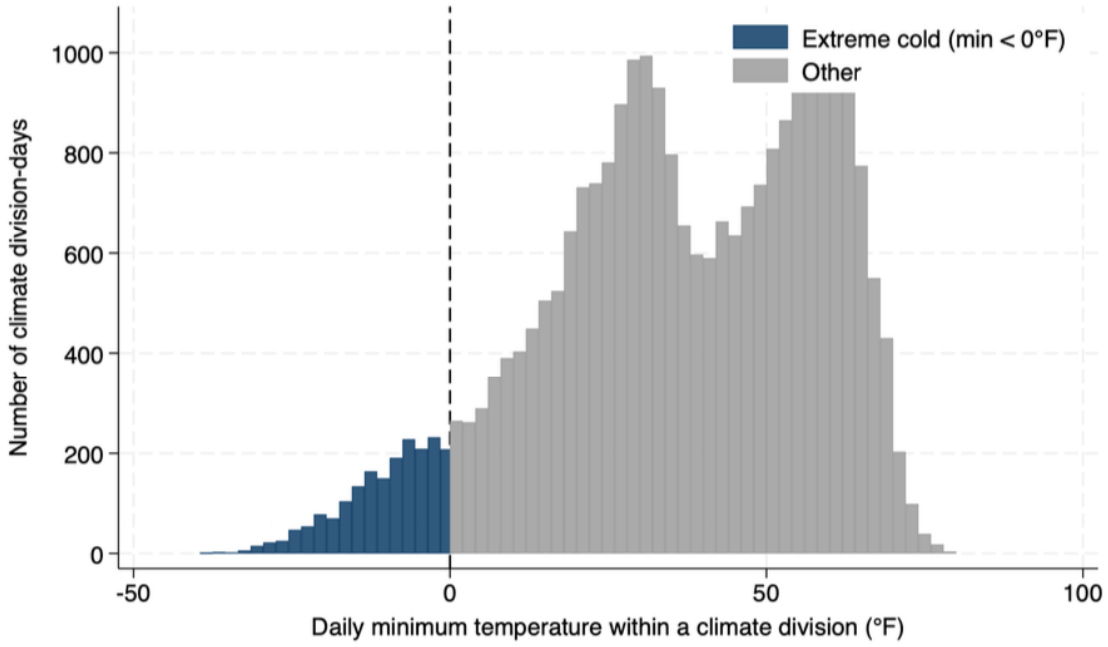
Extreme Cold Day Filter

Both Part A and Part B are restricted to extreme-cold days, defined as any day where at least one hour records a temperature below 0°F. All 24 hours of a qualifying day are retained. For Part A, this threshold is applied per climate division, so the qualifying dates differ across divisions. For Part B, it is applied to the Minnesota-wide hourly temperature series. Figure 2 shows the daily-minimum temperature distribution within climate divisions, with the 0°F cutoff marked, and the navy bars correspond to the climate division-days that enter the Part A regression.

Final Analysis Sample

After applying the filter, the Part A panel has 46,628 climate-division-hour observations across nine divisions and 2016–2023, used in Results Tables 3 and 4. The Part B panel has 5,832 MISO-hour observations (5,827 for net imports, due to a small number of

Figure 2: *Distribution of daily minimum temperature within Minnesota climate divisions, 2016–2023. Each observation is one climate division-day; the daily minimum is taken across the 24 hours within that division. The dashed vertical line at 0°F marks the cutoff used to define an extreme-cold climate division-day. Bars to the left of the cutoff (shaded navy) are the climate division-days that enter the Part A regression sample.*



missing values in the source), used in Results Tables 5 and 6. Summary statistics for both samples are reported in Tables 1 and 2, respectively.

A few patterns from Tables 1 and 2 are worth flagging before the regression results. From Table 1, hourly demand within a climate division-hour averages 878 MWh and ranges up to 5,065 MWh, a 5.8× spread that the division fixed effects in Part A absorb, and median income across the nine divisions ranges from $67k to $95k, supporting the income-interaction strategy in H2. From Table 2, the MISO wind generation mean (9,793 MWh) coincides exactly with the constant term in Table 5, a property of the mean-centered specification, and net imports occasionally run negative, indicating that the MISO region exports power on some hours within the sample.

Results

Part A: Extreme-Cold Demand Response by Income

Income is centered as:

$$\text{Income}_i^c = \text{Income}_i - \overline{\text{Income}} \tag{3}$$

so that the coefficient on temperature represents the effect of temperature on electricity

Table 1: *Summary Statistics for Part A Regression Sample (Extreme-Cold Climate Division-Days)*

	N	Mean	SD	Min	Max
Hourly demand (MWh)	46,628	878.20	1,182.27	130.51	5,065.17
Temperature (°F)	46,628	−0.95	10.61	−39.51	39.94
Wind speed (mph)	46,628	7.98	5.33	0.00	35.01
Median income ($1,000s)	46,628	77.18	10.70	67.26	95.17
Electric heat share	46,628	0.21	0.04	0.15	0.29

Sample restricted to climate division-hours where the within-division daily minimum temperature falls below 0°F, the extreme-cold filter used in Part A. Each observation is one climate division-hour, with 46,628 observations across nine climate divisions and the 2016–2023 period. Hourly demand is the division total in MWh; temperature is the division hourly average; wind speed is the division hourly average in miles per hour; median income and electric heat share are time-invariant at the division level.

Table 2: *Summary Statistics for Part B Regression Sample (MISO Hours on Extreme-Cold Days)*

	N	Mean	SD	Min	Max
Gas generation (MWh)	5,832	22,197.36	5,700.72	6,188.00	38,486.00
Coal generation (MWh)	5,832	28,059.13	6,942.66	11,110.00	43,316.00
Wind generation (MWh)	5,832	9,792.88	5,506.48	307.00	23,354.00
Load (MWh)	5,832	76,392.72	7,420.34	53,884.00	101,533.00
Net imports (MWh)	5,827	4,165.73	2,563.63	−4,497.00	12,837.00
Minnesota temperature (°F)	5,832	2.07	10.18	−30.88	35.41

Sample is the MISO hourly panel restricted to extreme-cold days as defined in Section 5.7. Each observation is one MISO hour, with 5,832 observations spanning the EIA Grid Monitor coverage window. Gas, coal, wind, load, and net imports are MISO-wide hourly generation totals in MWh from the EIA Grid Monitor (REG-MIDW). Minnesota temperature is the hourly average across the nine climate divisions, in Fahrenheit, taken from the Part A temperature panel. The slightly smaller N for net imports reflects a small number of missing values in the source data.

usage at mean income. Without this adjustment, the temperature coefficient would measure the effect when income equals zero, which is not a meaningful quantity.

Part B: Demand Response to Severe-Cold Days

Interpretation

From Column 6 of Table 3, every one-degree drop in temperature significantly increases electricity usage, with each degree of cooling leading to an average increase of 1.3 MWh in hourly electricity usage, holding all else constant. This confirms the baseline hypothesis that a cold shock generates a measurable electricity demand response. Column 6 is discussed here to address the heating-mix interaction, but Column 5 is the preferred specification used elsewhere in the paper, including the marginal effects in Table 4 and the headline numbers in the Conclusion.

The heating-mix control has little effect on the $T_{it} \times \text{Income}_i$ coefficient, likely because heating mix—the choice between electric and gas heating—is already closely correlated with income. Additionally, the scope of this paper is to measure total electricity usage during cold shocks, which encompasses all appliance-related consumption, not only the electric heating used by a subset of households.

To assess the effect of income on the electricity demand response to a cold shock, we examine the marginal effect of temperature on electricity usage evaluated at different income levels:

$$\frac{\partial Y_{it}}{\partial T_{it}} = \beta_1 + \beta_2(\text{Income}_i - \overline{\text{Income}}) \tag{4}$$

Using Column 5, $\beta_1 = -1.147$ and $\beta_2 = -0.058$, with income deviations taken relative to the sample mean.

Table 3: *Extreme-Cold Days: Climate Division Panel Regressions (Income Centered)*

	(1) *Baseline*	(2) *+ Wind*	(3) *+ Div FE*	(4) *+ Hour FE*	(5) *+ Date FE*	(6) *+ Heat Mix*
Temperature (°F) at mean income	0.186	2.435***	−1.186***	−2.598**	−1.147***	−1.307**
	(0.383)	(0.389)	(0.291)	(0.897)	(0.313)	(0.463)
Temp × Income centered ($1k)	0.147***	0.128**	−0.064**	−0.084**	−0.058**	−0.057**
	(0.053)	(0.053)	(0.027)	(0.036)	(0.022)	(0.022)
Wind speed		−15.871***	1.640	0.998*	0.513*	0.506
		(0.582)	(0.898)	(0.533)	(0.276)	(0.275)
Temp × Electric heat share						0.749
						(1.225)
Observations	46,628	46,628	46,628	46,628	46,628	46,628
R-squared	0.563	0.568	0.018	0.229	0.343	0.343
Within R-squared			0.018	0.229	0.343	0.343

Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 4 shows that, although standard errors are large, income has a substantial effect on the cold-shock demand response. The results are consistent with the hypothesis: higher-income divisions exhibit stronger responses to cold, as households have greater resources to remain indoors and maintain heating.

Table 4: *Marginal Effect of Temperature at Different Income Levels (Column 5)*

	b	**se**
Mean income − $20k	0.020	(0.390)
Mean income − $10k	−0.563*	(0.277)
Mean income	−1.147***	(0.313)
Mean income + $10k	−1.730***	(0.466)
Mean income + $20k	−2.314***	(0.658)

Each entry is the estimated effect of a $1°F$ *warmer hour on electricity usage. Negative values imply that colder temperatures are associated with higher usage.* * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Part C: Wind Displacement During Extreme-Cold Days

Interpretation

From the H3 regression in Table 5, temperature has a positive and statistically significant relationship with wind generation within the extreme-cold sample. A 1°F warmer hour is associated with approximately 83 additional MWh of wind generation, holding fixed effects constant. This implies that wind generation is actually lower during the coldest hours, not higher, so H3 is not supported in its original form. Instead, the result suggests that severe cold may suppress wind generation during the most system-stressed hours.

Table 5: *Part B — H3: Temperature and Wind Generation During Extreme-Cold Days*

	Wind generation
Temperature (°F)	83.223***
	(23.830)
Constant	9,792.878***
	(252.844)
Observations	5,832
Adj. R^2	0.244

Standard errors in parentheses. Dependent variable is hourly MISO wind generation (MWh). Temperature is in °F and mean-centered. All columns control for hour-of-day, day-of-week, month, and year fixed effects. Standard errors clustered at the daily level. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

The main specification in Table 6 (Columns 1–3) shows that each additional MWh of wind generation displaces 0.486 MWh of gas, 0.320 MWh of coal, and 0.134 MWh of net imports, all significant at the 1% level. The three coefficients sum to −0.940, close to the −1 implied by system balancing, with the remaining 6% absorbed by other generation sources (nuclear, hydro, biomass). Conversely, lower wind output is associated with greater reliance on gas, coal, and net imports during extreme-cold hours.

Turning to the other coefficients in Table 6, the load coefficient is large and positive for gas (0.571) and coal (0.549), meaning both sources ramp up with demand, and small and negative for net imports (−0.093), meaning rising demand is met primarily through in-region generation rather than imports. The temperature coefficient is positive and significant for gas (79.3 MWh per 1°F warmer), negative for net imports (−63.9), and insignificant for coal. The positive gas and negative net-imports coefficients together imply that the coldest hours within an extreme-cold event stress gas plants through cold-weather operational constraints and fuel-supply limits, forcing MISO to draw on neighboring grids, with gas output recovering and imports declining as temperatures rise within the same event.

Diagnostics and Robustness

To assess regression validity, I run two diagnostics: variance inflation factors for multicollinearity, and joint significance F-tests on the fixed effects. VIFs for the substantive Part A regressors (temperature, income, wind speed) and Part B regressors (wind generation, load, temperature) are all close to 1, with mean VIFs of 1.06 and 1.03 respectively, indicating that multicollinearity is not a concern. Joint significance F-tests confirm

Table 6: *Part B — Wind Displacement During Extreme-Cold Days*

	Main specification			*With wind × temp interaction*		
	Gas	**Coal**	**Net imp.**	**Gas**	**Coal**	**Net imp.**
Wind generation	−0.486***	−0.320***	−0.134***	−0.488***	−0.316***	−0.134***
	(0.026)	(0.030)	(0.017)	(0.026)	(0.030)	(0.017)
Load	0.571***	0.549***	−0.093***	0.570***	0.552***	−0.093***
	(0.025)	(0.025)	(0.020)	(0.025)	(0.025)	(0.020)
Temperature (°F)	79.303***	−21.060	−63.951***	67.724***	5.926	−65.587***
	(14.037)	(15.501)	(10.550)	(23.447)	(28.697)	(17.116)
Wind × Temperature				0.001	−0.003	0.000
				(0.002)	(0.002)	(0.001)
Constant	−16,682.013***	−10,759.264***	12,598.495***	−16,591.733***	−10,969.668***	12,611.243***
	(1,867.971)	(1,940.050)	(1,539.370)	(1,887.026)	(1,954.149)	(1,552.337)
Observations	5,832	5,832	5,827	5,832	5,832	5,827
Adj. R^2	0.811	0.831	0.521	0.811	0.831	0.521

Standard errors in parentheses. Dependent variables are hourly MISO generation in MWh. Wind generation and load are also in MWh; temperature is in °F and mean-centered. All columns control for hour-of-day, day-of-week, month, and year fixed effects. Standard errors clustered at the daily level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

that the fixed effects belong in each regression. For Part B, all four sets are jointly significant at $p < 0.001$: year effects are the most powerful ($F(5, 5790) = 2371.55$), followed by month ($F(4, 5790) = 415.38$), hour-of-day ($F(23, 5790) = 20.86$), and day-of-week ($F(6, 5790) = 6.68$). The Part A hour and date fixed effects are also jointly significant at $p < 0.001$.

For the robustness check, I re-estimate the Column 5 specification on the three northernmost climate divisions (2101, 2102, 2103) only, to verify that the income result is not driven by pooling colder northern divisions with milder southern ones. Table 7 shows that the temperature effect at mean income shrinks from −1.147 to −0.812, but the income interaction remains significant at −0.033 MWh per °F per $1,000 of additional median income. The within R^2 also rises from 0.343 to 0.897 once cross-division heterogeneity is removed, confirming that the income gradient is real and persists within the colder, more cold-day-dense northern third of the state.

Conclusion

On the demand side, a 1°F drop raises hourly demand by ∼1.15 MWh per climate division at mean income, confirming that severe cold significantly raises hourly electricity demand. The severe-cold demand response is also monotonically increasing in income: roughly zero at $20k below the mean, −1.15 MWh/°F at the mean, and −2.31 MWh/°F at $20k above the mean. This confirms that the severe-cold demand response varies systematically with median income, consistent with higher-income households having greater heating capacity. The heating-mix interaction is statistically insignificant once income is included, because electric-heating share and income are already highly correlated in Minnesota.

On the supply side, each 1 MWh of wind displaces 0.49 MWh of gas, 0.32 MWh of coal, and 0.13 MWh of net imports, summing to −0.94. The residual 6% is absorbed by nuclear, hydro, and biomass. Notably, colder hours within the extreme-cold sample

Table 7: *Part A Robustness Check: Northern Minnesota Climate Divisions Only*

	Northern MN only
Temperature (°F) at mean income	-0.812^{***}
	(0.036)
Temp × Income centered ($1k)	-0.033^{**}
	(0.004)
Wind speed	0.189^{**}
	(0.028)
Observations	20,950
Within R^2	0.897

Sample restricted to climate divisions 2101, 2102, and 2103 (Northwest, North Central, and Northeast Minnesota). Specification matches Column 5 of Table 3, run on this northern subsample only. Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

are associated with *lower* wind generation. At the same time, the wind × temperature interaction is small and insignificant, suggesting that the marginal displacement effect of wind does not vary substantially with temperature. This implies that even when wind output is weaker during the most system-stressed hours, the grid rebalances through greater reliance on conventional generation and net imports.

These findings connect to the existing literature in two ways. On the demand side, the income gradient corroborates Doremus et al. (6)'s finding that extreme weather creates unequal energy burdens across income groups, and the temperature response is consistent with Berkouwer (3), who links the cold demand response directly to the prevalence of electric heating. On the supply side, the lower wind generation during the coldest hours is consistent with the icing mechanism documented in Gao and Hong (10), and the displacement coefficients align with the wind intermittency literature (5; 8; 12).

Two directions for future research follow from these limitations. First, gathering temperature and demand data from all 15 MISO states would allow the Part B supply regressions to use a true MISO-wide temperature shock rather than the Minnesota proxy employed here. Second, decomposing the demand series into residential and industrial components would sharpen the test of H2, since residential and industrial loads respond to cold very differently but are currently pooled into the same total demand variable.

In conclusion, the Minnesota grid exhibits a partially self-correcting structure during cold shocks: when demand surges, lower wind output during the coldest hours is offset by greater reliance on gas, coal, and net imports. However, lower-income areas absorb cold shocks partly by under-heating, a welfare and health cost that is invisible in the MWh data.

References

[1] Alberini, A., Gans, W., and Velez-Lopez, D. (2011). Residential consumption of gas and electricity in the U.S.: The role of prices and income. *Energy Economics*, 33(5):870–881.

[2] Beltrami, F., Grossi, L., and Liebensteiner, M. (2025). Beyond borders: Estimating the marginal emission factor of electricity trade. *Energy Economics*, 148:108588.

[3] Berkouwer, S. B. (2020). Electric heating and the effects of temperature on household electricity consumption in South Africa. *The Energy Journal*, 41(4):209–230.

[4] Bessec, M. and Fouquau, J. (2008). The non-linear link between electricity consumption and temperature in Europe: A threshold panel approach. *Energy Economics*, 30(5):2705–2721.

[5] Cullen, J. A. and Mansur, E. T. (2017). Inferring carbon abatement costs in electricity markets: A revealed preference approach using the shale revolution. *American Economic Journal: Economic Policy*, 9(3):106–133.

[6] Doremus, J. M., Jacqz, I., and Johnston, S. (2022). Sweating the energy bill: Extreme weather, poor households, and the energy spending gap. *Journal of Environmental Economics and Management*, 112:102609.

[7] Dubin, J. A. and McFadden, D. L. (1984). An econometric analysis of residential electric appliance holdings and consumption. *Econometrica*, 52(2):345–362.

[8] Fell, H. and Kaffine, D. T. (2018). The fall of coal: Joint impacts of fuel prices and renewables on generation and emissions. *American Economic Journal: Economic Policy*, 10(2):90–116.

[9] Fell, H., Kaffine, D. T., and Novan, K. (2021). Emissions, transmission, and the environmental value of renewable energy. *American Economic Journal: Economic Policy*, 13(2):241–272.

[10] Gao, L. and Hong, J. (2021). Wind turbine performance in natural icing environments: A field characterization. *Cold Regions Science and Technology*, 181:103193.

[11] Iowa Environmental Mesonet (2026). Download ASOS/AWOS/METAR data: Minnesota ASOS network (MN_ASOS). Iowa State University. Retrieved March 28, 2026, from the IEM data portal.

[12] Kaffine, D. T., McBee, B. J., and Ericson, S. J. (2020). Intermittency and CO2 reductions from wind energy. *The Energy Journal*, 41(5):23–54.

[13] Lee, C.-C. and Chiu, Y.-B. (2011). Electricity demand elasticities and temperature: Evidence from panel smooth transition regression with instrumental variable approach. *Energy Economics*, 33(5):896–902.

[14] Lücke, L., Dent, C. J., Hegerl, G., Wilson, A. L., and Schurer, A. (2024). Severe compound events of low wind and cold temperature for the British power system. *Meteorological Applications*, 31(4):e2219.

[15] Nam, K. and Seo, W.-K. (2026). Nonlinear temperature sensitivity of residential electricity demand: Evidence from a distributional regression approach. *Energy Economics*, 153:109076.

[National Institute on Minority Health and Health Disparities] National Institute on Minority Health and Health Disparities. Income table for Minnesota counties. HDPulse Data Portal. Retrieved March 28, 2026.

[NOAA National Centers for Environmental Information] NOAA National Centers for Environmental Information. CONUS climate divisions shapefile. NOAA Physical Sciences Laboratory. Retrieved March 28, 2026.

[18] Obika, K., Cole, W., and Rivers, M. (2025). Hourly electricity demand profiles for each county in the contiguous United States. Open Energy Data Initiative, National Renewable Energy Laboratory. Data set.

[19] Ohler, A., Loomis, D. G., and Marquis, Y. (2022). The household appliance stock, income, and electricity demand elasticity. *The Energy Journal*, 43(1):241–262.

[20] Thornton, H. E., Scaife, A. A., Hoskins, B. J., and Brayshaw, D. J. (2017). The relationship between wind power, electricity demand and winter weather patterns in Great Britain. *Environmental Research Letters*, 12(6):064017.

[21] U.S. Census Bureau (2022). American community survey 5-year estimates: Table B25040, house heating fuel (Minnesota counties). Retrieved March 28, 2026.

[22] U.S. Energy Information Administration. Minnesota state energy profile. U.S. Department of Energy. Retrieved March 28, 2026.

[23] U.S. Energy Information Administration. Real-time operating grid: Midwest region (REG-MIDW). U.S. Energy Information Administration. Retrieved March 28, 2026.

[24] Vaage, K. (2000). Heating technology and energy use: A discrete/continuous choice approach to Norwegian household energy demand. *Energy Economics*, 22(6):649–666.

How Did the COVID-19 Pandemic Affect Household Entertainment Expenditure in the U.S.?

Anna Fratto

ECON 381: Introduction to Econometrics (Advisor: Amy Damon)

This paper investigates the effect of COVID-19 severity on household entertainment expenditure across 23 U.S. Metropolitan Statistical Areas from 2005 to 2024 using Consumer Expenditure Survey data. I model COVID-19 as a shadow cost that raises the effective price of entertainment consumption, generating a substitution effect away from in-person activities toward at-home alternatives. I estimate four panel regression specifications with MSA and year fixed effects, using cumulative COVID-19 cases as the primary explanatory variable and deaths and vaccination rates as robustness checks. Across all specifications, COVID-19 indicators are statistically insignificant, while household income is the only consistently significant predictor at the 1% level, consistent with entertainment being a normal good. I interpret the null result as reflecting data constraints: annual aggregation smooths away within-year behavioral responses, the effective pandemic sample is limited to 88 MSA-year observations, and year fixed effects absorb much of the identifying variation in COVID-19 severity. The directional patterns in the data nonetheless support the theoretical prediction, with entertainment declining relative to household necessities. These findings are consistent with other papers who find that lockdown policies rather than case counts drove the steepest declines in discretionary spending. Future work should use monthly data, expand the MSA sample, and incorporate lockdown policy indicators to better identify the causal effect of COVID-19 on entertainment expenditure.

Introduction

In 2020, the CDC recorded 384,536 deaths in the U.S. alone (2). The COVID-19 pandemic drastically altered consumer behavior by restricting in-person activities, and entertainment was one of the categories most directly affected given how much of it depends on physical attendance at venues and events. This paper examines entertainment expenditure, defined by the BLS Consumer Expenditure Survey as annual household spending on fees and admissions, audio and visual equipment, pets, toys, and recreational items. The average American spends about $3,600 annually on this category, with households in Minneapolis-St. Paul spending about $5,041 and Seattle spending the highest at about $7,108 (4). Understanding shifts in entertainment spending matters because it captures how households make discretionary choices under uncertainty, reflecting broader changes in consumer behavior during a public health crisis.

[3]Anna Fratto is a Junior majoring in Economics. Correspondence concerning this article should be addressed to afratto@macalester.edu.

For example, between 2010 and 2019, the U.S. box office averaged approximately $10.88 billion annually. In 2020, box offices generated approximately $2.11 billion, an 81.4% decrease from 2019 (7). As of 2024, the industry continues to recover and consumer behavior continues to evolve. In 2025, the average price of a movie ticket in the U.S. would cost about $16.08 (10), while the cheapest Netflix streaming plan is $7.99/month (17). This price differential, combined with pandemic-era closures, created a strong incentive to substitute toward streaming, evidenced by the simultaneous box office collapse and growth of streaming platforms (22). Thus, I explore the relationship between COVID-19 and entertainment demand in the U.S.

Several studies have examined how COVID-19 affected movie theaters and the broader entertainment industry. A Turkish study found results consistent with my hypothesis: increased access to online streaming reduced the need for in-person entertainment, with a global decrease of approximately 72% in cinema attendance and expectations that attendance would remain low beyond the pandemic (20). In contrast, a study of Korean movie theaters found that the negative effects of COVID-19 were short-lived, with attendance rebounding relatively quickly (14). Spectator entertainment specifically declined sharply as major sports and entertainment gatherings were banned, while e-commerce spending shifted toward household necessities (1). The mixed evidence across countries suggests that the magnitude and persistence of COVID-19's effect on entertainment may depend on local market conditions, motivating a U.S.-specific analysis.

In this paper, I explore the relationship between COVID-19, through reported positive cases, vaccinations, and number of deaths, and entertainment expenditure across 23 Metropolitan Statistical Areas (MSA) in the U.S. measured annually from 2005 to 2024, see full list in Appendix Table 4. My hypothesis is that as COVID-19 rates increase, household entertainment expenditure will decrease. I use annual Consumer Expenditure Survey data from the BLS and U.S. Census Bureau, alongside CDC records of COVID-19 cases, deaths, and vaccinations to analyze which pandemic indicators most strongly influence household entertainment expenditure across the U.S.

Literature Review

The entertainment industry is shaped by pricing based on supply and demand and consumer decision factors (12; 3). Most forms of entertainment depend on three important characteristics: physicality, synchronicity, and digitizability. Industries that were physical and synchronous, such as movie theaters, live concerts, and sporting events, suffered the most during the pandemic, while digitizable alternatives like streaming services and online gaming thrived (24; 21). In non-pandemic times, consumer demand for movies depends more on the movie itself than on traditional macroeconomic factors such as GDP or consumer confidence (27). The shift toward online entertainment was already underway before COVID-19, but the pandemic accelerated investment in subscription-based services and digital gaming ecosystems (18).

In the U.S., as case counts grow, households alter their spending across categories. The first category that increases is food, followed by retail (i.e. toilet paper, personal care items) due to stockpiling tendencies (6). Additionally, higher-income households chose to reduce their spending most sharply, particularly in sectors requiring in-person interaction, as their greater ability to work remotely and avoid health risk gave them more flexibility

to substitute away from contact-intensive consumption (9). One survey of over 10,000 U.S. households found that aggregate consumer spending dropped by about 31 percentage points, with the biggest declines in travel and clothing. The timing of lockdowns helps explain this steep decline in total household spending (11).

More indicators include household consumption, income, and savings, with higher percentiles of the consumption distribution not increasing before the pandemic and falling in 2020 (16). Germany was one of the countries without a mandated lockdown, but still found that households reduced spending on durable goods and social activities as infection rates rose (5). Overall, there was a big shift in household consumption during the pandemic, as many entertainment activities are primarily in person.

During the COVID-19 pandemic, health behaviors reduce in-person activities (29). Research from China shows that public health shocks directly decreased cultural consumption, including theater attendance, confirming an inverse relationship between pandemic severity and in-person attendance (15). Personal health safety concerns became a dominant decision factor for households, such as framing entertainment options based on risk of infection and crowding (21). The health shocks also shift household spending priorities; increasing spending on healthcare and infection prevention and decreasing non-essentials and discretionary activity (23). The pandemic demonstrates that consumer behavior in cultural industries shifts when public health shocks affect discretionary consumption (25).

Broader industry research shows that COVID-19 produced multi-billion-dollar losses and extensive disruptions across the global entertainment sector (18). Numerous surveys show that people reduced movie theater visits in favor of alternative media consumed at home (19; 26).

Many families even got "pandemic puppies" (8), and the BLS Consumer Expenditure Survey includes pets in the entertainment category. Specifically for young adults, the pandemic did not just increase entertainment demand, but transformed it into a more conscious choice between online vs. offline life and social vs. solitary experiences (13). The behavioral explanation is that our social and emotional needs were unsatisfied during the lockdowns so consumers substituted in person with digital alternatives (28). Overall, the pandemic accelerated the use of online streaming as a substitute for in-person entertainment, supported by company adaptations that incentivized at-home consumption. Given that activities like movies and concerts involve prolonged indoor exposure, health-risk perceptions played a key role in shaping consumer behavior, making it useful to examine entertainment consumption alongside COVID-19 indicators.

This snapshot provides empirical evidence that pandemic-induced behavior, beyond policy restrictions, significantly changed entertainment demand, highlighting the need to examine this effect in the U.S. Most studies focus either on smaller countries, or the broader consumption household changes. Few studies link COVID-19 indicators directly to entertainment expenditure, indicating a gap this study aims to fill. Thus I intend to explore which COVID-19 indicators most strongly affect entertainment expenditure and decision-making in the U.S. market.

Economic Theory

I assume households maximize utility over entertainment and non-entertainment consumption. During COVID-19, in-person entertainment became associated with in-

creased health risk, which I model as a shadow cost that rises with case counts and raises the implicit price of entertainment. This generates a substitution effect away from in-person activities toward at-home alternatives. If entertainment is additionally a normal or luxury good, pandemic-era income losses and uncertainty further decrease demand through an income effect.

A household maximizes utility of consumption as follows:

$$\max_{C_e, C_n} \; U(C_e, C_n) - \phi(COVID\text{-}19_{m,y}) \times C_e \tag{1}$$

Where C_e is entertainment consumption, C_n is all other consumption, and $\phi(COVID\text{-}19_{m,y})$ is the perceived risk from COVID-19 exposure. I assume $\phi \geq 0$ and $\phi' > 0$, so that perceived risk is non-negative and strictly increasing in case counts. The shadow cost $\phi(COVID\text{-}19_{m,y}) \times C_e$ captures the idea that each unit of entertainment consumption becomes increasingly costly in utility terms as the pandemic worsens. As COVID-19 cases rise in a given MSA and year, households face a higher cost of consuming entertainment, reducing its attractiveness relative to other goods.

The household budget constraint is defined by

$$P_e \cdot C_e + P_n \cdot C_n = Y \tag{2}$$

Where P_e is the price of entertainment and P_n is the price of all other goods respectively, and Y is household income. Because ϕ enters the function only through entertainment consumption, the price of entertainment rises with COVID-19 severity while the price of all other goods remains unchanged.

Setting up the Lagrangian we get:

$$\mathcal{L} = U(C_e, C_n) - \phi(COVID\text{-}19_{m,y}) \times C_e - \lambda(P_e C_e + P_n C_n - Y) \tag{3}$$

The first-order conditions for equation (3) are

$$\frac{\partial U}{\partial C_e} - \phi(COVID\text{-}19_{m,y}) = \lambda P_e$$

for the entertainment category, and

$$\frac{\partial U}{\partial C_n} = \lambda P_n$$

for all other consumption categories. Dividing and rearranging gives the optimality condition

$$\frac{\partial U / \partial C_e}{\partial U / \partial C_n} = \frac{P_e + \phi(COVID\text{-}19_{m,y})}{P_n}$$

which means that as COVID-19 rises, the shadow price ϕ also increases, decreasing household C_e to restore optimality given diminishing marginal utility. Income losses during the pandemic reinforce this substitution, further reducing C_e at the optimum.

I expect that as COVID-19 cases increase, there will be a significant decrease in entertainment spending relative to other categories. The substitution effect and utility

maximization model described above will apply: people will spend less on in-person entertainment as COVID-19 cases rise, weighing the health risk in their consumption decision. This framework predicts $\beta_1 < 0$ on COVID-19 cases in my regression. Conversely, as COVID-19 severity subsides, the shadow cost ϕ falls, and the effective price of entertainment returns toward P_e, and household entertainment spending is expected to recover toward its pre-pandemic level.

Empirical Strategy

To investigate the effect of COVID-19 on consumer expenditure for entertainment in the U.S., I will estimate this relationship using a regression model that controls for MSA-level characteristics. I will use the following specified regression model for entertainment and also the other categories for comparisons:

$$CE_{m,y} = \beta_0 + \beta_1 \, COVID\text{-}19_{m,y} + \mathbf{Z}'_{m,y}\mathbf{\Gamma} + \alpha_m + \gamma_y + \varepsilon_{m,y} \tag{4}$$

The first regression (4) models the effect of COVID-19 cases ($COVID\text{-}19_{m,y}$) on consumer expenditure in the entertainment category ($CE_{m,y}$) with two-way fixed effects. The subscript m represents each MSA, and y corresponds to the different years. $\mathbf{Z}'_{m,y}\mathbf{\Gamma}$ is a matrix of control variables including age, income, and household size, and $\varepsilon_{m,y}$ is the error term.

A negative and significant β_1 would indicate that higher COVID-19 case counts are associated with decreased household entertainment spending within an MSA, consistent with pandemic-driven behavioral restrictions.

To interpret β_1, the following assumptions must be true. In the absence of COVID-19, spending trends across MSAs would have followed parallel paths over time. I use 2005–2018 as a pre-pandemic baseline, assuming that MSAs with high COVID-19 case counts were not already diverging in entertainment spending prior to 2019. Additionally, there is no OVB after conditioning the MSAs with year fixed effects. I assume that COVID-19 case counts are random, but correlated with local conditions such as density or higher-income spending habits. Controlling for income helps, but this remains a limitation. There is no reverse causality of consumer expenditure and COVID-19 case count impacting regional case totals. Fixed effects and controls consume degrees of freedom quickly, so I am parsimonious in my choice of control variables. I cluster standard errors at the MSA level to account for within-MSA serial correlation over time.

I estimate four specifications that build progressively on one another. The first model (4) adds MSA and year fixed effects, allowing identification to come from within-MSA variation over time while controlling for unobserved, time-invariant differences across MSAs and common time shocks. The second model (5) extends this specification by including COVID-19 deaths, helping distinguish between the effects of disease spread (cases) and severity (deaths). The third model (6) further incorporates vaccination rates, measured as the share of individuals who completed the full vaccine series, to account for behavioral and policy responses to the pandemic. Finally, the fourth model (7) replaces deaths with the interaction term to avoid multicollinearity between the three COVID-19 variables, providing insight into potential mechanisms.

$$CE_{m,y} = \beta_0 + \beta_1 \, COVID\text{-}19_{m,y} + \beta_2 \, Deaths_{m,y} + \mathbf{Z}'_{m,y}\mathbf{\Gamma} + \alpha_m + \gamma_y + \varepsilon_{m,y} \tag{5}$$

$$CE_{m,y} = \beta_0 + \beta_1 \, COVID\text{-}19_{m,y} + \beta_2 \, Deaths_{m,y} + \beta_3 \, Vax_{m,y} + \mathbf{Z}'_{m,y}\mathbf{\Gamma} + \alpha_m + \gamma_y + \varepsilon_{m,y} \tag{6}$$

$$CE_{m,y} = \beta_0 + \beta_1 \, COVID\text{-}19_{m,y} + \beta_2 \, Vax_{m,y} + \beta_3 \, (COVID\text{-}19_{m,y} \times Vax_{m,y}) + \mathbf{Z}'_{m,y}\mathbf{\Gamma} + \alpha_m + \gamma_y + \varepsilon_{m,y} \tag{7}$$

The main identification challenge is that COVID-19 case counts are not randomly assigned across MSAs. Higher case count cities may differ from lower case count cities in ways that also affect entertainment spending, such as existing cultural preferences for in-person activities. While MSA fixed effects absorb time-invariant differences across cities and year fixed effects absorb common national shocks, any remaining time-varying confounders within an MSA would bias β_1. I cannot claim a causal interpretation of β_1 because of these limitations, and instead interpret the estimations as conditional correlations.

To assess how close I get to identification, I do three things. First, I use the pre-pandemic period as a baseline to visually inspect whether high and low COVID-19-exposure MSAs were already trending differently in entertainment spending. If pre-trends are parallel, this increases confidence in the two-way fixed effects strategy. Second, I run the deaths specification as a robustness check: if the sign and magnitude of β_1 are consistent when substituting deaths for cases, this suggests the result is not driven by measurement errors in case reporting. Third, the vaccination interaction in specification (7) provides a partial mechanism check. If vaccination rates moderate the negative relationship between cases and spending, this is consistent with the shadow price interpretation from my theoretical model rather than a spurious correlation.

Data Description

The main source of my consumer expenditure data from the U.S. is the Consumer Expenditure Survey (CES) collected by the U.S. Government Bureau of Labor Statistics. The CE collects detailed information on both household expenditures and income using a questionnaire and Diary Survey. My sample includes 23 Metropolitan Statistical Areas, split into four regions of the U.S. (Midwest, Northeast, South, and West) to provide a representative sample of the different U.S. households. The panel is unbalanced as not all MSAs report data for every year. I originally compiled 25 MSAs but dropped Pittsburgh and Portland due to insufficient time coverage, leaving a final sample of 23 MSAs across 20 years. The years 2005–2018 serve as a pre-pandemic baseline, 2020–2023 capture the pandemic period, and 2024 is treated as post-pandemic as COVID-19 had ceased to meaningfully restrict social activity. I also used the U.S. Census Bureau for added population controls, which helped to control for differences in city size and density, which are correlated with both COVID-19 case counts and baseline entertainment spending patterns.

The spending categories represent major household consumption groups. Food, household necessities, personal expenses, and transportation are composite categories constructed by aggregating related CE sub-variables, while entertainment is taken directly as

a single BLS category. I expect essentials to remain relatively stable during the pandemic while non-essentials decrease and savings increase. COVID-19 case and death counts come from the New York Times COVID-19 database aggregated to the MSA level, and vaccination records come from the Centers for Disease Control and Prevention (CDC). Cumulative cases are my main explanatory variable, with vaccination rates as a secondary recovery indicator and deaths as a robustness check. I also used data from the U.S. Census Bureau for controls.

Table 1: *Descriptive Statistics, U.S. Consumer Expenditure, 2005–2024*

Variable	# of Observations	Mean	Standard Deviation	Minimum	Maximum
COVID-19 Cases (1000)	400	239.4588	639.0682	0	4515.198
COVID-19 Deaths (1000)	400	2.697895	6.9831	0	43.969
Vaccinations (1000)	400	644.8087	1931.562	0	15497.58
Entertainment (U.S.D)	400	3222.997	910.3701	1352	7108
Food (U.S.D)	400	8111.285	1921.093	4324	14676
Household Necessities* (U.S.D)	400	26605.44	12486.78	0	56839
Personal Expenses (U.S.D)	400	4214.752	834.7545	2066	7477
Transportation (U.S.D)	400	10297.83	2287.588	5799	19624

*Household necessities contain missing values for select MSA-years where CE sub-components were not reported and replaced with 0, and should be interpreted with caution. COVID-19 cases, deaths, and vaccinations are scaled in thousands. COVID-19 variables are coded as zero for pre-pandemic years 2005–2019.

Table 1 presents descriptive statistics for the COVID-19 variables of interest and CE consumption categories across 400 MSA-year observations from 2005 to 2024. Entertainment expenditure averages $3,222 annually per household, ranging from $1,352 to $7,108, reflecting meaningful variation across MSAs and years that will help identify the effect of COVID-19 severity on spending. Food is the second largest consumption category at an average of $8,111, consistent with its status as a non-discretionary essential. Household necessities average $26,649 but have a minimum of zero due to missing sub-components replaced with zero, so this category should be interpreted with caution. Transportation averages $10,297 and personal expenses average $4,214, both showing moderate variation across the sample.

For the COVID-19 variables, cumulative cases average 240,662 per MSA-year but range from zero for pre-pandemic years up to 4,515,198, reflecting dramatic variation in pandemic intensity across both time and geography. Deaths follow a similar pattern, averaging 2,711 per MSA-year but reaching as high as 43,969 in the hardest-hit MSAs. Vaccination counts average 648,049 per MSA-year but are highly right-skewed, with a maximum of 15,497,580 reflecting the largest MSAs. In the regressions, vaccination is normalized as the share of dose-1 recipients who completed the full vaccine series to account for differences in MSA population size. All COVID-19 variables are coded as zero for pre-pandemic years 2005–2019, meaning the non-zero variation is concentrated in 2020–2023.

Figure 1 plots the average annual household expenditure across all five consumption categories from 2005 to 2024, Figure 2 shows consumer expenditure alone. Vertical dashed lines mark the pandemic period 2020 to 2023. Entertainment spending grows steadily from 2005 through 2019, consistent with rising incomes and expanding leisure options. Food and household necessities follow a similar upward trend, while transportation and personal expenses remain relatively stable. Beginning in 2020, entertainment shows a visible dip relative to its pre-pandemic trend, while food spending appears relatively stable, consistent

Figure 1: *U.S. Consumer Expenditure by Category, 2005–2024*

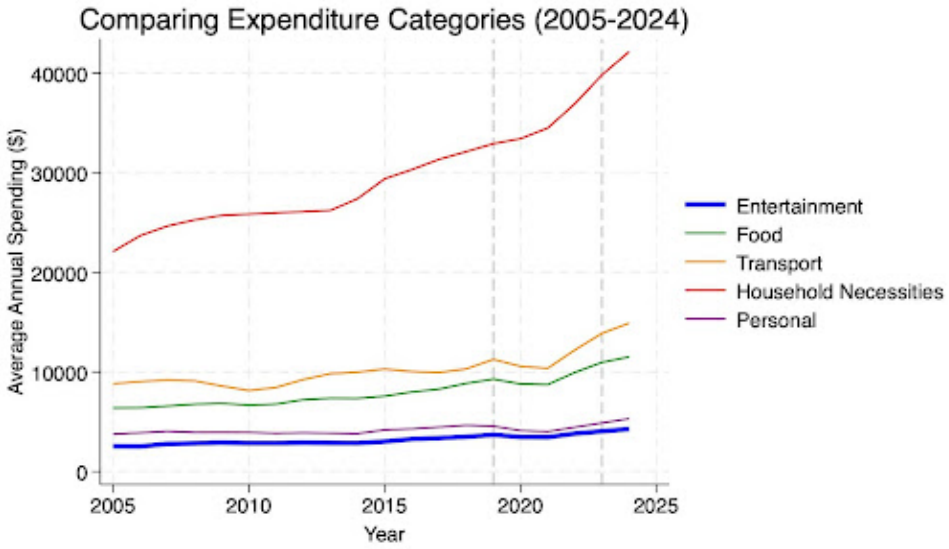
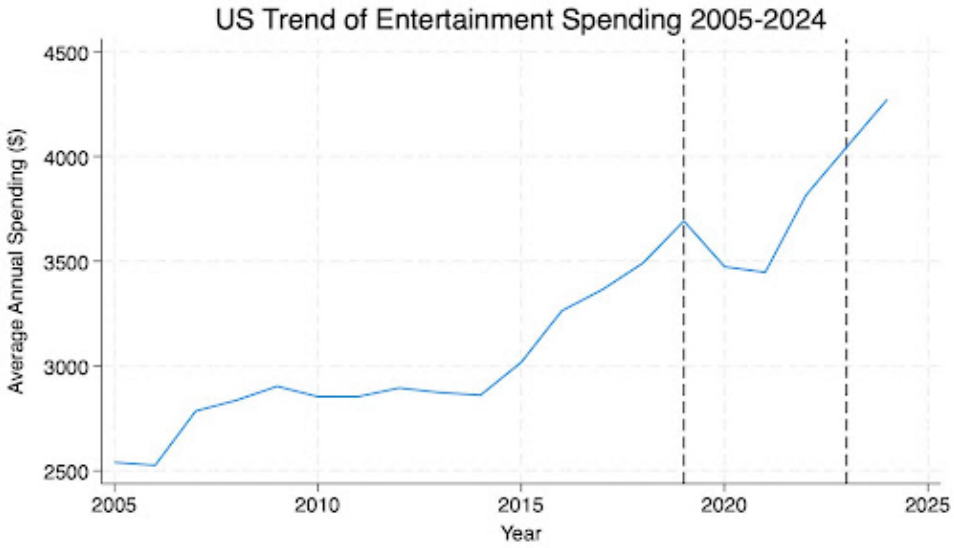


Figure 2: *U.S. Entertainment Consumer Expenditure, 2005–2024*



with its status as a non-discretionary essential.

Results

Table 4 shows the results of four regression specifications examining the effect of COVID-19 severity on household entertainment expenditure across 23 MSAs from 2005 to 2024. Across all four specifications, the COVID-19 variables are statistically insignificant. The coefficients on COVID-19 cases are small and negative in specifications (1) through (3), ranging from −0.006 to −0.065 per thousand cases, but none are statistically signif-

Table 2: *Effect of COVID-19 on Entertainment Expenditure*

	(1)	(2)	(3)	(4)
Variables	Baseline	Main Spec	w/ Deaths	Full Model
COVID-19 Cases (1000)	−0.032	−0.006	−0.065	0.000
	(0.098)	(0.235)	(0.243)	(0.291)
COVID-19 Deaths (1000)		−2.397	−4.900	−6.740
		(24.661)	(25.881)	(26.321)
Vaccination (1000)			0.029	0.056
			(0.043)	(0.081)
Interaction				0.000
				(0.000)
Income Before Taxes	0.019***	0.019***	0.019***	0.019***
	(0.006)	(0.006)	(0.005)	(0.006)
Age of Reference Person	−12.296	−12.405	−11.791	−11.590
	(25.173)	(25.183)	(25.256)	(25.455)
Earners	−715.651	−711.597	−694.784	−693.183
	(448.772)	(452.276)	(442.156)	(440.580)
Constant	2,829.500*	2,831.929*	2,801.926*	2,784.404*
	(1,432.474)	(1,429.470)	(1,433.756)	(1,445.135)
Observations	400	400	400	400
R-squared	0.531	0.531	0.532	0.532
Controls	Yes	Yes	Yes	Yes

Note: The dependent variable is annual household entertainment expenditure in USD. All regressions include MSA and year fixed effects. COVID-19 cases, deaths, and vaccinations are scaled in thousands. Specification (4) replaces deaths with a cases × vaccination interaction term. Household controls include income before taxes, age of reference person, and number of earners. The sample covers 23 MSAs from 2005–2024, with COVID-19 variables coded as zero for pre-pandemic years. Robust standard errors in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

icant given the large standard errors. Deaths are similarly negative across specifications (2) through (4), with coefficients ranging from −2.397 to −6.740 per thousand deaths, but remain insignificant. Vaccination shows a small positive coefficient in specifications (3) and (4), consistent with the theoretical prediction that pandemic recovery would increase entertainment spending, but is also statistically insignificant. The interaction term in specification (4) is also zero, likely reflecting a scaling issue and should be interpreted with caution.

The only statistically significant predictor across all four specifications is the control income before taxes, with a coefficient of approximately 0.019 significant at the 1% level in every specification. This indicates that a one-dollar increase in pre-tax income is associated with approximately $0.019 in additional annual entertainment expenditure, holding all else equal. This is consistent with entertainment being a normal good as predicted by my utility maximization model. The age of the reference person and number of earners in the household are both negative but statistically insignificant across all specifications.

The absence of statistically significant COVID-19 coefficients warrants further interpretation with three factors. First, there is a multicollinearity issue between the year fixed effects and the COVID-19 variables. Since COVID-19 cases, deaths, and vaccinations only vary during 2020–2023, the year dummy variables for those four years absorb much of the variation in COVID-19 severity before the COVID-19 variables can explain it. This leaves very little independent variation for the COVID-19 coefficients to identify. Second, the sample for COVID-19 is small, which substantially limits statistical power. Third, there is heterogeneity across MSAs where some larger cities were hit worse than others. Thus, it may average out in the aggregate, masking effects that would be detectable with more granular data.

Table 3: *Effect of COVID-19 on Other Expenditure Categories*

	(1)	(2)	(3)	(4)
Variables	Food	Household Necessities	Personal Expenses	Transportation
COVID-19 Cases (1000)	−0.270	−0.865***	−0.042	−0.023
	(0.166)	(0.292)	(0.084)	(0.221)
Income Before Taxes	0.039***	0.176***	0.024**	0.042***
	(0.008)	(0.017)	(0.009)	(0.012)
Age of Reference Person	20.571	94.195	63.124*	62.665
	(52.035)	(75.688)	(31.515)	(80.512)
Earners	560.758	647.047	−418.997	1,886.433*
	(555.609)	(2141.718)	(572.822)	(1039.970)
Constant	1,963.168	4,887.912	−418.480	364.710
	(2587.512)	(4788.624)	(1944.785)	(4416.717)
Observations	400	400	400	400
R-squared	0.824	0.980	0.509	0.738
Controls	Yes	Yes	Yes	Yes

Note: The dependent variable is annual household expenditure in U.S.D. for each category as defined by the BLS Consumer Expenditure Survey. All regressions include MSA and year fixed effects. COVID-19 cases are scaled in thousands. Food, household necessities, personal expenses, and transportation are composite categories constructed by aggregating related CE sub-variables. Household controls include income before taxes, age of reference person, and number of earners. The sample covers 400 MSA-year observations across 23 MSAs from 2005–2024, with COVID-19 cases coded as zero for pre-pandemic years. Robust standard errors in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table 5 presents results for the other expenditure categories. Household necessities is the only category with a statistically significant COVID-19 coefficient, at −0.865 per thousand cases significant at the 1% level. This is a surprising finding given the theoretical prediction that essentials would remain stable during the pandemic. One possible explanation is the zero-replacement for missing sub-components discussed in the data section, which may be driving this result rather than a true behavioral response. Food, personal expenses, and transportation all show negative but insignificant coefficients on COVID-19 cases. Income remains positive and significant at the 1% level across all four categories, consistent with all spending categories being normal goods.

The null results on COVID-19 variables contrast with findings from several related papers, though important data differences help explain the divergence. Baker et al. (6) found that as case counts grew, households radically altered spending across categories, with initial stockpiling spikes in food and retail followed by sharp declines in travel, entertainment, and contact-intensive services. This reflects a clear reallocation of spending rather than a uniform decline, consistent with the substitution effect predicted by my theoretical model

(6). However, this study uses transaction-level data at daily and weekly frequency, which captures the sharp within-month behavioral responses that annual MSA-level averages in my data smooth out entirely.

Another study finds that high-income households drove the initial spending decline in sectors requiring in-person interaction, consistent with my finding that income is the only significant predictor of entertainment expenditure across all specifications (9). Most relevant to my null result, Coibion et al. (11) find that lockdown policies and expectation shifts, rather than case counts themselves, drove the steepest declines in discretionary spending. This suggests that cumulative case counts may simply be the wrong variable to capture the mechanism, even if the underlying effect is real.

Limitations

The most significant limitation is that the CE data is collected annually at the MSA level. COVID-19's behavioral effects were concentrated in specific weeks and months—the initial lockdown shock in March 2020, the Delta wave in late 2021, and the Omicron surge in early 2022—all of which are smoothed away by annual aggregation.

Additionally, several important confounders are not captured in my specifications. Most notably, I did not control for local lockdown policies or government restrictions, which Coibion et al. (11) find to be more predictive of spending declines than case counts themselves. Employment levels are another likely confounder: households facing unemployment will reduce discretionary spending regardless of COVID-19 severity, and failing to control for this may bias my estimates. Political beliefs and partisanship represent a further omitted variable, as several studies find that partisan differences in risk perception significantly affected behavioral responses to the pandemic.

Two forms of measurement error affect my COVID-19 variables. First, the CE data is self-reported annually, meaning respondents must recall and estimate expenditures across broad categories, introducing recall bias. Second, cumulative COVID-19 case counts systematically undercount true infections because they only capture cases that were officially tested and confirmed. For example, home testing was largely unreported, meaning true case counts were likely substantially higher than recorded. This attenuation bias would push my COVID-19 coefficients toward zero even if a true effect exists.

There is a potential reverse causality concern between spending and COVID-19 case counts. Higher in-person spending could increase contact rates and drive case growth, biasing the estimated coefficient on cases. This concern is most relevant for entertainment and food away from home, where spending directly involves in-person interaction. It is less of a concern for household necessities, where spending is largely non-discretionary and less contact-intensive.

Lastly, with only 23 MSAs and four pandemic years, the effective sample for identifying COVID-19 effects is 92 MSA-year observations. This limits statistical power considerably. Averaging across MSAs may mask heterogeneity, as lower-income households likely reallocated spending very differently from higher-income households, and cities with stricter lockdown policies likely experienced sharper entertainment declines. Future work should use monthly data, incorporate lockdown policy indicators, expand the MSA sample, and use household-level microdata to better identify heterogeneous effects across income groups.

Conclusion

I investigated the effect of COVID-19 severity on household entertainment expenditure across 23 U.S. Metropolitan Statistical Areas from 2005 to 2024 using Consumer Expenditure Survey data. Across all four panel regression specifications, I fail to reject the null hypothesis that COVID-19 indicators have no statistically significant effect on entertainment expenditure. Cases, deaths, and vaccination rates all produce small, insignificant coefficients, while household income is the only consistently significant predictor at the 1% level, consistent with entertainment being a normal good.

This null result likely reflects data constraints—annual aggregation, a small pandemic sample, and missing policy controls—rather than a true absence of effect. However, the broader pattern in the data is nonetheless directionally consistent with the existing literature. Household necessities remain relatively stable across the pandemic period while entertainment shows a relative decline during 2020–2023, consistent with the substitution effect predicted by my theoretical model and the reallocation of spending as lockdown policies and behavioral expectation shifts drove the steepest declines in discretionary spending.

This paper contributes to the growing literature on pandemic-induced consumption shifts by providing an MSA-level analysis of entertainment expenditure using CE data over a twenty-year window. The directional patterns nonetheless support the theoretical prediction: entertainment declines relative to household necessities during 2020–2023, consistent with the substitution effect from my model. Future work should use monthly data, expand the MSA sample, and incorporate lockdown policy indicators and household-level microdata to better identify the effect of COVID-19 on entertainment expenditure across income groups.

References

[1] Abdul Hussein, A., Cozzarin, B., and Dimitrov, S. (2022). Changes in consumer spending behavior during the COVID-19 pandemic across product categories. *Electronic Commerce Research*, 24.

[2] Ahmad, F. B. (2023). COVID-19 mortality update — United States, 2022. *MMWR: Morbidity and Mortality Weekly Report*, 72(18):493–496.

[3] Apostolakis, A. (2017). Analysis of the economic impact of cultural festivals in the local economy. *Cultural Management: Science and Education*.

[4] Arky, J. (2024). Americans spend $3.6k annually on fun, see which cities top the list. Yahoo Finance.

[5] Bachmann, R., Bayer, C., and Kornejew, M. (2025). Pandemic consumption. *Journal of Economic Dynamics and Control*, 180:105185.

[6] Baker, S. R., Farrokhnia, R. A., Meyer, S., Pagel, M., and Yannelis, C. (2020). How does household spending respond to an epidemic? consumption during the 2020 COVID-19 pandemic. *The Review of Asset Pricing Studies*, 10(4).

[7] Box Office Mojo (2024). Domestic yearly box office.

[8] Brand, C. L., O'Neill, D. G., Belshaw, Z., Pegram, C. L., Stevens, K. B., and Packer, R. M. A. (2022). Pandemic puppies: Demographic characteristics, health and early life experiences of puppies acquired during the 2020 phase of the COVID-19 pandemic in the UK. *Animals*, 12(5):629.

[9] Chetty, R., Friedman, J. N., Hendren, N., and Stepner, M. (2020). How did COVID-19 and stabilization policies affect spending and employment? a new real-time economic tracker based on private sector data. Technical report, National Bureau of Economic Research.

[10] Cinemark (2025). The ultimate guide to movie ticket prices.

[11] Coibion, O., Gorodnichenko, Y., and Weber, M. (2020). The cost of the COVID-19 crisis: Lockdowns, macroeconomic expectations, and consumer spending. *SSRN Electronic Journal*.

[12] Courty, P. (2000). An economic guide to ticket pricing in the entertainment industry. *Recherches Économiques de Louvain*, 66(2):167–192.

[13] Ganter, S. A. (2024). Young adults' perceptions of entertainment consumption in their everyday lives during the COVID-19 pandemic: Negotiating versatility, emotions, and agency in times of limited choice. *Media, Culture & Society*.

[14] Kim, I. K. (2020). The impact of social distancing on box-office revenue: Evidence from the COVID-19 pandemic. *Quantitative Marketing and Economics*, 19.

[15] Marin-Lopez, B. A., Jimenez-Gomez, D., and Abellán-Perpiñán, J.-M. (2022). Behavioral economics in the epidemiology of the COVID-19 pandemic: Theory and simulations. *International Journal of Environmental Research and Public Health*, 19(15):9557.

[16] Meyer, B., Murphy, C., and Sullivan, J. X. (2022). Changes in the distribution of economic well-being during the COVID-19 pandemic: Evidence from nationally representative consumption data. *SSRN Electronic Journal*.

[17] Netflix (2025). Plans and pricing.

[18] Nhamo, G., Dube, K., and Chikodzi, D. (2020). Implications of COVID-19 on gaming, leisure and entertainment industry. In *Counting the Cost of COVID-19 on the Global Tourism Industry*, pages 273–295.

[19] Nie, W. (2022). The impact of COVID-19 on the film industry in China. *International Journal of Frontiers in Sociology*, 4(7).

[20] Okumuş, M. S. (2022). The effects of COVID-19 pandemic on audience practices in cinema, television, and OTT platforms. *İstanbul Ticaret Üniversitesi Sosyal Bilimler Dergisi*, 21(43).

[21] Pretto, R., Huang, A., Ridderstaat, J., de La Mora, E., and Haney, A. (2023). Consumers' behavioral trends in the arts, entertainment, and recreation sector amid a global pandemic: A qualitative study. *Tourism and Hospitality*, 4(2):233–243.

[22] Ptak, M. (2025). From luxury to commodity and back to luxury: The media and entertainment industry. Michigan Journal of Economics.

[23] Qiu, Y. and Zhang, F. (2024). Impact of health shocks on household consumption structure. *Frontiers in Public Health*, 12.

[24] Ryu, S. and Cho, D. (2022). The show must go on? the entertainment industry during (and after) COVID-19. *Media, Culture & Society*, 44(3):591–600.

[25] Tache-Codreanu, A. (2020). The consumption of cinema in the context of the COVID-19 pandemic. *Concept*, 21(2):203–215.

[26] Volchenkov, W., Bichard, S., and Bowman, N. D. (2022). COVID-19 and movie theaters. *Journal of Student Research*, 11(4).

[27] von Rimscha, M. B. (2013). It's not the economy, stupid! external effects on the supply and demand of cinema entertainment. *Journal of Cultural Economics*, 37(4):433–455.

[28] Wang, J. (2022). The impact of COVID-19 pandemic on consumer behaviours in the entertainment industries. In *Proceedings of the 2022 International Conference*.

[29] Zhang, R., Ji, H., Pang, Y., and Suo, L. (2022). The impact of COVID-19 on cultural industries: An empirical research based on stock market returns. *Frontiers in Public Health*, 10.

Data Sources

- U.S. Bureau of Labor Statistics. Consumer Expenditure Surveys. https://www.bls.gov/cex/csxmsa.htm
- U.S. Bureau of Labor Statistics. Consumer Expenditure Survey microdata time series. https://download.bls.gov/pub/time.series/cx/
- The New York Times. COVID-19 data (county-level). GitHub. https://github.com/nytimes/COVID-19-data
- Centers for Disease Control and Prevention. COVID-19 vaccinations in the United States, county-level dataset. https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-County/8xkx-amqh
- U.S. Census Bureau. Metropolitan and micropolitan statistical area delineation files. https://www.census.gov/geographies/reference-files/time-series/demo/metro-micro/delineation-files.html
- U.S. Census Bureau. Population estimates program. https://www.census.gov/programs-surveys/popest.html

Appendix

Table 4: *Metropolitan Statistical Areas Included in Sample*

#	MSA	Region
1	Anchorage	West
2	Atlanta	South
3	Baltimore	South
4	Boston	Northeast
5	Chicago	Midwest
6	Cleveland	Midwest
7	Dallas-Fort Worth	South
8	Denver	West
9	Detroit	Midwest
10	Honolulu	West
11	Houston	South
12	Los Angeles	West
13	Miami	South
14	Minneapolis-St. Paul	Midwest
15	New York	Northeast
16	Philadelphia	Northeast
17	Phoenix	West
18	San Diego	West
19	San Francisco	West
20	Seattle	West
21	St. Louis	Midwest
22	Tampa	South
23	Washington DC	South

Note: Pittsburgh and Portland were excluded due to insufficient time coverage. MSA regions follow the U.S. Census Bureau regional classification.

Investigating Industry Sensitivity to Citywide Speed Limit Reductions

Aram Petrosyan

ECON 381: Introduction to Econometrics (Advisor: Amy Damon)

Introduction

There is a growing adoption of the Vision Zero strategy in the US, which entails a strong commitment to the principle that no one should die or suffer serious injury in road traffic crashes. This strategy prioritizes lower and safer speed limits, leading to several US cities implementing citywide speed limit reductions in the past decades. However, opponents of citywide speed limit reductions have argued that it harms the economy (4).

This paper studies whether there is evidence of various industries slowing down or speeding up in affected cities due to the enforcement of lower speeds on the roads. Previous research on speed limit reductions has shown that they are associated with major drops in fatal injuries, reduction in noise levels, reduction in carbon dioxide and nitrogen oxide emissions and promotion of walking and cycling. Research focused on economic outcomes has also shown 8-9 fold benefits over implementation costs of speed limit reductions, and $15 per resident cost-savings to the healthcare system (6). The contribution of this paper is in directly assessing different industries and how they perform before and after the implementation of a citywide speed limit reduction. Results provide no statistically significant evidence of citywide speed limit reductions impacting local performance in 9 out of 10 industries. The exception is the Leisure & Hospitality industry, where a positive and statistically significant effect is found. Cities that implemented citywide speed limit reductions saw the industry's local employment growth outperform the national trend by approximately 3.8 percentage points, consistent with the theoretical prediction that slower, more pedestrian-friendly streets stimulate demand for restaurants, cafes, and hospitality businesses.

I use data from the American Community Survey as well as the Bureau of Labor Statistics to look at local-level NAICS (North American Industry Classification System) industry performance prior and following treatment, using the staggered Difference-in-Differences approach. This paper starts with an overview of speed limit-related literature. The next two sections briefly suggest a theory for speed limits having a causal relationship with industry performance and describe the data used in the study. The main section walks the reader through 3 different models - univariate, multivariate regressions and a staggered DiD approach that determine the final results. The last two sections address limitations and specific focus areas for future research in the same research question.

[3]Aram Petrosyan is a Junior majoring in Economics and Geography. Correspondence concerning this article should be addressed to apetrosy@macalester.edu.

Literature Review

Speed is one of the most important contributing factors of crashes, injuries and fatalities - among distraction, drunk driving and absence of seatbelt and helmet use. The kinetic energy transferred to the vehicle occupants is determined to be the causal agent of injuries (8). Since late 1990s and early 2000s when many European cities were implementing citywide speed limit reduction policies, there has been a growing size of literature that takes speed limit reduction as a predictor variable and tries to see changes in various economic and non-economic outcomes.

Research on speed limit reductions and non-economic outcomes has been wide from both European and North American cities (17). In Toronto, Canada, a speed limit reduction from 40km/h to 30km/h resulted in a 67% drop in major and fatal injuries (7). Road casualties decreased by 8.74% after the implementation of 20 mph speed limit areas in New York City (10). Speed limit reductions have also been shown to reduce noise levels in a number of cities (2; 3) and in carbon dioxide (CO2) and nitrogen oxide (NOX) emissions (9; 12). Karndacharuk and McTiernan (11) showed that reduced speed limits promote walking, cycling and public transport which improves physical and mental health and lead to better livability in the treated city. Hypothetically, these positive outcomes should produce direct and indirect positive economic benefits for all industries.

There are two established categories of economic outcomes discussed in speed limit reduction research - fuel consumption and overall cost-benefits. In the literature about fuel consumption, there is no definite consensus among researchers. Some research finds that cars have less efficient fuel consumption at lower speeds and that fuel consumption may increase under lower speed limits (14; 16). However newer studies show that decreasing maximum speed by 20% results in a 3.4% reduction in fuel consumption (13). Regarding overall cost-benefits, Davis and Jones (6) find that over a three-year period, the benefits of introducing the 30km/h default speed limit are estimated to be 8-9 times higher than the implementation costs. Jiao et al. (10) showed that reduced speed limits were associated with cost-savings of $15 per resident to the healthcare system. Additionally, implementation of city-wide 30km/h speed limit policy positively influences overall job accessibility by car (1). For a more comprehensive and systematic literature review of effectiveness of speed limit reductions, please refer to Yannis and Michelaraki (17).

Based on the above literature, one gap I identified is assessing industry responsiveness to citywide reductions in speed limits, which is the aim of this paper. This idea was motivated by a study in a closely related parallel literature - research that has congestion as its predictor, rather than speed limits. It has been shown that sensitivity to traffic congestion varies by industry sector depending on differences in each sector's mix of required inputs and hence its reliance on access to skilled labor and specialized inputs (15).

Theory

The hypothetical causal relationship between citywide reduction of speed limits and change in industry performance is informed by aforementioned work on economic costs of urban traffic congestion to different industries. The varying sensitivity to traffic congestion by industry sector is attributable to differences in each industry sector's mix of required inputs and hence its reliance on access to skilled labor, access to specialized inputs, and

access to a large, transportation-based market area (15). Congestion disrupts the access to different inputs for businesses, and Weisbrod essentially identifies a varying sensitivity to congestion by using data to assess the different ways access to inputs changed for different businesses. A reduction in speed limits and congestion are different in practice, but both operate through the mechanism of reducing average speed of cars in the city. Speed limit reduction is exogenous as it is a policy change, while congestion as a predictor in previous studies might have had endogeneity challenges since it can be a product of economic activity as well. Slower traffic can impact different industries' access to labor and inputs. Below is a production function for a given industry:

$$Y_i = F_i(K_i, L_i, A_i) \tag{1}$$

Where Y_i is the output of industry i. K_i is the required capital inputs for industry i. L_i is the required labor inputs for a given industry. I propose A_i to be an access input that depends on road conditions. Worse road conditions for vehicle dependent industries means A_i decreases. On the other hand, for pedestrian-accessible industries worse road conditions might mean that A_i increases. Finally, F_i is the industry specific function that combines the inputs.

Data Description

There are two primary sources in this study: the Quarterly Census of Employment and Wages (QCEW) by the Bureau of Labor Statistics, and the American Community Survey (ACS) by the United States Census Bureau. More specifically, I use data from the State and County Employment and Wages database from the QCEW, and the Selected Economic Characteristics (DP03) database from the ACS. I use 16 cities in my study - 5 in the treatment group, and 11 in the control group. Since speed limit reductions in the US have been carried out primarily citywide, not countywide, it would be ideal to have the aforementioned datasets provide city-level information. Having two alternatives (counties and metropolitan statistical areas), I choose counties as they are generally closer in size to the city boundaries, and often match the city sizes ideally (e.g. Denver & Philadelphia). I use the years 2010 to 2024 due to different cities implementing citywide speed limit reductions in different years during this timeframe. Table 1 gives a summary of the cities included in the study. There is no dedicated dataset that has all US cities with their speed limit policies summarized. The treated group of Seattle, Portland, Minneapolis, Saint Paul and Denver is not exhaustive - there is also New York City that pioneered the citywide speed limit reduction policy in the US in 2014. I excluded NYC due to it possibly distorting the sample because of its special geographic and economic characteristics. NYC spans 5 counties and choosing one county out of the 5 as a proxy for the entire city is problematic. For example, if one chose Manhattan county - its industry mix would heavily be mostly financial or information activities, and not representative of the overall New York City.

The scope of this study only includes the 10 NAICS defined supersectors. Table 2 presents the NAICS industries whose responsiveness and sensitivities to speed limit reductions will be tested. My outcome variable will be based on employment growth differential locally and nationally in a given industry. This is based on an assumption that an industry's

Table 1: *Summary of selected cities for treatment and control groups*

Treated Group of Cities (policy implementation year)	**Control Group of Cities**
Seattle, WA (2016)	Chicago, IL
Portland, OR (2018)	Columbus, OH
Minneapolis, MN (2020)	Indianapolis, IN
Saint Paul, MN (2020)	Jacksonville, FL
Denver, CO (2021)	Milwaukee, WI
	Nashville, TN
	Philadelphia, PA
	Phoenix, AZ
	Sacramento, CA
	Salt Lake City, UT
	St. Louis, MO

responsiveness to changes in its Access input from Equation 0 will be visible through its employment numbers.

Table 2: *Overview of the 10 NAICS Supersectors*

NAICS Supersectors included in the study (private ownership sector)
Natural Resources and Mining
Construction
Manufacturing
Trade, Transportation and Utilities
Information
Financial Activities
Professional and Business Services
Education and Health Services
Leisure and Hospitality
Other Services

Table 3 gives summary statistics for variables in interest. Additional tables are provided in the appendix with summaries by industry (Appendix Table 8), by treated cities only (Appendix Table 9) and by controlled cities only (Appendix Table 10). It's important to note that due to limited data availability - several variables are measured in different levels. Citywide speed limit reduction, median household income, percent of workers who drove alone to work, mean travel time to work - all have county and year level data. The number of establishments and average weekly wage have county level data. The outcome variable - Difference in Local and National Employment Growth Rates in Industry has county, industry and quarter level data.

Citywide speed limit reduction is a binary variable, where 0 at a given county and year means "no citywide speed limit reduction", and 1 means "citywide speed limit reduction has been implemented". The rest of the variables are control variables. They are discussed in detail in the next section. The final cleaned dataset has 9600 observations, a panel combination of 15 years, 4 quarters, 16 cities and 10 industries.

Table 3: *Summary statistics for variables of interest*

	N	Mean	SD	Median	Max	Min
Local performance*	9,283	−0.004	0.084	−0.001	0.999	−1.202
Citywide speed limit reduction	9,600	0.104	0.305	0	1	0
Median household income	9,600	62,015.93	16,908.54	59,233.5	122,080	31,997
% workers drove alone to work	9,600	69.067	9.494	70.9	84	44.6
Number of establishments	9,600	4,557.75	5,666.21	2,733	39,832	2
Average weekly wage	9,600	1,250.78	644.40	1,169	7,203	0
Mean travel time to work	9,600	25.312	3.360	24.55	34.4	20.7

*Difference in Local and National Employment Growth Rates in Industry.

Empirical Modeling and Results

I start with running the following preliminary robust univariate regression model, where y_{cit} is the local performance of an industry — employment growth differential between local and national levels in a given quarter. x_{1ct} is the speed limit dummy variable.

$$y_{cit} = \beta_0 + \beta_1 x_{1ct} + \varepsilon \tag{1}$$

Figure 1 displays β_1 results for all industries. β_1 can be interpreted as the average percentage point change in the local-national growth rate difference associated with a citywide speed limit reduction — in other words, the average change in the local performance of an industry associated with a citywide speed limit reduction. Table 4 shows three statistically significant results for the speed limit reduction effect — in the Manufacturing, Financial Activities, and Education & Health Services industries. Their local performance worsens when citywide speed limit reductions are implemented.

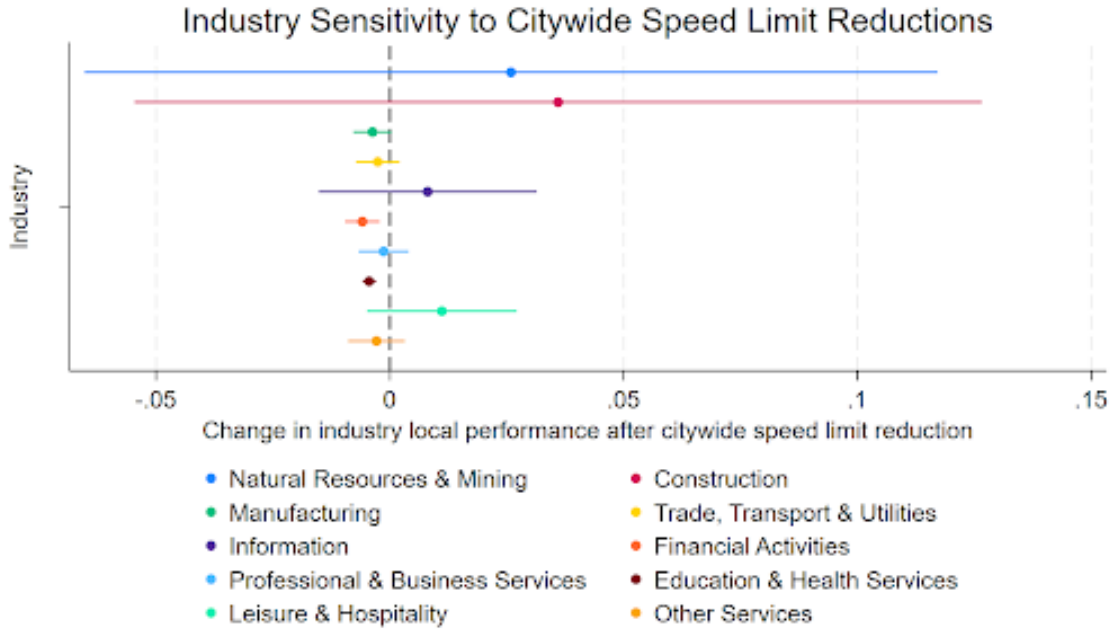
Table 4: *Results of univariate robust regression per industry*

	(1)	(2)	(3)	(4)	(5)
VARIABLES	Nat. Res.	Construction	Mfg.	Trade, Trans. & Util.	Information
Speed Limit Reduction	0.0259	0.0360	−0.00372*	−0.00258	0.00809
	(0.0428)	(0.0425)	(0.00192)	(0.00221)	(0.0109)
Constant	−0.00937*	−0.0191***	−0.00108***	−0.000344	−0.0120***
	(0.00490)	(0.00486)	(0.000203)	(0.000234)	(0.00117)
Observations	874	874	944	944	933
R-squared	0.001	0.006	0.005	0.002	0.000
Cities	16	16	16	16	16

	(6)	(7)	(8)	(9)	(10)
VARIABLES	Financial	Prof. & Bus.	Educ. & Health	Leisure & Hosp.	Other
Speed Limit Reduction	−0.00588***	−0.00132	−0.00441***	0.0111	−0.00286
	(0.00173)	(0.00250)	(0.000676)	(0.00751)	(0.00283)
Constant	−0.00136***	−0.000504*	0.000989***	−0.000236	−0.00477***
	(0.000183)	(0.000265)	(7.16e-05)	(0.000795)	(0.000302)
Observations	944	944	944	944	938
R-squared	0.011	0.000	0.007	0.005	0.000
Cities	16	16	16	16	16

Robust standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Figure 1: *Visualizing β_1 results for all industries from the univariate regression.*



Next, I run Equation 2 for the multivariate regression model that incorporates control variables as well:

$$y_{cit} = \beta_0 + \beta_1 x_{1ct} + \beta_2 x_{2cit} + \beta_3 x_{3cit} + \beta_4 x_{4ct} + \beta_5 x_{5ct} + \beta_6 (x_{1ct} \times x_{4ct}) + \alpha_{ci} + \alpha_t + \varepsilon \quad (2)$$

Where:

y_{cit}**: Local Industry Performance (Employment growth difference between a given industry locally and nationally in a quarter):** This is the outcome variable that captures the local performance of an industry. If employment in Manufacturing grew by 10% nationally, but only 5% in a given city, then the industry's local performance was lagging behind due to some local factors different from national factors.

x_{1ct}**: A dummy variable that is 0 when citywide speed limit reduction has not been implemented, and 1 when it has been implemented:** This is the main predictor variable. The goal of this study is to investigate a possible causal relationship between speed limit reductions and the employment growth differential y_{cit} — in other words, how much of the aforementioned local factors the speed limit reduction explains.

x_{2cit}**: Number of establishments in a local industry in a quarter:** This control variable is the actual size of a local industry. The same industry might grow differently in two different cities based on its current size. An initially large Leisure and Hospitality industry in one city might grow faster than a smaller one in another city because of economies of scale, more competition, more supplier networks, and greater labor availability.

x_{3cit}**: Average weekly wage in a local industry in a quarter:** Another industry level control variable. Including this prevents structural differences between industries, as

it can serve as a proxy for industry skill intensity and productivity. Industries that are high-skill (such as financial services) behave and grow differently than those with low-skill occupations (such as leisure and hospitality).

x_{4ct}**: Percent of workers driving alone to work locally in a year:** This is a proxy for how car-dependent a city is. If a city is very car dependent (e.g. 90% of workers drive to work), then speed limit reductions impacting commute times or customer accessibility will have stronger effects on industry growth than in places where many workers do not depend on cars.

x_{5ct}**: Median household income locally in a year:** This controls for the overall wealth and prosperity of a city. Wealthier cities have different consumer behavior, better roads, and possibly a more educated workforce. Those characteristics independently drive industry growth, and controlling for city prosperity removes the risk of attributing effects to the speed limit policy when they are really just the effects of being a richer or poorer city.

$(x_{1ct} \times x_{4ct})$**: Interaction term between the speed limit dummy variable and the percent of workers driving alone to work:** Hypothetically, the effect of speed limit reductions on industry growth should be larger in car-dependent cities. As one independent variable's effect on the dependent variable depends on another independent variable, an interaction term is utilized. If the coefficient of this term is negative, then cities where more people drive alone to work experience a stronger negative impact on industry growth due to speed limit reductions.

$\alpha_{ci} + \alpha_t$**: Fixed effects:** Entity and time fixed effects are also added to the model, where one entity is one industry in one county, and one time period is one quarter. These capture time-invariant and city-invariant unobserved and observed characteristics.

Table **??** summarizes the results of Equation 1 (with and without fixed effects) and Equation 2. It is notable that the regression with fixed effects and controls has a statistically significant (at the 10% level) coefficient of -0.113 for the speed limit variable. Since the outcome variable is expressed in decimal form, this result can be interpreted through percentage terms — on average, there is an 11.3 percentage point decrease in local performance when a citywide speed limit reduction policy is implemented.

Table 6 presents the results of the multivariate regression separated by industry. We see 3 industries where the coefficient of speed limit reduction is statistically significant, and only 2 at the 1% significance level. In the case of a negative coefficient, if the industry was initially overperforming locally compared to the nation, this overperformance either becomes underperformance, or stays as overperformance but in a smaller magnitude after the citywide speed limit reduction. If the local industry was already initially underperforming compared to the nation, then it means the underperformance worsens. Hence we cannot tell whether an industry shifted from underperforming to overperforming, or vice versa, by just using our coefficient estimates, but we can still say that an industry's local performance went down or went up. We have statistically significant negative effects for the Natural Resources (-0.532^{*}) and the Trade, Transportation and Utilities (-0.0743^{***}) industries. On the other hand, the Leisure and Hospitality ($+0.157^{***}$) industry is estimated to start performing better after the citywide speed limit reduction. The large drop (53.2 percentage points) in the performance of the Natural Resources industry might be explained by the initially small employment in that industry in urban areas, as the industry

Table 5: *Summary of univariate (with and without FE) and multivariate regression results*

Variables	Eq. (1) Naive OLS	Eq. (1) FE	Eq. (2) FE + Controls
Speed Limit Reduction (S)	0.00514	0.00598	-0.113^{*}
	(0.00317)	(0.00619)	(0.0669)
% driving alone to work (D)			4.61×10^{-5}
			(0.000275)
Interaction ($S \times D$)			0.00192
			(0.00116)
No. of establishments			-2.86×10^{-7}
			(4.43×10^{-7})
Average weekly wage			$7.31 \times 10^{-5**}$
			(2.91×10^{-5})
Median household income			$-1.07 \times 10^{-6**}$
			(4.12×10^{-7})
Constant	-0.00473^{**}	-0.00462^{***}	-0.0312
	(0.00211)	(0.000919)	(0.0286)
Observations	9,283	9,283	9,283
R-squared		0.000	0.057
Number of panel units	160	160	160

Standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Clustered errors included in Eq. (2).

is more prevalent in rural areas. Appendix Table 8 also shows that the standard deviation in the outcome variable for the Natural Resources & Mining and Construction industries are substantially large (0.189 and 0.136 respectively) compared to more stable industries such as Manufacturing (0.013) and Financial Activities (0.015). This likely reflects the fact that these industries have very few establishments in urban counties, and explains their large and unstable coefficients in the results seen in Figure 2, which visualizes the coefficients for the predictor variable from Table 6.

The same aforementioned industries also have statistically significant coefficient results for the interaction term. The coefficients are surprising. For example, the positive interaction coefficient for Natural Resources and Trade, Transportation and Utilities means the negative effect of the speed limit reduction gets smaller as the percentage of workers driving alone to work increases. In other words, on average, the more car dependent the city, the less sensitive the mentioned industries are to citywide speed limit reductions. These results do not necessarily make sense theoretically, as one would expect less car-dependent cities to be less sensitive to driving-related changes. Due to this, I proceed with a staggered Difference-in-Differences approach (5). A staggered DiD approach is better suited here because different cities in the treatment group received treatment in different years.

Table 7 presents the average treatment effect on the treated for each of the 10 industries. We see a new striking result — the only statistically significant coefficient remaining is the one for the Leisure and Hospitality industry. This industry, on average among the treated cities, grew by 3.8 percentage points after the implementation of citywide

Table 6: *Multivariate regression results by industry*

	(1)	(2)	(3)	(4)	(5)
VARIABLES	Nat. Res.	Construction	Mfg.	Trade, Trans. & Util.	Information
Speed Limit (S)	-0.532^{*}	-0.0321	-0.00397	-0.0743^{***}	-0.310
	(0.268)	(0.238)	(0.00721)	(0.0157)	(0.177)
% driving alone (D)	-0.00142	0.00225^{**}	-1.02×10^{-5}	0.000225	0.000115
	(0.00113)	(0.000898)	(0.000187)	(0.000158)	(0.00152)
Interaction ($S \times D$)	0.00923^{*}	0.000493	-3.25×10^{-5}	0.00127^{***}	0.00471^{*}
	(0.00474)	(0.00380)	(0.000125)	(0.000248)	(0.00263)
No. of establishments	-0.000379	-1.21×10^{-6}	-2.24×10^{-6}	$1.17 \times 10^{-6***}$	-2.20×10^{-5}
	(0.000354)	(6.27×10^{-6})	(1.51×10^{-6})	(3.50×10^{-7})	(1.80×10^{-5})
Avg. weekly wage	6.09×10^{-5}	0.000497^{***}	1.20×10^{-6}	6.93×10^{-6}	0.000106
	(5.06×10^{-5})	(0.000108)	(3.34×10^{-6})	(5.58×10^{-6})	(7.18×10^{-5})
Median HH income	-3.26×10^{-7}	$-7.42 \times 10^{-6***}$	7.47×10^{-8}	-9.41×10^{-8}	-2.40×10^{-6}
	(5.43×10^{-7})	(1.64×10^{-6})	(6.73×10^{-8})	(6.00×10^{-8})	(1.78×10^{-6})
Constant	0.0854	-0.346^{***}	-0.00297	-0.0270^{*}	-0.0454
	(0.103)	(0.0925)	(0.0182)	(0.0150)	(0.131)
Observations	874	874	944	944	933
R-squared	0.037	0.626	0.011	0.030	0.132
Cities	16	16	16	16	16

	(6)	(7)	(8)	(9)	(10)
VARIABLES	Financial	Prof. & Bus.	Educ. & Health	Leisure & Hosp.	Other
Speed Limit (S)	-0.0110	0.0624	-0.00224	0.157^{***}	0.0252
	(0.0312)	(0.0376)	(0.00321)	(0.0358)	(0.136)
% driving alone (D)	0.000582^{***}	0.000137	7.97×10^{-5}	-0.00162^{*}	0.00427^{**}
	(0.000164)	(0.000155)	(0.000114)	(0.000849)	(0.00150)
Interaction ($S \times D$)	0.000167	-0.00105	-6.55×10^{-6}	-0.00250^{***}	-0.000740
	(0.000507)	(0.000608)	(4.94×10^{-5})	(0.000607)	(0.00225)
No. of establishments	-6.18×10^{-7}	$3.21 \times 10^{-7*}$	1.31×10^{-7}	4.74×10^{-6}	$5.49 \times 10^{-6**}$
	(4.34×10^{-7})	(1.73×10^{-7})	(1.06×10^{-7})	(3.26×10^{-6})	(2.14×10^{-6})
Avg. weekly wage	6.55×10^{-7}	-4.70×10^{-6}	1.25×10^{-6}	$8.71 \times 10^{-5**}$	0.000659^{**}
	(1.40×10^{-6})	(5.72×10^{-6})	(7.72×10^{-6})	(3.57×10^{-5})	(0.000234)
Median HH income	3.58×10^{-8}	4.89×10^{-9}	-6.45×10^{-8}	$-1.52 \times 10^{-6***}$	$-6.39 \times 10^{-6**}$
	(7.18×10^{-8})	(1.06×10^{-7})	(7.00×10^{-8})	(2.48×10^{-7})	(2.46×10^{-6})
Constant	-0.0425^{**}	-0.00666	-0.00305	0.139^{*}	-0.455^{***}
	(0.0147)	(0.0139)	(0.0136)	(0.0784)	(0.152)
Observations	944	944	944	944	938
R-squared	0.039	0.012	0.010	0.046	0.415
Cities	16	16	16	16	16

Robust standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

speed limit policies ($p = 0.006$). This means that the difference between local employment growth and national employment growth in this industry increased by 3.8 percentage points after the policy change. This finding makes sense theoretically for a number of reasons. Slower speeds promote walkability and a nicer atmosphere in commercial corridor-streets in the city. Slower speeds also improve the perception of safety, reduce noise levels, and increase foot traffic. These changes make commercial corridors with restaurants and cafes more attractive and allow those businesses to expand services by accepting more consumers and placing serving tables outside on the sidewalks. Figure 3 shows a visualization of Leisure and Hospitality local performance before and after the treatment. There is visual evidence that after citywide speed limit reduction, the industry starts performing better.

Figure 2: *Visualizing β_1 results of the multivariate regression for all industries.*

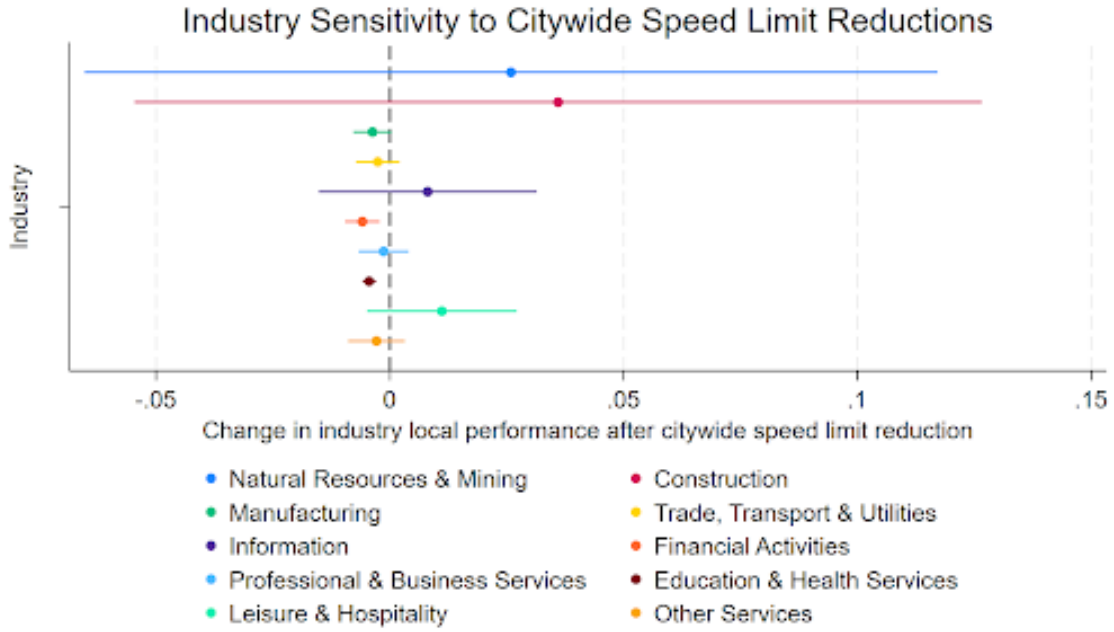


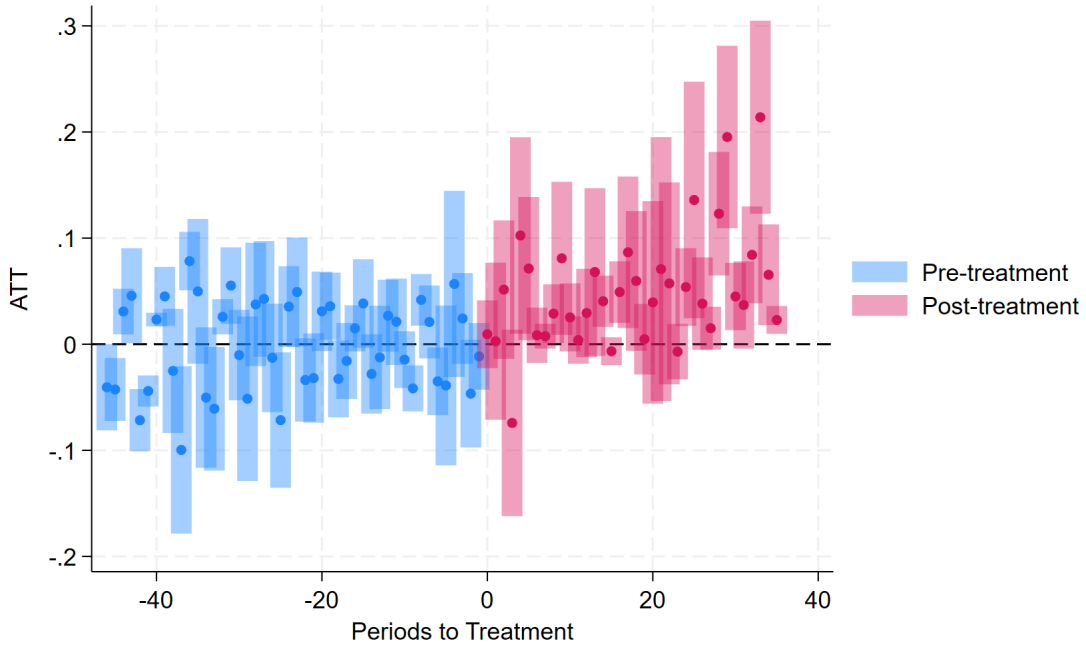
Table 7: *Industry level results of the staggered DiD approach*

Average treatment effect	Coefficient	Std. Error	z	$p > \|z\|$
Natural Resources and Mining	−0.0011	0.0253	−0.04	0.967
Construction	0.0014	0.0272	−0.05	0.959
Manufacturing	0.0120	0.0086	1.40	0.162
Trade, Transportation, and Utilities	−0.0034	0.0072	−0.46	0.642
Information	−0.0154	0.0309	−0.50	0.619
Financial Activities	0.0049	0.0039	1.24	0.215
Professional and Business Services	−0.0032	0.0063	−0.50	0.614
Education and Health Services	0.0007	0.0029	0.25	0.803
Leisure and Hospitality	0.0389**	0.0141	2.76	0.006
Other Services	−0.0264	0.0171	−1.54	0.123

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Clustered errors included.

I run two robustness checks for the above result. First, I try a staggered Difference-in-Differences with an alternative command package in my statistical software. Growth in Leisure and Hospitality is still maintained ($p = 0.022$). Additionally, I run a placebo test for the Leisure and Hospitality industry by creating a fake treatment where a "treatment" happens 1 year before the actual speed limit reduction. There was no statistically significant difference ($p = 0.167$ for the average treatment effect, compared to $p = 0.006$ when the treatment was assigned at the actual speed limit reduction time).

Figure 3: *Before-after of Leisure and Hospitality's local performance. 40 periods is 10 years, as one period is one quarter.*



Discussion and Limitations

Using the data I have, it is exciting to find that the Leisure and Hospitality industry generally benefits from a citywide speed limit reduction, and that there is no statistically significant evidence of any industry being harmed as a result of the policy. This paper does not try to argue that a city should implement citywide speed limit reductions to boost the Leisure and Hospitality industry, particularly because there are challenging limitations that call for a deeper analysis of this topic in the future.

One fundamental limitation stems from the treatment and control groups. Not only is the treatment group small in size, the cities in the group are not random. The treated cities tend to be cities that are actively implementing other road safety measures as well — bike lanes, speed bumps, speed cameras — making them systematically different from cities that do not focus on road safety to the same extent. On the other hand, it is difficult to argue that cities in the control group have no speed-reducing measures at all. The use of speed bumps, speed cameras, and other measures that slow effective speeds is not controlled in the study, leading to an omitted variable bias. Additionally, in Tables 5 and 6, we can see that the treated and controlled cities are similar to each other in terms of mean travel time to work, but are quite different in percent of workers who drove alone to work. In future research, the selection of control cities should be done more carefully to ensure parallel trends and confident comparability between the treated and the untreated.

The other important drawback is the geographic and temporal approximation used in the analysis. Due to limited data, I have used counties as proxies for cities. In many

cases the county boundaries extend beyond the city limits, meaning that a portion of the data captured by the QCEW is located outside the city and not directly exposed to the citywide speed limit reductions. Additionally, in my data, the local performance is captured quarterly, while some other control variables are only measured annually. Annual values from the ACS were assigned uniformly to all 4 quarters in a given year, which has reduced true within-year variation of industry performance.

Lastly, for cities that were treated fairly recently, in 2020 and 2021, this study only sees the short-run effects. It would be reasonable to assume that industries take many years to show responsiveness (if any) to citywide speed limit reductions. Further research should investigate how much of the growth in the local performance of the Leisure and Hospitality industry was influenced by other external factors, such as the regrowth of the economy after the pandemic — this is especially important for the cases of Minneapolis, Saint Paul, and Denver, where the policy change happened in 2020 and 2021. Moreover, additional placebo tests covering further time periods might also strengthen or weaken the results.

Conclusion

This paper investigated whether citywide speed limit reductions lead to local industry performance changes across the 10 NAICS supersectors in 5 treated and 11 control US cities over the period 2010–2024. Using a staggered Difference-in-Differences approach, I find that the only statistically significant change is seen in the Leisure and Hospitality industry, where there is a positive effect on its local performance. This industry's local performance compared to the national trend improved by approximately 3.8 percentage points following citywide speed limit reductions. This finding makes sense generally, since slower streets improve pedestrian safety perception, reduce noise and emissions, increase foot traffic, and allow restaurants and other hospitality businesses to expand onto the sidewalks — all of which stimulate growth for the industry.

For the remaining 9 industries, I find no statistically significant evidence of either positive or negative effects on local performance. Concerns that speed limit reductions would harm vehicle-dependent industries such as Trade, Transportation & Utilities are not supported by the data, at least within this study. Policymakers who reason that their city is similar to the cities involved in this study and that results can be generalized should not expect significant disruption to local industry performance as a consequence of implementing citywide speed limit reductions to prevent fatal car accidents.

There are several limitations to be considered. The treatment group of cities is small and not random. There is potential omitted variable bias due to not controlling for the implementation of other road safety measures in both treated and control cities. There are problematic geographical and temporal approximations used while merging different datasets. Additionally, we only see short-run effects for several cities in the treatment group, while it is possible that industries' response to speed limit policy changes takes many years. Future research on this particular topic can consider other forms of economic performance that might be more directly causally related with speed limits. Research can also try identifying responsiveness to speed limit changes in more specific NAICS defined sectors, rather than the 10 supersectors used in this study.

References

[1] Asadi, M., Ulak, M. B., Geurs, K. T., Weijermars, W., Zantema, K., and Van Veenen, E. (2025). Integrating traffic safety and accessibility: Spatial equity impacts of city-wide speed limit reduction in Amsterdam. *Journal of Transport Geography*, 128:104391.

[2] Borowska-Stefańska, M., Felcyn, J., Gałuszka, M., Kowalski, M., Majchrowska, A., and Wiśniewski, S. (2023). Effects of speed limits introduced to curb road noise on the performance of the urban transport system. *Journal of Transport & Health*, 30:101592.

[3] Brink, M., Mathieu, S., and Rüttener, S. (2022). Lowering urban speed limits to 30 km/h reduces noise annoyance and shifts exposure–response relationships: Evidence from a field study in Zurich. *Environment International*, 170:107651.

[4] Brown, S. (2024). Government delivers sensible approach to speed limits. New Zealand Government.

[5] Callaway, B. and Sant'Anna, P. H. C. (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics*, 225(2):200–230.

[6] Davis, A. and Jones, S. (2022). The value of prevention – casualty reductions in switching from a 30mph to 20mph speed limit default in Wales. Road Safety Knowledge Centre.

[7] Fridman, L., Ling, R., Rothman, L., Cloutier, M.-S., Macarthur, C., Hagel, B., and Howard, A. (2020). Effect of reducing the posted speed limit to 30 km per hour on pedestrian motor vehicle collisions in Toronto, Canada — a quasi experimental, pre-post study. *BMC Public Health*, 20(1):56.

[8] Friedman, L. S., Hedeker, D., and Richter, E. D. (2009). Long-term effects of repealing the national maximum speed limit in the United States. *American Journal of Public Health*, 99(9):1626–1631.

[9] Int Panis, L., Beckx, C., Broekx, S., De Vlieger, I., Schrooten, L., Degraeuwe, B., and Pelkmans, L. (2011). PM, NOx and CO2 emission reductions from speed management policies in Europe. *Transport Policy*, 18(1):32–37.

[10] Jiao, B., Kim, S., Hagen, J., and Muennig, P. A. (2019). Cost-effectiveness of neighbourhood slow zones in New York City. *Injury Prevention*, 25(2):98–103.

[11] Karndacharuk, A. and McTiernan, D. (2019). Implementation principles for 30 km/h speed limits and zones. *Journal of Road Safety*, 30(2).

[12] Madireddy, M., De Coensel, B., Can, A., Degraeuwe, B., Beusen, B., De Vlieger, I., and Botteldooren, D. (2011). Assessment of the impact of speed limit reduction and traffic signal coordination on vehicle emissions using an integrated approach. *Transportation Research Part D: Transport and Environment*, 16(7):504–508.

[13] Rowland, T. and McLeod, D. (2017). Time and fuel effects of different travel speeds. Technical Report 582.

[14] Taylor, M. (2000). Network modeling of the traffic, environmental and energy effects of lower urban speed limits. ResearchGate.

[15] Weisbrod, G., Vary, D., and Treyz, G. (2003). Measuring economic costs of urban traffic congestion to business. *Transportation Research Record*, 1839(1):98–106.

[16] Woolley, J. E., Zito, R., Dyson, C. B., Stazic, B., and Taylor, M. A. P. (2002). Impacts of lower speed limits in South Australia. *IATSS Research*, 26(2):6–17.

[17] Yannis, G. and Michelaraki, E. (2025). Effectiveness of 30 km/h speed limit – a literature review. *Journal of Safety Research*, 92:490–503.

Appendix

Table 8: *Summary statistics by NAICS supersector*

	N	Mean	SD	Median	Max	Min
Natural Resources and Mining						
Local performance	874	−0.006	0.189	0.003	0.999	−1.202
Number of establishments	960	134.879	144.112	73.5	534	2
Average weekly wage	960	1,266.10	941.92	1,069.5	6,678	0
Construction						
Local performance	874	−0.015	0.136	0	0.152	−1.096
Number of establishments	960	3,248.60	3,034.52	2,069	13,237	379
Average weekly wage	960	1,236.88	437.58	1,268	2,462	0
Manufacturing						
Local performance	944	−0.001	0.013	−0.001	0.104	−0.113
Number of establishments	960	1,566.06	1,392.79	955.5	6,712	434
Average weekly wage	960	1,389.90	303.29	1,352	3,871	863
Trade, Transportation and Utilities						
Local performance	944	−0.001	0.014	0	0.058	−0.191
Number of establishments	960	8,478.35	6,981.05	6,068.5	31,596	1,327
Average weekly wage	960	1,007.88	249.20	948.5	2,461	670
Information						
Local performance	933	−0.011	0.104	−0.003	0.417	−1.02
Number of establishments	960	1,022.08	908.19	620.5	4,223	138
Average weekly wage	960	1,906.66	865.53	1,691	7,203	0
Financial Activities						
Local performance	944	−0.002	0.015	−0.002	0.143	−0.085
Number of establishments	960	4,640.56	3,922.27	3,212.5	19,318	681
Average weekly wage	960	1,771.68	568.51	1,655	5,064	988
Professional and Business Services						
Local performance	944	−0.001	0.017	−0.001	0.254	−0.073
Number of establishments	960	9,818.49	8,365.03	6,764.5	39,832	1,430
Average weekly wage	960	1,516.00	420.88	1,462	4,270	850
Education and Health Services						
Local performance	944	0.001	0.014	0	0.199	−0.058
Number of establishments	960	7,783.61	6,924.86	4,750.5	38,607	824
Average weekly wage	960	1,065.80	178.49	1,033.5	1,724	703
Leisure and Hospitality						
Local performance	944	0.001	0.042	0.002	0.251	−0.273
Number of establishments	960	4,078.35	3,110.78	3,098	14,401	984
Average weekly wage	960	548.70	150.82	524	1,623	317
Other Services						
Local performance	938	−0.005	0.078	0	0.159	−1.007
Number of establishments	960	4,806.55	4,589.60	3,095.5	25,224	759
Average weekly wage	960	798.18	187.77	772	1,385	0

Table 9: *Summary statistics for the treated counties*

	N	Mean	SD	Median	Max	Min
Local performance	1,000	−0.002	0.038	−0.002	0.328	−0.234
Citywide speed limit reduction	1,000	1	0.000	1	1	1
Median household income	1,000	90,891.12	14,492.12	89,675	122,080	70,518
% workers drove alone to work	1,000	57.424	4.743	58.3	65.7	46.3
Number of establishments	1,000	5,191.45	5,112.18	3,656.5	34,823	13
Average weekly wage	1,000	1,736.10	942.87	1,591.5	7,203	527
Mean travel time to work	1,000	25.104	3.145	24.9	30.2	20.7

Table 10: *Summary statistics for controlled cities*

	N	Mean	SD	Median	Max	Min
Local performance	8,283	−0.004	0.088	0	0.999	−1.202
Citywide speed limit reduction	8,600	0	0.000	0	0	0
Median household income	8,600	58,658.35	13,656.88	56,769	99,353	31,997
% workers drove alone to work	8,600	70.421	8.967	72.8	84	44.6
Number of establishments	8,600	4,484.07	5,722.87	2,681	39,832	2
Average weekly wage	8,600	1,194.35	574.18	1,134	6,678	0
Mean travel time to work	8,600	25.336	3.384	24.4	34.4	20.8

The Promise of Designation: Long-term Impact of Opportunity Zone Designation on Employment Outcomes

Patrick Angelo Sursano

ECON 381: Introduction to Econometrics (Advisor: Amy Damon)

Introduction

Over 50 million Americans live in distressed communities characterized by high poverty rates and unemployment (7). Economic inequality remains a problem for distressed areas because of weaker employment growth compared to more prosperous areas (2; 20). These areas often face structural barriers that constrain residents' employment opportunities (18). Examples include reduced access to capital and fewer business establishments.

Policymakers advocate for place-based policies to respond to geographic disparities. Fratesi (11) defines them as policies targeting places that require specific interventions with the main goal of helping people. An example of this place-based approach is the Opportunity Zone program established through the 2017 Tax Cuts and Jobs Act. This program provides tax incentives for capital gains invested in an opportunity zone. It aims to attract capital to underinvested areas to promote local economic growth. The next section provides a more detailed description of the policy.

By the end of 2019, the Council of Economic Advisors reported that investors had directed $75 billion into opportunity zones. The largest share of these investments was in real estate, followed by the finance and insurance sector, and professional, scientific, and technical services (6). While opportunity zones have attracted billions of dollars in investment, these statistics alone do not indicate whether the influx of capital has improved employment outcomes. Thus, this paper asks: how does an opportunity zone designation affect employment outcomes? In particular, this paper examines the long-run effect of an opportunity zone designation on employment outcomes across census tracts designated as opportunity zones[4] and those eligible to be opportunity zones but not selected. I use the total number of jobs and resident workers as proxies for employment outcomes.

The paper is structured as follows: the policy description section details the specifics of the program, and the economic theory section provides a theoretical framework for understanding how opportunity zone designation leads to better employment outcomes, the empirical theory section outlines the models used to estimate to treatment effect, the data description section describes the data used and provides summary tables and figures for key variables of interest, the results section covers the outcomes from different specifications and their interpretations, and the conclusion section summarize the findings and future directions of this paper.

[3]Patrick Angelo Aguilar Sursano is a Junior majoring in Economics and Geography. Correspondence concerning this article should be addressed to psursano@macalester.edu.

[4]TODO: restore original footnote text.

Policy Description

The program provides tax benefits for investing capital gains in opportunity zones. Investors typically pay capital gains tax on the profit made from selling an asset (e.g., stocks, real estate, collectibles) that has increased in value. However, this program offers tax benefits to individuals who invest in a qualified opportunity fund (QOF). A QOF is an investment vehicle organized to invest in an opportunity zone business or property. Table 1 outlines the three tax benefits investors could receive from investing in an opportunity zone.

Table 1: *Tax Benefits in Different Scenarios*

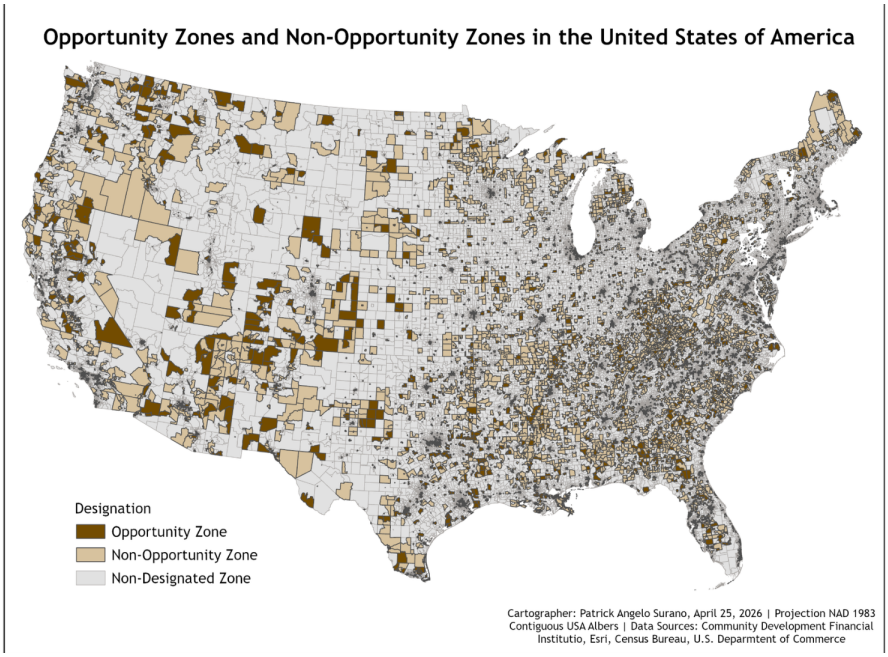
Tax Benefits	**Scenario**
Temporary tax deferral until the end of 2026 or until the asset is disposed of.	Investors holding their capital gains in a QOF within 180 days after realizing gains.
5 years: 10% of the capital gains is tax-free. 7 years: 15% of the capital gains is tax-free.	Capital gains are held in a QOF for at least 5 years or 7 years.[5]
Appreciation of QOF investments is tax-free. Investors must still pay tax on their original capital gains by 2026, but 10% or 15% of those gains are tax-free.	Capital gains are held for at least 10 years.

Upon its enactment, the U.S. Department of the Treasury selected census tracts to be designated as opportunity zones. The Treasury identified over 40,000 eligible census tracts in the U.S. and U.S. territories. Approximately 75% of these eligible tracts were considered low-income communities (LICs) because of the following criteria: (1) poverty rate of above 20% and/or (2) median household income less than 80% of the local median household income. The remaining 25% were eligible because they were contiguous to an LIC tract and had relatively low household income. Governors nominated at most 25% of the eligible census tracts as opportunity zones. In June 2018, the Treasury certified the final list of opportunity zones. Roughly 40,000 census tracts were eligible for an opportunity zone designation, but only 8,762 were designated, comprising 8,532 LIC tracts and 230 contiguous communities. Figure 1 shows the opportunity and non-opportunity zones in the contiguous U.S. The map shows that opportunity zones are distributed across states in the U.S. There are about 170 designated zones per state on average. As also shown on the map, opportunity and non-opportunity zones are almost next to each other, which could suggest clustering of economic distress. Yet, some of the census tracts were designated and some were not.

Literature Review

The U.S. government adopted place-based policies such as the Empowerment and Enterprise Zones in the mid-to-late 1990s, before the Opportunity Zone program, which

Figure 1: *Opportunity and Non-Opportunity Zones in the Contiguous US*



had mixed effects on employment outcomes. Busso et al. (5) find that neighborhoods designated as empowerment zones experienced a substantial increase in total employment and in weekly wages for zone residents working within the zone. Similarly, Hanson (15) finds the same results, where empowerment zones increased resident employment rates and reduced poverty rates. However, the instrumental variables results show otherwise: there are no statistically significant effects on either employment or poverty. This suggests that the positive estimates may reflect selection bias rather than causal effects. In contrast, evidence suggests that enterprise zones do not improve employment. Researchers examine the short- and long-run effects of the program on employment and find no consistent evidence of benefits. If anything, the long-run effects on employment are negative (21; 22). The mixed effects from previous place-based policies suggest that they can stimulate economic activity, but their effects on employment outcomes are inconsistent. While these findings provide valuable evidence for understanding earlier place-based policies, it remains unclear whether newer programs, such as Opportunity Zones, generate similar employment outcomes.

As we delve into opportunity zones, it is necessary to understand the investment patterns in these census tracts. Research shows that investment has increased but remains spatially and sectorally concentrated. Early evidence from Kennedy and Wheeler (17) suggests that opportunity zone capital is highly spatially concentrated, with only 16% of the census tracts receiving funding. Additionally, capital gravitates toward neighborhoods with relatively higher educational attainment, higher incomes, higher home values, and greater concentrations of professional and amenity services. Opportunity zone investments are also sectorally concentrated, with the highest investment in residential real estate. Feldman and Corinth (9) conclude that after almost five years of opportunity zone implementation,

private investment has focused on multifamily housing. Similar evidence shows that investors concentrate their investments in the equity of businesses related to real estate and construction, primarily targeting high-end real estate (3; 17). These papers establish that opportunity zones attract investment, but as shown, concentrated. However, these investments still have the potential to stimulate employment, which this paper seeks to examine.

Focusing on the effect of opportunity zones on employment, economists find little to no evidence that they improve employment outcomes. Freedman et al. (12) find little or no evidence of positive effects of the opportunity zone program on employment rate, average earnings, or poverty rate among residents. They employ a difference-in-differences (DiD) approach using microdata from the American Community Survey (ACS) from 2017 to 2019. Similarly, Arefeva et al. (1) examine the effects of tax incentives on business creation and employment in the first two years of the program's implementation, using establishment- and tract-level data from the Your-economy Time Series and the ACS. They find that the designation of an opportunity zone increased the employment rate by 3.0 to 4.5 percent compared to similar tracts in the metropolitan area. Interestingly, the results also show that residents outside the designated opportunity zone tracts likely filled most of the jobs created by the program. While these results show that an opportunity zone designation has little to no effect on employment outcomes, both studies cover only the first two years of implementation. I argue that the effects of new investments take time due to lags arising from regulatory scrutiny and the planning and construction process (16; 19). Additionally, transitioning from one job to another comes with financial, psychological, moving, and time costs (8; 23). People consider these factors before making decisions that could cause delays when transitioning to a job that is in an opportunity zone.

This paper is closely related to the working paper by Freedman et al. (13), as it explores the longer-run impact of the designation on labor market outcomes using ACS and Longitudinal Employer-Household Dynamics (LEHD) Origin-Destination Employment Data Statistics (LODES) microdata, and finds improvements in employment outcomes. However, I diverge by using only LODES microdata to examine changes in the total number of jobs in each census tract from 2013 to 2023, which may not be possible with publicly available ACS data, as it provides tract-level data averaged over 5 years. I also limit my analysis to the LIC census tracts only and omit some states with data gaps. Therefore, I contribute a small difference in the growing literature on opportunity zones and employment outcomes. Using my dataset, the outcome variable is the total number of jobs, serving as a proxy for employment outcomes. I then look at the number of resident workers to assess whether the opportunity zone designation benefited its residents. Ultimately, this paper seeks to evaluate the long-term impact of opportunity zones on employment outcomes.

Economic Theory

The Opportunity Zone program's effect on employment outcomes can be understood through Hall and Jorgenson (14)'s User Cost of Capital Theory and Labor Demand Theory. The former states that investment tax incentives decrease the user cost of capital, making investments more profitable and encouraging firms to increase investment. To see how tax incentives stimulate investment, I follow the model:

$$UCC = (r + \delta)P - \pi(1 - T) \tag{1}$$

where UCC is the user cost of capital, r is the real interest rate, δ is the depreciation rate, P is the price of investment, π is the expected capital gains from the investment, and T is the tax rate. Essentially, the model shows that the user cost of capital is equal to the difference between the price of investment subject to the real interest rate and depreciation rate, and the after-tax capital gains. As the tax rate decreases, the after-tax expected capital gains increase, which leads to a lower user cost of capital, holding capital gains and other factors constant. The tax advantages the program provides would increase the after-tax capital gains and lower the user cost of capital, which incentivizes investors to invest in a QOF that funnels the capital towards opportunity zones.

Theories of investment and of labor and capital inputs can explain how investment influences potential increases in employment. Even though financial investments occur instantaneously, the physical accumulation of capital occurs in the next period because it is a lagged process. When capital increases in the next period, the marginal product of labor (MPL) rises, assuming that labor and capital are complements. There are two ways to look at this in the context of the Opportunity Zone program. First, a QOF invests capital to create a new business, therefore increasing the demand for labor to manage the capital. Second, a QOF invests capital in an existing establishment, which increases the value of MPL (Price × MPL). The company benefits from increased output and revenue; therefore, it will take advantage of higher productivity by hiring more workers. All these theories help explain how tax incentives from the Opportunity Zone program can lead to better employment outcomes.

Empirical Strategy

In examining the effect of the Opportunity Zone program on employment outcomes, I use two regression models to estimate the effects of the opportunity zone designation on total jobs and resident workers, which serve as proxies for employment at the census tract level. The models' specifications are listed below:

$$\ln(EMP_{CT}) = \beta_0 + \beta_1 OZ_{CT} + \beta_2 EDUC_{CT} + \beta_3 RACE_{CT} + \epsilon_{CT} \tag{2}$$

$$RES_{CT} = \beta_0 + \beta_1 OZ_{CT} + \beta_2 EDUC_{CT} + \beta_3 RACE_{CT} + \epsilon_{CT} \tag{3}$$

where the opportunity zone designation (OZ), measured at the census tract level, is the primary regressor in both regressions. OZ is a binary variable that takes the value of 1 if a census tract is designated as an opportunity zone. It varies across census tracts (C) depending on whether they were eligible and chosen or just eligible but not chosen. OZ also varies across years (T) as the Treasury implemented the designation in 2018.

Equation (2) models the effect of opportunity zone designation on total jobs. The dependent variable is the natural log of total jobs $\ln(EMP)$, which varies across census tracts and years. LODES defines a job based on stable employment between quarters rather than a one-time snapshot. It measures total jobs using unemployment insurance payroll data to count jobs reported by employers and traced to workplace locations. I take the natural log of the total number of jobs because raw counts often have high variance, which helps reduce heteroskedasticity.

Equation (3) models the effect of opportunity zone designation on resident workers. The dependent variable is the percentage of resident workers (RES) that varies across census tracts and years. I examine this effect to see whether the designation enabled residents to find jobs in their census tracts and whether they benefited from this policy. LODES provides raw counts of both workers living in the census tract and total workers employed in the tract, which allows me to calculate the share of resident workers as the ratio of workers who both live and work in the tract over the total number of workers employed in the tract, both sourced from the same dataset.

In the regression models above, I control for factors that influence employment outcomes, such as education and race, to reduce omitted variable bias (OVB) in the estimates. The second regressor in the models ($EDUC$) is the share of workers in the tract with a bachelor's degree. The third regressor in the models ($RACE$) represents the share of minority workers. Both control variables vary across census tracts and years. Despite accounting for these control variables, a potential variable that poses OVB is the amount of investment. Investment data directly affects employment because it determines whether businesses receive capital to hire new workers. It is also correlated with the independent variable, as an opportunity zone designation would encourage more investment due to the tax benefits. Given this variable's positive relationship with the independent and dependent variables, the omission of this variable would lead to an upward bias in the coefficient.

I also employ a DiD with two-way fixed effects to analyze changes within treated census tracts relative to changes in control census tracts, using 2018 as the treatment year when all census tracts were designated as opportunity zones. The two-way fixed effects control for observable and unobservable characteristics that are time-invariant but tract-varying, and for those that are tract-invariant but time-varying, which can cause OVB. The DiD specification is in Equation (4). The interaction term ($OZ \times Post$) captures the additional change in outcome variables for opportunity zone tracts post-implementation of the program. Control variables for education, race, and fixed effects are also included for census tracts (α_C) and time in years (λ_T).

$$\ln(EMP_{CT}) \,/\, RES_{CT} = \beta_0 + \beta_1 (OZ_{CT} \times Post) + \beta_2 EDUC_{CT} + \beta_3 RACE_{CT} + \alpha_C + \lambda_T + \epsilon_{CT} \tag{4}$$

Another concern in this analysis is the violation of the independent and identically distributed data. Despite accounting for fixed effects, the error in this model can be correlated over time within the same census tracts. The economic conditions of a particular census tract in the year before are correlated with its economic conditions in the current year. Thus, I use clustering of standard errors to avoid small standard errors and inflated t-statistics.

Data Description

The data I use is the LODES microdata by the U.S. Census Bureau for the employment data. This dataset reports the number of jobs by place of work and place of residence. It also includes statistics on many worker characteristics, such as race, gender, education, income, and industry. The employment data is provided at the block level, but I aggregated it into the census tract level to match the units for opportunity zones. This

dataset is missing some observations for Alaska and Michigan from 2017 to 2023 because they were out of the LEHD partnership, resulting in these data gaps and exclusion from the analysis in this paper.

For the opportunity zone data, I use data from the Community Development Financial Institutions to identify census tracts designated as opportunity zones and LIC tracts eligible for opportunity zone designation but not selected for designation. The former serves as the treatment group in the analysis, while the latter is the control group. In this paper, I limit my analysis to census tracts within US states only, given the differences in economic and social structures between US states and US Territories. Additionally, I restrict my analysis to low-income census tracts and exclude non-low-income tracts to ensure that all tracts are comparable in baseline socioeconomic conditions.

Each employment variable was matched with the corresponding census tract and year. Upon data cleaning, I have a final sample of 16,936 census tracts with employment variables for each year spanning from 2013 to 2023, including 3,743 opportunity zones and 13,193 non-opportunity zones.

Table 2 presents the descriptive statistics for key variables of interest for the treatment (OZ) and control (Non-OZ) groups in 2017, before the Opportunity Zone program was signed. The table also shows the t-value, which tests for the difference in means between the treatment and control groups. As shown below, the average total number of jobs in opportunity zones is around 2,200, which is statistically significantly higher than the total number of jobs in non-opportunity zones at approximately 1,400. The composition of the resident and nonresident workers between the two groups is roughly the same. The percentage of workers with a bachelor's degree is nearly the same, but the values remain statistically significantly different between the groups. In both groups, people who identify as White make up the largest share of workers, followed by the Black population, Asian, and eventually Native and Pacific Islander populations.

Table 2: *Descriptive Statistics for Opportunity and Non-Opportunity Zones in 2017*

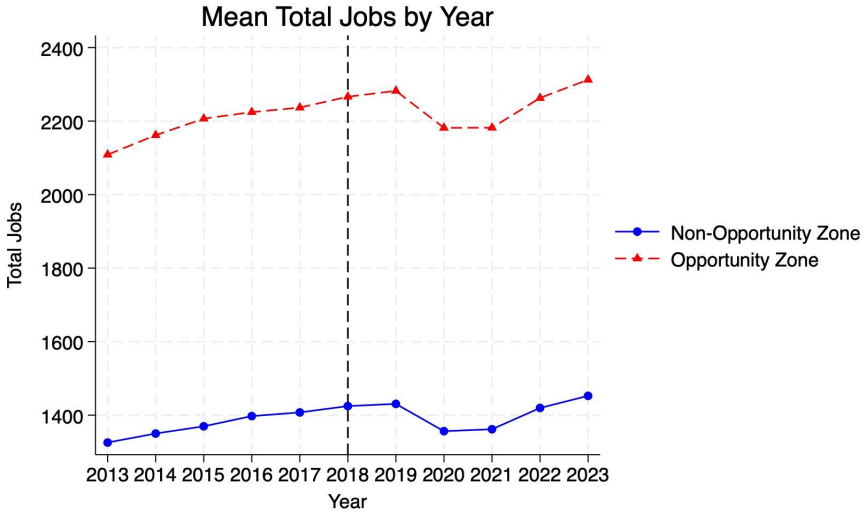
	Mean		Standard Error		t-value
Variable	OZ	Non-OZ	OZ	Non-OZ	
Total Jobs	2236.814	1407.401	62.07	26.14	−14.01
Nonresident Workers (%)	18.394	19.075	.3070	.1727	1.87
Resident Workers (%)	81.606	80.925	.3070	.1727	−1.87
Bachelor's Degree (%)	18.41	18.013	.1050	.0580	−3.24
Population, White (%)	70.914	73.297	.3246	.1661	6.67
Population, Black (%)	20.383	17.64	.3008	.1498	−8.47
Population, Asian (%)	5.454	5.925	.1209	.0755	3.02
Population, Native (%)	1.249	1.159	.0719	.0324	−1.25
Population, Pacific (%)	.221	.206	.0105	.0046	−1.38
Observations	3743	13193	3743	13193	3743

Note: t-value is the t-statistic from conducting a two-sample difference of means test.

The key variable of interest is the total number of jobs in the treatment and con-

trol groups to assess whether the opportunity zone program created more jobs. Figure 2 presents the trend graph comparing the total number of jobs from 2013 to 2023. Before implementation, opportunity zone tracts have consistently had more total jobs than non-OZ tracts. Post-2018, both groups continue to grow modestly. In 2019, opportunity zone and non-opportunity zone tracts peaked, but at different magnitudes. Again, the COVID-19 pandemic led to declines in total employment for both groups. However, the opportunity zone tracts recovered faster and have widened the gap, which may suggest a delayed effect of the opportunity zone investment.

Figure 2: *Mean Total Jobs by Year*



Another key variable of interest is the percentage of resident workers in the opportunity and non-opportunity zones, to assess whether the policy has enabled residents to benefit. Figure 3 shows the trend graph comparing the mean share of workers who live and work in the same census tract from 2013 to 2023. Before the policy's implementation, the share of resident workers was declining, and this trend continued even afterward. The figures show that both groups move similarly in total jobs and resident workers, which may suggest a parallel trend, but I cannot rule it out based on these figures alone.

Results

The regression models above try to estimate the effects of an opportunity zone designation on two employment proxies, the natural log of total jobs and the percentage of resident workers. I also included variables for education and race as control variables. To answer this question, I use two different approaches: a naive regression model and a DiD regression model with entity and year fixed effects. Table 3 presents the results from these regressions.

The naive regression model shows that a census tract designated as an opportunity zone ($OZ = 1$) expects total jobs to increase by 46%, significant at the 1% level. Meanwhile, an opportunity zone designation is associated with a decrease of 0.753 percentage points for the share of resident workers, which is also significant at the 1% level. However, these

Figure 3: *Mean Share of Resident Workers by Year*

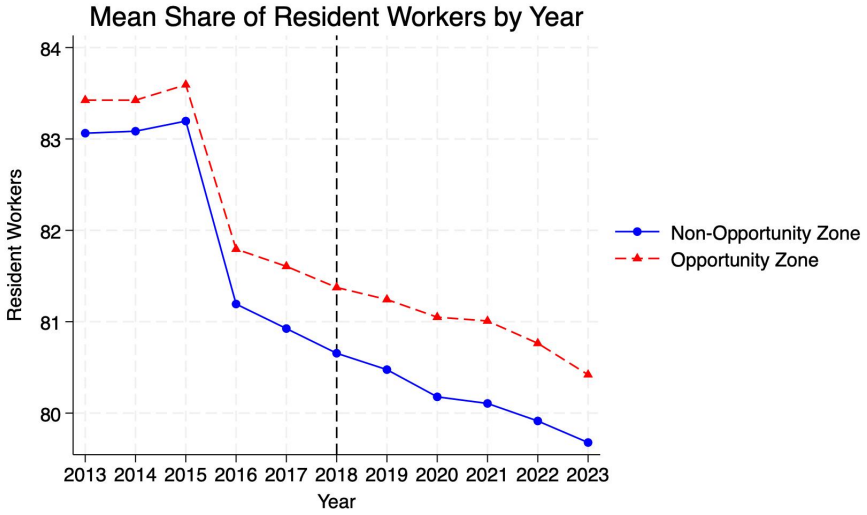


Table 3: *Results from Naive and Difference-in-Differences Models*

	(1) Naive 1	(2) Naive 2	(3) DiD 1	(4) DiD 2
VARIABLES	Total Jobs	Resident Workers	Total Jobs	Resident Workers
Opportunity Zone (OZ)	0.460***	−0.753***	−0.006	0.274***
	(0.008)	(0.130)	(0.005)	(0.060)
OZ × Post			0.004	0.144***
			(0.005)	(0.052)
Bachelors (%)	0.038***	0.497***	0.010***	0.250***
	(0.001)	(0.007)	(0.001)	(0.008)
Minority (%)	−0.015***	0.057***	0.004***	0.173***
	(0.000)	(0.002)	(0.001)	(0.009)
Constant	6.235***	70.759***	6.288***	72.024***
	(0.011)	(0.165)	(0.026)	(0.309)
Observations	186,296	186,296	186,296	186,296
R-squared	0.092	0.035	0.963	0.977
Tract Fixed Effects	No	No	Yes	Yes
Year Fixed Effects	No	No	Yes	Yes
Cluster of SE	No	No	Yes	Yes

Note: Robust standard errors in parentheses. Significance level: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$. Clustering of standard errors is employed in the DiD regression models.

models do not account for differences across census tracts and years. These estimates are also subject to selection bias because of the non-random assignment of opportunity and non-opportunity zones.

By construction, the designation was targeted at census tracts with high poverty rates and low incomes. Additionally, the governors' methods for nominating census tracts as opportunity zones were highly variable. Frank et al. (10) find that governors are more likely

to nominate a census tract as an opportunity zone when the tract's state representative is a member of the governor's political party. Therefore, the negative coefficients from the naive model can be a product of the selection bias.

The DiD specification with two-way fixed effects provides a better estimate of the treatment effect. This approach compares changes within treated census tracts to changes in the control census tracts, while controlling for time-invariant but tract-specific heterogeneity and year-specific characteristics, but tract-invariant. However, the internal validity of the estimates depends on the parallel trends assumption that, in the absence of the opportunity zone designation, treatment and control census tracts would follow the same trend over time.

Results from the DiD approach show that opportunity zones do not statistically significantly differ in the baseline total jobs ($\beta = -0.006$, $p > 0.01$) from non-opportunity zones, even accounting for tract and time fixed effects. However, opportunity zone tracts are associated with a higher baseline share of resident workers, statistically significant at the 1% level. In terms of the post-policy effect, the model indicates that an opportunity zone designation is associated with a 0.4% increase in total jobs relative to non-opportunity zones, but the effect is not statistically significant. On the other hand, results show that an opportunity zone designation is associated with a 0.144 percentage point increase in the share of resident workers after the policy implementation, and is statistically significant at the 1% level.

Given the result, I model non-resident workers to examine the effect of the opportunity zone designation on their share to check the robustness. Table 4 shows the results for both the naive and DiD models. Results indicate a statistically significant decline in nonresident workers in opportunity zone tracts, mirroring the increase observed in resident workers. This provides evidence of a potential reallocation of employment from nonresident to resident employees.

Table 4: *Results from Naive and Difference-in-Differences Models for Nonresident Workers*

	(1)	(2)
VARIABLES	Naive Nonresident Workers	DiD Nonresident Workers
Opportunity Zone	0.753^{***}	-0.274^{***}
	(0.130)	(0.060)
OZ × Post		-0.144^{***}
		(0.052)
Bachelors (%)	-0.497^{***}	-0.250^{***}
	(0.007)	(0.008)
Minority (%)	-0.057^{***}	-0.173^{***}
	(0.002)	(0.009)
Constant	29.241^{***}	27.976^{***}
	(0.165)	(0.309)
Observations	186,296	186,296
R-squared	0.035	0.977
Time Fixed Effect	No	Yes
Year Fixed Effect	No	Yes
Cluster of SE	No	Yes

Note: Robust standard errors in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$. Clustering of standard errors is employed in the DiD regression models.

Despite finding statistical significance for the effect of opportunity zones on resident

workers, these findings are not robust due to a violation of the parallel-trends assumption. Figure 4 shows the event study results for the treatment effect before the opportunity zone designation. From 2013 to 2015, the coefficients are significantly negative, and their confidence intervals do not cross zero. Additionally, the coefficient already increases in 2016, and its confidence interval barely touches zero. These results suggest that treatment and control groups were already diverging even before the designation, which violates the parallel trends assumption. Therefore, the effects of the increase in the share of resident workers in opportunity zones may stem from the continuation of the upward trend that began before the designation.

Figure 4: *Event Study Panel of the Effects of OZ Designation on Resident Workers*

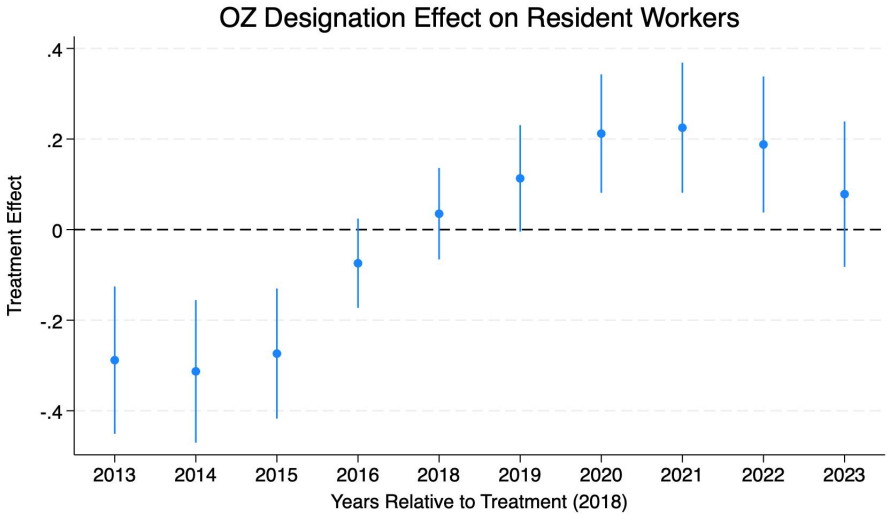


Figure 4 shows a negative pre-treatment trend in the share of residents for treated tracts relative to control tracts, which weakens the parallel trends assumption and necessitates caution in interpreting the DiD estimates. With this caveat in mind, the results tentatively suggest a small but statistically significant increase in the share of resident workers, which may suggest improved equity for zone residents. There might be a shift in the employment composition favoring residents. However, an increase in the share of resident workers may also be a result of the movement of new and more skilled workers into those census tracts, who may be more qualified for the jobs. Brazil and Portier (4) find that gentrifying census tracts are more likely to be eligible for an opportunity zone designation than non-gentrifying tracts. Additionally, a Washington, DC, case study reveals that gentrification in OZ-designated tracts has led to an increase in the in-migration of higher-income residents and the displacement of lower-income residents (**?**). The eligible census tracts for designation were already experiencing gentrification. Additionally, the influx of investment in distressed areas could also have intensified it, especially since most of the investment goes to real estate, a primary driver of gentrification. These new residents may benefit more from the program than existing or displaced residents, which defeats the purpose of improving economic conditions for these distressed areas experiencing persistent geographic disparities.

Limitations

This paper has limitations in its analysis that are worth acknowledging. First, LODES data do not provide comprehensive coverage for some demographic groups, such as Hispanics, who are categorized by ethnicity rather than race. Additionally, Alaska and Michigan are excluded from the analysis due to data gaps. Second, the DiD statistical model depends on the parallel trends assumption that may not fully hold in this context. As shown in Figure 4, a negative pre-treatment trend exists before the implementation of the Opportunity Zone program. This divergence may reflect the fact that eligible tracts for opportunity zones were already on a different trajectory prior to the designation, which makes it challenging to attribute the post-treatment effect solely to the program. Third, while I controlled for education and race, variables such as investment, urbanity or rurality, and gentrification status could potentially pose OVB. Finally, the analysis excludes non-LIC contiguous census tracts that may offer additional avenues for analyzing and interpreting the policy's effect.

Conclusion

This paper aims to examine how an opportunity zone designation affects employment outcomes over the long term. Using census-tract-level opportunity zone and employment data, I find that the designation has no significant effect on the number of jobs in a census tract. In contrast, the share of resident workers mildly increased in census tracts designated as opportunity zones. Yet this result does not hold because it violates the parallel trends assumption. The findings in this paper suggest that, even in the long run, opportunity zones have little to no effect on employment outcomes, consistent with prior literature on short-term employment effects.

Gentrification is a potential explanation for the increase in resident workers, even before and after the implementation of the treatment. Therefore, further research into the potential impacts of opportunity zones on demographic shifts and their influence on employment outcomes is a reasonable direction. Additionally, the Treasury is selecting a new list of opportunity zones as the program becomes permanent, rather than terminating it by the end of 2026 as originally planned. Thus, future research on the effect of new opportunity zone designations on employment would help assess whether improvements have been made to ensure the policy fulfills its promise to help Americans in distressed communities.

References

[1] Arefeva, A., Davis, M. A., Ghent, A. C., and Park, M. (2023). The effect of capital gains taxes on business creation and employment: The case of opportunity zones. SSRN Scholarly Paper 3645507, Social Science Research Network.

[2] Austin, B., Glaeser, E., and Summers, L. (2018). Jobs for the heartland: Place-based policies in 21st-century america. *Brookings Papers on Economic Activity*, pages 151–232.

[3] Bekkerman, R., Cohen, M. C., Liu, X., Maiden, J., and Mitrofanov, D. (2024). The impact of the opportunity zone program on residential real estate. *Manufacturing & Service Operations Management*, 26(6):2142–2159.

[4] Brazil, N. and Portier, A. (2022). Investing in gentrification: The eligibility of gentrifying neighborhoods for federal place-based economic investment in U.S. cities. *Urban Affairs Review*, 58(5):1234–1276.

[5] Busso, M., Gregory, J., and Kline, P. (2013). Assessing the incidence and efficiency of a prominent place based policy. *American Economic Review*, 103(2):897–947.

[6] Coyne, D. and Johnson, C. E. (2022). Use of the opportunity zone tax incentive: What the tax data tell us. SSRN Scholarly Paper 4308263, Social Science Research Network.

[7] Economic Innovation Group (2025). 2025 distressed communities index. Retrieved March 1, 2026.

[8] Edwards, K. A. (2022). Worker mobility in practice: Is quitting a right, or a luxury? SSRN Scholarly Paper 4567457, Social Science Research Network.

[9] Feldman, N. and Corinth, K. (2023). The impact of opportunity zones on commercial investment and economic activity. SSRN Scholarly Paper 4086056, Social Science Research Network.

[10] Frank, M., Hoopes, J. L., and Lester, R. (2022). What determines where opportunity knocks? political affiliation in the selection of opportunity zones. *Journal of Public Economics*, 206:104588.

[11] Fratesi, U. (2025). Two definitions of place-based policies: American and European approaches to fiscal equalization and regional development. *Global Challenges & Regional Science*, 1:100006.

[12] Freedman, M., Khanna, S., and Neumark, D. (2023). JUE insight: The impacts of opportunity zones on zone residents. *Journal of Urban Economics*, 133:103407.

[13] Freedman, M., Kouchekinia, N. A., and Neumark, D. (2025). Opportunity zones and economic outcomes. Working paper, National Bureau of Economic Research.

[14] Hall, R. E. and Jorgenson, D. W. (1967). Tax policy and investment behavior. *The American Economic Review*, 57(3):391–414.

[15] Hanson, A. (2009). Local employment, poverty, and property value effects of geographically-targeted tax incentives: An instrumental variables approach. *Regional Science & Urban Economics*, 39(6):721–731.

[16] Hawkins, R. G. and Lapham, I. (1972). Investment time lags. RBA Research Discussion Papers.

[17] Kennedy, P. and Wheeler, H. (2021). Neighborhood-level investment from the U.S. opportunity zone program: Early evidence. SSRN Scholarly Paper 4024514, Social Science Research Network.

[18] Kline, P. and Moretti, E. (2013). People, places and public policy: Some simple welfare economics of local economic development programs. Working Paper 19659, National Bureau of Economic Research.

[19] Millar, J. N., Oliner, S. D., and Sichel, D. E. (2012). Time-to-plan lags for commercial construction projects. Finance and Economics Discussion Series, Article 2012–34.

[20] Moretti, E. (2013). *The New Geography of Jobs*. Mariner Books/Houghton Mifflin Harcourt, 1st mariner books edition.

[21] Neumark, D. and Kolko, J. (2010). Do enterprise zones create jobs? Evidence from California's enterprise zone program. *Journal of Urban Economics*, 68(1):1–19.

[22] Neumark, D. and Young, T. (2020). Heterogeneous effects of state enterprise zone programs in the shorter run and longer run. Working Paper 27545, National Bureau of Economic Research.

[23] Ransom, T. (2022). Labor market frictions and moving costs of the employed and unemployed. *Journal of Human Resources*.