The MIT Undergraduate Journal of Economics

Volume XXI

2021-2022

The MIT Undergraduate Journal of Economics Volume XXI

2021-2022

Mailing Address:

The MIT Undergraduate Journal of Economics
Massachusetts Institute of Technology, Building E52-301
Cambridge, MA 02139

Foreword

"Money… must always be scarce with those who have neither wherewithal to buy it, nor credit to borrow it."

- *Adam Smith*

As MIT undergraduate economics students progress through their coursework, they are continuously introduced to new economic topics, constantly learning the ideas and models of established economists, and relentlessly being challenged to think differently about the observable phenomena around them. It is this enthusiasm for learning that led undergraduates at MIT to proceed in their own research—to experience the excitement of asking a question and striving to answer it. We hope that this year's papers highlight the vigor with which our undergraduate students pursue economic research and the rigor with which they present their ideas.

The publication of this Journal is made possible by the support of many people. We especially thank Professor Dave Donaldson for selecting the articles for this year's publication.

These relevant student papers demonstrate the enduring importance of rigorous economic research in the days ahead.

The MIT Undergraduate Journal of Economics Volume XXI

2021-2022

Contents

Effects of the Kansas Tax Reform of 2012 on Establishment Growth: Evidence from a Synthetic Controls Method

Aaron Lu

December 2021

Abstract

In 2012, Kansas Governor Sam Brownback passed arguably one of the largest state tax cut policies in the United States, cutting income taxes across the board. The intention was to boost the state's economic activity by keeping more money on the supply side. However, the general consensus was that the tax cuts failed to stimulate the economy, and thus were repealed in 2017. This paper aims to provide evidence of economic effects on business establishment growth from the tax policy's enactment in 2012. Using a synthetic controls method with data from 2003 to the policy's repeal in 2017, we find that there is statistically significant evidence that the tax policy harmed business establishment growth by around 3.3% per year relative to its neighbors.

1 Introduction

The Kansas tax cuts of 2012 was an unprecedented sweeping tax reform, nearly cutting all state corporate taxes to zero. Specifically, in addition to adjustments in personal income taxes, the tax policy no longer taxed pass-through income, which includes any income that sole proprietorships, limited liability companies, and S corporations pass onto their owners. This was a relatively extreme test of supply-side economic theory which argues that tax cuts lead to more work incentives and create more businesses, resulting in more jobs and higher economic activity (Canto, Joines, & Laffer, 1983). Many studies have been done on the effects of taxation on economic activity (Gale, Krupkin, & Rueben 2015, Reed 2008), but the topic is still largely debated and no clear consensus is in sight.

With the 2017 Tax Cuts and Jobs Act cutting corporate taxes down to 21% from 35%, it seems as crucial as ever to study if lowering corporate taxes is beneficial for businesses. Historically, there have been many tax incentives in favor of small businesses as they are seen as a driving force in the economy. So, it seems critical to study how tax cuts for small businesses in particular affect their business growth and the overall economic activity. But as with most economic policies and questions, it is difficult to find an appropriate experiment/data to provide analysis into economic effects. There are numerous reasons both from a theoretical and practical standpoint which make studies biased and/or inconclusive. However, the Kansas tax cuts are seen as one of the cleanest tax cut experiments on economic growth (Gale 2017). As such, I will be conducting a synthetic controls method of analysis to measure the effect of the tax cut on business growth.

The question I will be trying to answer in this research project is how did the Kansas tax cuts of 2012 affect business formation. Specifically, I will be defining business formation

as the formation of new sources of economic activity, which may be new establishments and expansions of labor associated with these establishments (not necessarily just the formation of new companies). The data I will be using will be from both the US Census (specifically the Business Dynamics Study) and the Federal Reserve Economic Data database.

For my empirical analysis, I will be using a synthetic controls method to produce a counterfactual Kansas as a control group. Then, the target treatment effect will just be the effect on Kansas minus this synthetic group. I will be using data from 2003 to 2012 to set up a synthetic control group, and I will use the policy's full range from 2012 to 2017 to measure the average treatment effect. In setting up the control group, I will be using the covariates log number of firms, log of state gdp in millions, and the state unemployment rate. There is a significant amount of noise in calculating the synthetic control, but after adjusting the averages to null out any pre-treatment effects we find that the resulting treatment period has statistically significant differences between Kansas and the control. In short, my analysis suggests that the tax policy caused a decrease of around 3.3% in establishment growth in Kansas relative to its neighboring states.

A decent amount of research has already been done on the topic, with the general consensus being that the Kansas tax cut was unsuccessful in jump-starting the economy as intended. State revenues plummeted and many public programs struggled to stay alive. According to the Center on Budget and Policy Priorities think tank, job creation and economic output severely lagged behind its neighboring states and the entire US as a whole. Labor force participation and new pass-through business formation (which is not quite the same as what I am studying) did not significantly improve and stayed in the middle relative to its neighbors. These are all factors pointing to the tax cut failing as a policy to stimulate economic growth.

More directly related to my research, there is a paper (Debacker et al. 2017) that performs rigorous economic analysis on the impact of the Kansas state tax policy on pass-through businesses. The paper finds that the primary result of the policy were the recharacterizations of reporting income rather than the formation of physical establishments. They run estimates of the change in establishments forming and exiting, but only from 2012 to 2014 and find that they do not have any statistical significance versus its neighbors. My research will be building upon that in that I will be using more data (from 2012 to 2017) and will also be studying the effects from a synthetic controls perspective. My research will also focus on just the establishment formations and exits rather than personal income or other more macroeconomic factors. To my knowledge, this is the first paper providing a rigorous analysis of business establishment data surrounding the Kansas tax cut.

The first area of research my paper aims to contribute to relates to studies on the supply-side tax policies - specifically studies on the Kansas tax cut. Other research centering on this tax policy (Debacker et al. 2017, Turner & Blagg 2017) focus on the effects of this policy on personal income and employment. Both papers find that the policy did not spur economic growth. My research on business establishment growth provide additional evidence that the policy did not jump-start economic activity as intended, as Kansas had a statistically significant negative change in establishment growth following the tax policy relative to the controls. Further research can be done on the corporate side by studying firm specific changes following the tax cut and looking at effects across different industries (all industries are aggregated together for my paper).

The second area of research my paper aims to contribute to is the general methodology of synthetic controls. Synthetic controls have recently become much more popular in the econo-

metrics field, but one major issue that arises when working with economic data is the lack of "clean" data, specifically to estimate the weights for the synthetic controls. Other papers (Hayes 2017, Abadie, A., Diamond, A., & Hainmueller, J. 2010) have used synthetic controls to study economic policies, and their synthetic control is largely accurate in replicating the treatment group pre-treatment. My paper aims to provide a method in utilizing synthetic controls when the synthetic control calculated has significant noise in the pre-treatment period. In particular, my paper demonstrates a method when the general trend between the synthetic control and the treatment group is similar, but variation in the estimates lead to deviations resulting in a positive treatment effect in the pre-treatment period. My method averages and nullifies the noise before the treatment takes place and subsequently adjusts the calculated average treatment effect by the pre-treatment average. This leads to a more accurate point estimate for the mean of the effect, but can lead to a higher standard error than if no adjustment was made (which is expected due to more variance incorporated into the final estimate).

The remainder of the paper is structured as follows. Section 2 gives more background on the Kansas tax policy and its intended effects. Section 3 discusses the data sources I used, and section 4 describes the empirical methods and tests I performed in order to investigate this topic. Section 5 covers the results and implications of my findings. Finally, I conclude in section 6.

2 Background

The Kansas Tax Reform of 2012, which cut taxes on pass-through income to zero, was one of the most radical tax policies passed in the state's history. During that time period, many believed that Kansas's economy was lacking behind its neighbors after the 2008 financial

crisis. As such, many conservatives advocated for a supply-side economic stimulant, and pushed for large tax cuts. They believed that a large tax cut could "boost investment, raise employment, and jump-start the economy" (Gale, 2017). Brownback signed the bill in May 2012, and the policy took effect on July 1st, 2012, applying to the 2013 tax year.

The policy had cuts on personal income taxes and a reduction in the number of tax brackets. Specifically, the policy cut the top two tax rates of 6.45% and 6.25% down to 4.9% and reduced the bottom rate from 3.5% to 3%. This cut heavily favored the higher earning taxpayers. However, the main focus of the bill was the elimination of taxes on "pass-through" income, which is defined as any income that businesses pass onto their owners. These pass-through businesses pay no corporate income taxes and pass the tax burden onto the individual (to avoid double taxation). Prior to the tax cut, the pass-through business income tax was 7%. This business tax cut was expected to impact almost 200,000 business owners in Kansas. Supply-side advocates were enthusiastic, projecting large increases in job growth and tax revenue.

However, as time went on, the state of Kansas faced numerous budget cuts, credit downgrades, and missed state payments. The budget cuts had heavy implications on future state spending, as budget constraints forced lawmakers to tap into state reserves. Medicaid, construction projects, and education funding were all hit hard by these cuts. Furthermore, Kansas's economic activity was still seen as lagging behind the same neighboring states it aimed to surpass with this policy. It was viewed that Kansas was lagging in every major economic category: job creation, unemployment, gross domestic product, and tax revenue (Ritholtz 2017).

The tax cuts were repealed in 2017 despite Brownback's continued support for them. A

new bill was written to raise personal income taxes and reinstate the pass-through income tax. Brownback vetoed multiple attempts of this bill but in June 2017, two-thirds majority in both the House and Senate overrode his veto and signed the repeal policy into law.

Many factors were pointed to as reasons for the failure of the tax cuts. As mentioned before, the tax policy was meant to affect around 200,000 business owners, but about 330,000 entities ended up using the pass-through tax benefit. Debacker et al. (2017) finds evidence suggesting that the response to the tax policy was largely behavioral towards tax avoidance, rather than economic output. This is consistent with the general views of the tax policy, as state revenues plummeted and economic activity was seen as stalled. Conversely, proponents of the tax policy maintained their optimistic views, faulting other reasons such as a rural recession and pointing out successes such as low unemployment and an increase in small business formation. This last point is closely related to what I aim to examine in this paper.

3 Data

The first data source I will be using is the US Census. This data is publicly available and easily accessible online at data.census.gov. The US Census reports a business dynamics study from 1978-2019. This dataset is gathered through a combination of administrative and survey-collected data versus a more typical probability sample. So, the data for a given year is more reliable with a few years of buffer in collecting late filers and verifying previous years' data. For my purposes where I use data from 2003 to 2017, any inaccuracies can be assumed to be resolved as there is sufficient buffer time. I have collected the BDS data for the states Kansas, Colorado, Missouri, Oklahoma, and Nebraska. The primary variables I will be using are firms, establishments, establishments entries, and establishment exits (and their corresponding natural logs).

The second data source I will be using is the Federal Reserve Economic Data database. This data is also publicly available at fred.stlouisfed.org. The data that I will be pulling from this will be the natural log of state gdp in millions (annual, not seasonally adjusted) and the unemployment rate (monthly, seasonally adjusted). For natural log of state gdp, I have chosen to take the natural log of gdp in millions as then the scale will be similar to the other covariates for the empirical test. For the unemployment rate, to compute the annual rate I will just be taking an arithmetic average of the unemployment rate for the 12 months of a given year.

The key outcome variables that I will be looking at are the number of establishments born in the past 12 months and the number of establishments exiting in the past 12 months. The US Census defines an establishment as any fixed physical location where economic activity occurs. A single firm (company) may have one or many establishments. The data is measured from mid March so each year's of data is the past 12 months, ending March 12 of the reported year. This data is reported annually by state, and aggregates all industry sectors together which is fine for the purpose for my analysis.

4 Empirical Test

The main empirical test that I will be conducting is a synthetic controls test (Abadie et al. 2010). The states that I will be using as controls will be Kansas's neighboring/similar states, as used in previous studies on this policy. This includes Colorado, Missouri, Nebraska, and Oklahoma. The years that I will be looking at will be from 2003 to 2017, and I will be splitting them up into two groups. The first will be from 2003-2012, which is the pre-treatment period before the tax cut was implemented. Then, from 2013-2017, the tax

cut was in effect only in Kansas, and I will use this period as the treatment period. I have chosen the years to have an appropriate amount of data to calculate each period's effect (10 years for pre-treatment, 5 years of enactment).

Some assumptions that I will need to make for the test and its results to be valid is that absent the tax policy, the comparable states and Kansas would have evolved similarly over time. In other words, a major assumption is that the other states are valid controls for this test. If this is not the case, then the synthetic controls test could result in a significant treatment effect, but there may be other factors either in addition to or instead of the tax cut policy causing the effect.

In order to test if the comparable states are indeed valid control groups, I analyzed the behavior of the states before the tax policy is enacted (from 2003 to 2012). From an initial look (Figure 1), the trends in business establishment growth are similar across all of the control states and Kansas without the tax cuts. In addition, taking a look at Figure 2 and Figure 3, we see that the control states all behave similarly from 2013 to 2017, which suggests that similar behavior continues throughout the period of the study. After conducting the synthetic controls for each of the two outcome variables (establishment entries and exits), it also looks like the synthetic control behaves similarly in the pre-period to Kansas and similarly in the post-treatment to the other control states (Figure 2 and Figure 3). All of this suggests that the chosen control states are valid controls.

The covariates I will be using for computing the synthetic control weights are going to be the natural log of firms, natural log of annual state gdp in millions, and the state unemployment rate. I will be scaling the unemployment rate covariate by 100 (so 5.2% remains 5.2, not 0.052). This way, all of the variables are in a similar range to estimate the synthetic

control weights (all values are on the order of 10). One issue with this method is the sparsity of covariates for the synthetic controls group. With more relevant variables, the synthetic controls method will produce more accurate weights and thus produce a more effective control.

The synthetic controls method is appropriate for my data as I have a single treated group (Kansas) and multiple other states as controls. Building off of other papers that have validated these states as controls, the resulting synthetic control should be suitable for my analysis. The method for finding the synthetic control is going to involve solving the optimization equation where the Kansas counterfactual $Y_{1,t}^N$ is some vector of weights w_j times the other states $Y_{j,t}$.

To find the weight vector W for the synthetic control, for $t \leq t_0$ where the treatment takes place in t_0, we will be trying to solve the optimization equation:

$$\min_W ||X_1 - X_0 W||^2 \tag{1}$$

Where W is the weight vector $[w_2, w_3, w_4, w_5]^T$, X_1 is our concatenated covariates (ln firms, ln gdp millions, and unemployment rate) for Kansas, and X_0 is the covariates for the other 4 states (both over time). Our concatenated X_1 will have dimension 30x1 (10 years of the pre-intervention period and three covariates which we stack on top of each other). Similarly, X_0 will have dimension 30x4.

To solve this equation, we will use a Lasso regression. The general Lasso regression aims

to minimize the objective function:

$$(1/(2 * n_{samples})) * ||y - Xw||_2^2 + alpha * ||w||_1 \tag{2}$$

By running this regression with $y = X_1, X = X_0$, we can output the weight vector $w = W$ that minimizes the squared distance of $X_1 - X_0W$ (with a regularization term $||w||_1$). With our data, the weight vector outputted is as follows:

$$[w_{Colorado}, w_{Missouri}, w_{Nebraska}, w_{Oklahoma}] = [0.109, 0.388, 0.108, 0.351]$$

Note that the weights do not sum to 1 (they sum to approximately 0.956), but this is appropriate as the predictors used to find the synthetic control are not exactly of the same magnitude (Abadie 2021) - two of the covariates are in logs but one is in percentages. Unemployment rate ranges from around 3 to 10 while log firms and log gdp in millions are all around 11 to 12. The slight difference in scale leads us to relax the constraint that the weights must sum to 1.

Assuming the weights W hold true post intervention, then the counterfactual for $t > t_0$ will be:

$$Y_1^N = Y_0W \tag{3}$$

Where Y is our outcome variable, either establishments born or exited (Y_1 is Kansas, Y_0 is all other states). Now we can perform our estimate for the average treatment effect, aiming to estimate $\beta = Y_1 - Y_1^N$, where Y_1^N is the counterfactual of Kansas (absent of the tax policy effect).

Our final model is as follows for $t > t_0$:

$$\begin{aligned} & \beta = Y_1 - Y_1^N \\ \implies & \beta = Y_1 - Y_0 W \text{ where } W = \min_W ||X_1 - X_0 W||^2 \text{ from pre-period data} \end{aligned} \tag{4}$$

β is going to be our estimated average treatment effect on the treated (Kansas). The model will be regressed on the data from 2003-2017, with $t_0 = 2012$. After conducting this analysis, the synthetic controls model should allow us to see the average treatment effect of the Kansas tax cut in establishment formations. If there was a significant impact of the Kansas tax cut on business formations, then we should see a significant change due to the enactment of the policy.

One main concern with this method is that synthetic controls typically requires a large number of covariates. Because there are only a few states that are deemed as controls, each survey data point is annual, and there are few relevant predictors of establishment entries/exits, there is a lot of room for variability in the estimates. This is seen in our synthetic control estimate (Figure 2 and Figure 3) in that even in the pre-period, our synthetic control does not behave exactly like Kansas (although much better than a simple average or any individual state). As a result, when calculating the average treatment effect, the absolute value will be skewed as the original synthetic was not a perfect approximation of Kansas. In other words, the bare analysis shows that there exists an average treatment effect in the pre-period ($\beta \neq 0$ in pre-period), which is impossible as no treatment has been done.

To resolve this, an average of the "average treatment effect" in the pre-period is computed, and every year's treatment effect (Kansas minus the synthetic control) is adjusted by this average. This is to account for the non-zero "treatment effect" before the treat-

ment period, and after adjusting into this Corrected ATE, the point estimates of the average treatment effect in the pre-period is now averaged to be zero (which is as expected). The resulting estimates of the mean and standard deviation of the corrected ATE, post-treatment, are shown in Table 6 for log entries and log exits.

In terms of calculating a value to test our hypothesis, I have done a simple t test of whether the mean of the Corrected ATE post-treatment is 0. I will be using a normal standard error calculated from the post-treatment time range. Specifically, the standard error will be $\frac{sd}{\sqrt{n}} = \frac{sd}{\sqrt{5}}$. The resulting standard errors and significance levels are shown in Table 7. Note that the coefficients are the same as in Table 6, as I am testing whether or not these coefficients are significantly different than 0.

One acknowledgement that I must make is that because we are adjusting the average treatment effect into the Corrected ATE, our standard errors are going to be smaller than expected. This is because there is going to be unaccounted for variance in how we estimate the mean of the pre-treatment effect, which we adjust by for the Corrected ATE. Measuring the correct standard error is difficult, and may be expanded upon with more research. However, looking at the standard errors and significance levels, for the purposes of this paper we are going to assume that both mean values are likely to be rejected at a statistically significant level. This acknowledgement is mainly to note that the reported standard errors could be misleading and are likely an underestimate of the true standard error.

5 Results

Following the method outlined in the empirical test section, the synthetic control weights have been computed and our synthetic control has been created. However, as Figure 2 and

Figure 3 show, the synthetic control does not exactly mirror Kansas in the pre-treatment period, and so simply calculating our average treatment effect after the treatment period is misleading. To resolve this, I have computed an average of the "average treatment effect" in the pre-period, and subtracted it from the entire period. Table 6 shows the mean and standard deviation of the corrected average treatment effect of log entries and log exits in the post-treatment period. Table 7 shows the results of the t test using normal standard errors, and reports the corresponding significance levels.

Looking at the results, we see that the average treatment effect for both log establishment entries and exits are statistically significant. The estimate for log entries is significant at the 0.001 level, and the estimate for log exits is significant at the 0.05 level. Acknowledging the underestimated standard errors, we are still treating the point estimates as reasonably statistically significant. Looking at the point estimates, we can see that the tax policy had an average effect of decreasing establishment entries in Kansas by around 9.7% per year (interpreting the log scale), but also had an average effect of decreasing establishment exits by around 6.4%. By defining establishment growth as a simple difference of entries minus exits, we see that our results indicate a decrease of around 3.3% of establishments per year, relative to the control states.

6 Conclusion

This paper uses the synthetic controls method on data from 2003 to 2017 to examine the effects of the Kansas tax cut of 2012 on business establishment growth. In summary, we see statistically significant evidence that the tax cuts had a negative effect on the growth of business establishments by around 3.3% per year relative to Kansas's neighbors. This finding is relatively consistent with other results such as Debacker et al. (2017) and Blagg (2018)

which find the tax cuts not causing any significant growth in economic activity. These results further oppose the efficacy of supply-side trickle-down theory on economic growth. I admit that the estimates made in this research paper may end up being statistically insignificant with more analysis into precise standard errors, but from this initial research the estimates stand. Further research on both tax policy effects and synthetic controls can be used to improve upon this paper.

7 References

Abadie, Alberto. (June 2021). Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects. Journal of Economic Literature, 59 (2), 391-425.

Abadie, A., Diamond, A., & Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. Journal of the American Statistical Association, 105, 493-505.

Barro, Josh (June 27, 2014). Yes, if You Cut Taxes, You Get Less Tax Revenue. The New York Times.

Canto, V.A., Joines, D.A., & Laffer, A.B. (1983). Foundations of Supply-Side Economics. New York, NY: Academic Press.

DeBacker, J., Heim, B.T., Ramnath, S.P, & Ross, J.M. (2016). The impact of state taxes on pass-through businesses: Evidence from the 2012 Kansas income tax reform.

Gale, William G. (July 11, 2017). The Kansas tax cut experiment. Brookings Institution.

Hayes, Michael S. (2017). Effects of Kansas' Tax Reform of 2012: Evidence from a Synthetic Control Method. Rutgers University-Camden.

Mazerov, Michael (2018). Kansas Provides Compelling Evidence of Failure of "Supply-Side" Tax Cuts. Center on Budget and Policy Priorities.

Reed, R.W. (2008). The robust relationship between taxes and U.S. state income growth. National Tax Journal.

Ritholtz, Barry (March 17, 2017). Commentary: Why Sam Brownback's tax cuts failed to make Kansas thrive. Chicago Tribune.

Turner, T.M, & Blagg, B. (2017). The short-term effects of the Kansas income tax cuts on employment growth. Public Finance Review.

8 Tables and Figures

List of Tables

List of Figures

Table 1: Summary Statistics by State 2003-2017: Kansas

	(1)				
	mean	sd	min	max	count
NET ESTABLISHMENTS	82.53	602.25	-1086.00	856.00	15
LN FIRMS	10.89	0.03	10.86	10.93	15
LN ESTABLISHMENTS	11.14	0.01	11.12	11.16	15
LN ESTABLISHMENT ENTRIES	8.71	0.11	8.59	8.91	15
LN ESTABLISHMENT EXITS	8.70	0.09	8.55	8.81	15
LN GDP*	11.77	0.17	11.47	12.01	15
UNEMPLOYMENT RATE(%)**	5.13	1.04	3.63	6.93	15

*Gdp is not seasonally adjusted, calculated as log of gdp in millions

**Unemployment rate is seasonally adjusted and its annual value is an average of the reported monthly values

Table 2: Summary Statistics by State 2003-2017: Colorado

	(1)				
	mean	sd	min	max	count
NET ESTABLISHMENTS	1632.07	2219.85	-3623.00	4405.00	15
LN FIRMS	11.60	0.04	11.54	11.68	15
LN ESTABLISHMENTS	11.81	0.05	11.73	11.90	15
LN ESTABLISHMENT ENTRIES	9.71	0.08	9.59	9.86	15
LN ESTABLISHMENT EXITS	9.60	0.09	9.48	9.81	15
LN GDP*	12.48	0.17	12.19	12.76	15
UNEMPLOYMENT RATE(%)**	5.64	2.06	2.62	9.14	15

*Gdp is not seasonally adjusted, calculated as log of gdp in millions

**Unemployment rate is seasonally adjusted and its annual value is an average of the reported monthly values

Table 3: Summary Statistics by State 2003-2017: Missouri

	(1)				
	mean	sd	min	max	count
NET ESTABLISHMENTS	912.33	1882.53	-2788.00	3550.00	15
LN FIRMS	11.58	0.03	11.53	11.61	15
LN ESTABLISHMENTS	11.84	0.03	11.80	11.88	15
LN ESTABLISHMENT ENTRIES	9.57	0.10	9.41	9.70	15
LN ESTABLISHMENT EXITS	9.51	0.08	9.40	9.63	15
LN GDP*	12.46	0.12	12.23	12.64	15
UNEMPLOYMENT RATE(%)**	6.25	1.69	3.77	9.51	15

*Gdp is not seasonally adjusted, calculated as log of gdp in millions

**Unemployment rate is seasonally adjusted and its annual value is an average of the reported monthly values

Table 4: Summary Statistics by State 2003-2017: Oklahoma***

	(1)				
	mean	sd	min	max	count
NET ESTABLISHMENTS	538.00	664.25	-623.00	1558.00	15
LN FIRMS	11.08	0.01	11.05	11.10	15
LN ESTABLISHMENTS	11.33	0.03	11.26	11.36	15
LN ESTABLISHMENT ENTRIES	8.99	0.08	8.85	9.14	15
LN ESTABLISHMENT EXITS	8.92	0.05	8.83	8.99	15
LN GDP*	11.95	0.19	11.57	12.19	15
UNEMPLOYMENT RATE(%)**	4.81	0.90	3.62	6.56	15

*Gdp is not seasonally adjusted, calculated as log of gdp in millions

**Unemployment rate is seasonally adjusted and its annual value is an average of the reported monthly values

Table 5: Summary Statistics by State 2003-2017: Nebraska

	(1)				
	mean	sd	min	max	count
NET ESTABLISHMENTS	310.80	311.79	-369.00	723.00	15
LN FIRMS	10.54	0.01	10.52	10.57	15
LN ESTABLISHMENTS	10.77	0.02	10.73	10.82	15
LN ESTABLISHMENT ENTRIES	8.32	0.07	8.22	8.48	15
LN ESTABLISHMENT EXITS	8.24	0.07	8.11	8.34	15
LN GDP*	11.44	0.20	11.10	11.71	15
UNEMPLOYMENT RATE(%)**	3.63	0.62	2.92	4.70	15

*Gdp is not seasonally adjusted, calculated as log of gdp in millions

**Unemployment rate is seasonally adjusted and its annual value is an average of the reported monthly values

Table 6: Average Treatment Effect via Synthetic Controls (2013-2017)

	Log Entries			Log Exits		
	(1)			(2)		
	mean	sd	count	mean	sd	count
Corrected ATE*	-0.0970	0.0144	5	-0.0640	0.0391	5

*Corrected ATE is our synthetic controls ATE minus the mean of the pre-treatment ATE

Table 7: T-Test of Log Establishment Entries and Log Establishment Exits

	Log Entries	Log Exits
	(1)	(2)
	Corrected ATE*	Corrected ATE*
Constant	-0.0970***	-0.0640*
	$(0.0064)^{+}$	$(0.0175)^{+}$
Observations	5	5

Standard errors in parentheses

+Reported standard errors are an underestimate of the true standard errors

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

Figure 1: Net Change of Establishments Over Time by State

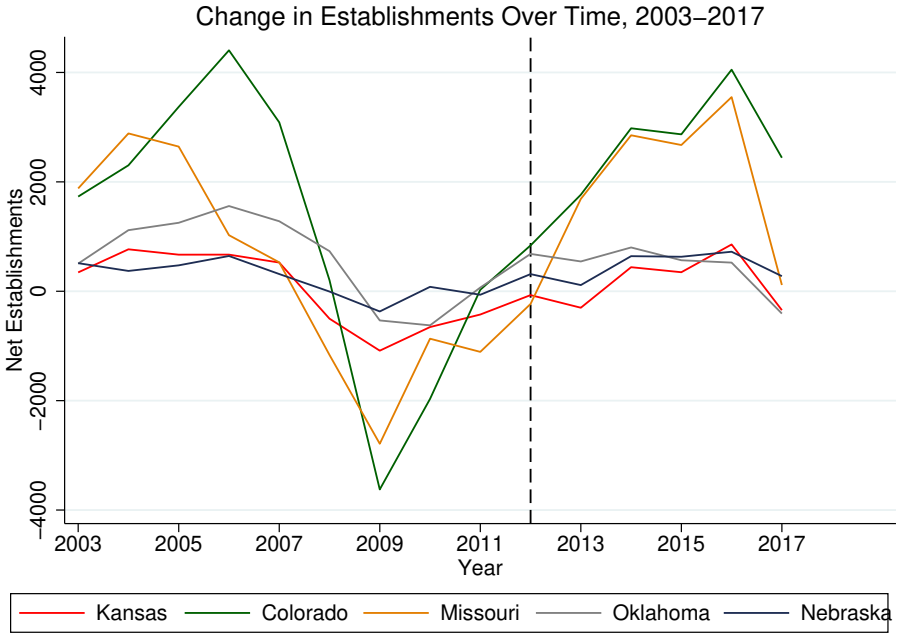
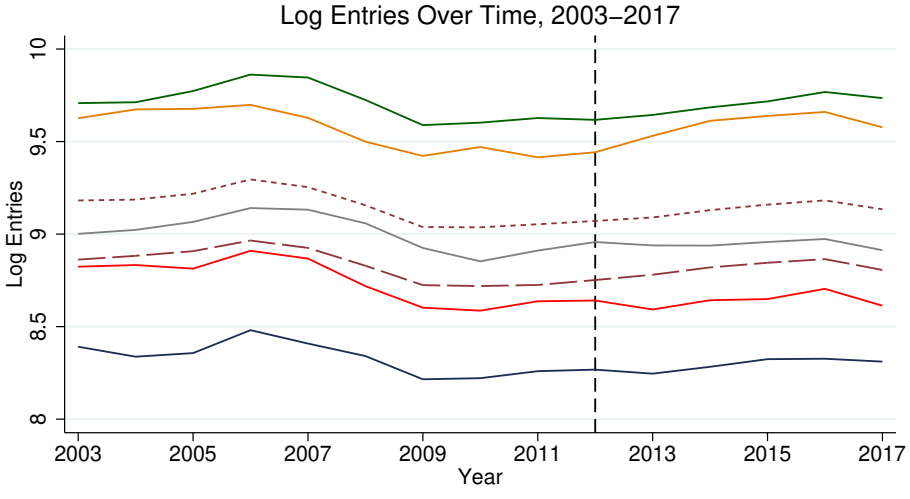
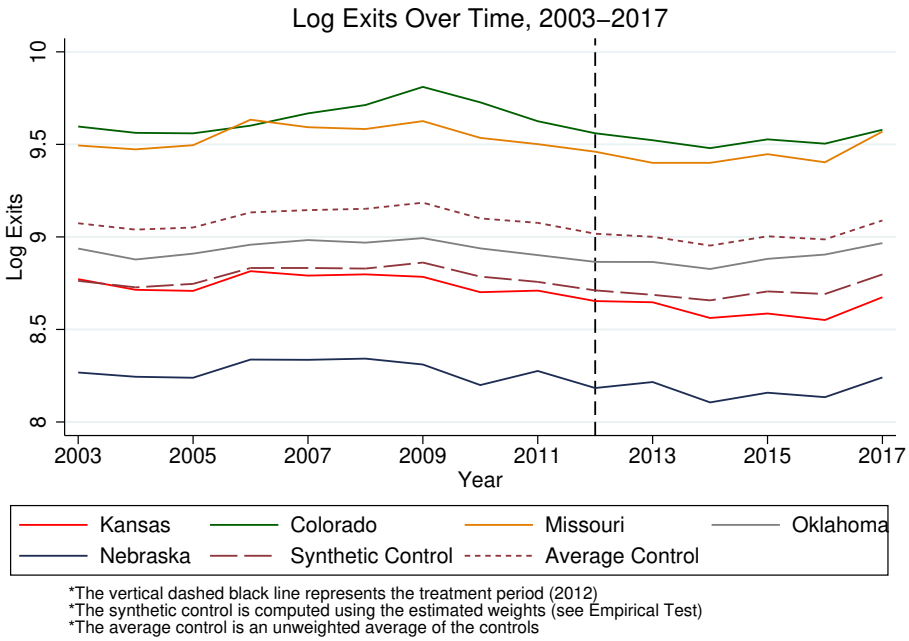


Figure 2: Synthetic Controls and Baseline Average Control versus States (Log Entries)



*The vertical dashed black line represents the treatment period (2012)
*The synthetic control is computed using the estimated weights (see Empirical Test)
*The average control is an unweighted average of the controls

Figure 3: Synthetic Controls and Baseline Average Control versus States (Log Exits)



*The vertical dashed black line represents the treatment period (2012)
*The synthetic control is computed using the estimated weights (see Empirical Test)
*The average control is an unweighted average of the controls

Impact of Mask Mandates on Yelp Review Significance for SMB Survival Rates

Giovanni Ahern | gahern@mit.edu

Abstract

What is the influence of mask mandate restrictions on the incremental effect of Yelp rating increases for restaurant survival rates? Using the self-published Yelp Dataset, this paper employs a difference-in-discontinuities empirical design based on two cities with a mask mandate treatment effect to determine the impact of an incrementally higher Yelp rating on restaurant survival rates. The goal of the experiment is to determine the extent to which digitization during Covid lent more credibility to online sources such as Yelp to facilitate restaurant discovery and decision-making. Across a range of different star cutoffs and income segmentations, I found no statistically significant evidence of a shift in the importance of Yelp ratings on survival rates before and after mandates took place. This implies coronavirus restrictions likely had less of an influence than anticipated on digital adoption, and online platforms did not gain substantially greater significance in the realm of real-world decision making. That being said, one interesting finding was a 15% increase in survival rating for low-income restaurants having a higher Yelp rating post-mandates, so income stratification could be an exciting area of future research.

I. Introduction

The coronavirus pandemic played a major role in the digitization of the economy as physical stores closed and government restrictions were put into place. One important potential confounding effect on my project related to these policies was the passage of the $953bn

Paycheck Protection Program (PPP) via the CARES Act in 2020. The PPP allowed entities, particularly small and medium sized businesses, to apply for low-interest private loans to subsidize payroll and other operational costs which could be fully forgiven if employee counts remained stable. Given this bill was introduced during the same time period as my experiment, there could be a potential confounding factor on restaurant survival rates related to CARES Act legislation as opposed to the significance of an incrementally higher Yelp star rating. Because the PPP ameliorated payroll, one of the highest restaurant operating costs, surviving after lockdowns could have been much easier than without such a program in place. Another impact of the pandemic was the rapid adoption of Yelp in the dining and entertainment space over the last 18 months. Founded in 2004, Yelp publishes crowd sourced reviews on various businesses for over 77 million unique users per month, making it one of the most popular applications on the internet. As more and more experiences transition online, especially as a result of Covid, it is important to consider the relative strength of Yelp in the restaurant discovery process and its influence on decision making. Yelp posts restaurants showing individual reviews along with a 1-5 star rating. The overall question this paper examines is the effect of restrictive government policies during the pandemic on the incremental effect of Yelp ratings for small business survival rates. While earlier papers (Luca 2016) have demonstrated that stronger Yelp reviews have a causal impact on revenue boosts for restaurants, this paper takes a different approach by discerning the relative change in review importance on survival rates once a restrictive policy is introduced and greater digital adoption takes place.

Specifically, the city of Atlanta issued a mask mandate on July 8, 2020 while Boston and Cambridge implemented an identical policy more than two months prior on May 6, 2020. Using a combination of data from the open source Yelp Dataset, I was able to determine during which

months a particular establishment was open along with an average star rating. A difference-in-discontinuities (DID) approach was utilized to specifically isolate the impact of mask mandates on the incremental effect of an additional half Yelp star on survival rates. This analysis was completed for restaurants open in 2020 before the initial Boston mandate and only includes those receiving a sufficient number of 20 reviews, corresponding to approximately 3 per month. Finally, census tract level data on metro area median household income per capita was used in a heterogeneity analysis along with a "donut" approach to compare effects for restaurants located in high and low income areas. The "donut" approach in particular selected for restaurants in the top and bottom 25th percentile in terms of median household income for their location in a given city. This could determine whether or not restaurants in low income (and likely low-traffic) areas could obtain meaningful marketing exposure without having to purchase expensive advertisements to reach customers. A potential confounding factor in this analysis has to do with the fact that income fell faster in lower versus higher income households during the pandemic despite a faster fall in spending for higher income households. Different changes in income and spending between these two groups as a result of the pandemic could have an impact on the survival rate of restaurants unrelated to a Yelp rating increase. This is especially true given that restaurant success is highly dependent on disposable income, and the disposable income of these subgroups diverged during the time period of the experiment. After running the empirical test described, I found no statistically significant results related to the effect of a mask mandate implementation on increased Yelp influence for small business survival rates across all heterogeneity tests. As a result, there is not enough evidence to suggest mask mandates had an outsized influence in increasing the significance of Yelp reviews for restaurant survival rates.

The paper proceeds as follows. Section II provides a brief background on related Yelp literature. Section III contains a discussion of the data and cleaning procedures used to conduct the empirical analysis. Section IV describes the DID empirical design and execution in the context of the Yelp data. Section V interprets the empirical results to provide a clearer picture of Yelp review impact on survival rates in a post-pandemic world. Section VI serves as a conclusion to wrap up findings and point to further research on potential coronavirus policy threshold experiments and income stratifications.

II. Related Literature

Several sources were relied upon as a basis for this paper. “Reviews, Reputation, and Revenue: The Case of Yelp.com” (Luca 2016) used a regression discontinuity design to determine the influence of a marginal star rating on revenue, and found a one star increase led to a 5 - 9% lift in sales based on data from the Washington Department of Revenue. One aspect of this paper which was taken into account was the usage of rounded Yelp ratings, or the idea that two businesses with a ratings difference of 0.01 stars should be similar in quality, but one could receive a rating that is an entire star greater. Another relevant paper comes from Limin Fang (2019) titled “The Effects of Online Review Platforms on Restaurant Revenue, Survival Rate, Consumer Learning, and Welfare”. This analysis deals specifically with survival rates in a RD context, and found doubling Yelp exposure can raise the survival rate of a restaurant by anywhere from 7 - 19 basis points. It led me to consider survival rates as the more important consequence of Yelp reviews relative to simply revenue or profit in the context of this paper.

III. Data Discussion

The "Yelp Dataset" contains more than 8.6 million reviews of 160,000 businesses in 8 metropolitan areas and is released by Yelp for academic purposes. The cleaned dataset consists of 6,275 business entries filtered to contain more than 20 reviews each in either Boston, Cambridge, or Atlanta. An additional filter was used to only select for businesses which were open at the start of 2020. In particular, 3,571 entries were based in Atlanta and 2,704 entries were based in Boston/Cambridge. Fortunately, these cities represented three of the largest eight available in the entire dataset with each containing more than 10,000 entries before filtering. Each entry comes with information related to postal code, average star rating, number of reviews, and current open status.

Arguably the most important step of the data cleaning process involved creating indicator variables to represent whether or not a given business was open for a particular month during the time period of interest. The months of interest were defined from January 2020 until July 2020, which is when the Atlanta mask mandate was introduced. Given the log history of each business describes exactly when all of the individual reviews were written, I classified a business as open or closed in a given month based on whether or not a review was given. While this method is not perfect, I am making the assumption that when most users are leaving a review on Yelp, it is being done almost immediately after visiting the restaurant as opposed to well after the fact, which seems consistent with common practice. In this example, if I saw a business had no reviews from July 2020 onward, I would assume it closed in July. The more readily available "is_open" attribute that comes with each data entry was not useful given it only alerts whether or not the business is currently closed without any indication of prior opening status. Figure 6 displays the survival rates in a given month for Atlanta and Boston before and after the onset of the pandemic for context around how survival rates changed over time. Columns were included

for each month with indicator variables assigned based on whether or not a review was given in that particular month for each of the entries. One potential confounding factor related to this classification method is that it is agnostic to the reason for a particular closure. Some proprietors could have expressed health related concerns due to the virus and decided to close their restaurant unrelated to the current demand for their product, and this could impact survival rates in a way that has nothing to do with Yelp reviews. The reason for a closure is quite relevant in tracking restaurant survival rates in that a health-related closure would not necessarily be demand-induced and could impact results. Given I did not have access to the underlying reason for business closure, this confounding influence cannot be directly fixed, but is addressed here as a relevant consideration.

Table 1 contains summary statistics of the dataset including the average star rating, review count, and monthly indicator variables from January to July indicating which percentage of businesses were open by city of interest. All metrics shown in the table are unitless with standard deviation, minimum value, and maximum value reported along with mean figures. Figure 1 displays the breakdown of star ratings by city, Figure 2 displays the breakdown of review counts by city, and Figure 3 shows the percentage of businesses open over time by city during the period of interest.

IV. Empirical Methods

The proposed empirical test for this project involves a difference-in-discontinuities (DID) design. This design in particular was chosen given the topic deals with the RD effect of an incremental star rating on business survival for two cities before and after one of them introduces a mask mandate. For these specific purposes, the running variable will be given as the micro

unrounded star rating at a given time on a monthly basis before and after the May 6th Massachusetts mask mandate implementation. The outcome variable has to do with the survival rate of businesses listed on Yelp for a given time period taken as the percentage of businesses that continued having reviews relative to the overall initial pool. In terms of the time parameters, each open indicator variable is taken on a monthly basis, and January was chosen as a starting point to represent a sufficient amount of time before any coronavirus restrictions were introduced in either state.

The policy threshold in this setup is the introduction of a mask mandate in Boston/Cambridge more than two months before the same policy in Atlanta. In terms of the specific star ratings, I will exploit the fact that half star ratings are given in the dataset while whole star ratings are reported to create a control for continuous star ratings while testing the discontinuous jump in star rating influence on survival rates. Consequently, this approach involves 4 different RD setups, each with the following equation:

$$R_{ij} = \gamma_1 S_{ij} + \gamma_2 S_{ij}^{\,2} + \beta\,(S_{ij} \geq T) + \delta_1[S_{ij} * 1(S_{ij} \geq T)] + \delta_2[S_{ij}^{\,2} * 1(S_{ij} \geq T)] + \epsilon_{ij}$$

In the context of this equation, R_{ij} refers to the survival rate for a particular restaurant i during a given time period j (either before or after the Boston mandate). S_{ij} refers to the micro, unrounded star rating for a particular restaurant i during a given time period j. T represents the star threshold of interest, or in this case 3.5 stars. 3.5 was chosen as the cutoff barrier because it contained the most data points by a large margin as shown in Figure 1, and it is relatively close to the median restaurant rating. A heterogeneity test was also conducted based on the specific star cutoff rating used, and these results are presented in Tables 4 and 5. From this equation, β

represents the coefficient of interest and signifies the impact on survival rate for a restaurant in a given time period based on a half star ratings jump. Additionally, the β term is reported in the results section from running the various RD trials in this project. Because the γ and δ terms are used to determine a line of best fit before and after mandates take place, the β term specifically deals with the impact of the mandates themselves on Yelp significance. The γ_1 and γ_2 terms are used to set the second degree polynomial line of best fit before the policy threshold, T. The δ_1 and δ_2 terms are used to set the second degree polynomial line of best fit after the policy threshold, T. Finally, the ϵ_{ij} coefficient represents the error term for a particular restaurant i during a given time period j. It is important to note this equation is only applicable for restaurants i in a specific city (either Boston/Cambridge or Atlanta) during a specific time period j (either before or after the Massachusetts mask mandate).

As a result, this same RD setup would be used in four different scenarios where the city contains restaurants, i, and the time period represents the interval of interest, j. A pass at the diff-in-disc estimate would then be represented as:

$$(\beta_{b,x} - \beta_{b,y}) - (\beta_{a,x} - \beta_{a,y})$$

Where β represents the coefficient of interest for each of the four RD equations, the first subscript refers to the city of interest (a for Atlanta and b for Boston), and the second subscript refers to the timing (x for before the Boston mandate and y for after).

There are three main assumptions necessary to identify the causal effect of an incremental half star rating on survival rate as a result of mask mandate restrictions:

I. The eligibility index of which restaurants are grouped into a whole number star rating should be continuous around the cutoff. In this case, there should not be any evidence that restaurants are purposely manipulating ratings in order to observe a star rating increase. The primary method by which to test this assumption is the McCrary Density Test which thoroughly examines the density function around the cutoff to ensure consistency.

II. Data points close to the cutoff should be similar to one another. In the context of this experimental design, this means restaurants receiving 3 stars versus 3.5 stars should be closely related to one another in both observed and unobserved characteristics. Proving this assumption tends to be more difficult. One could use observed characteristics as covariates in the equation above to check for and quantify similarity, but there is no valid way to measure unobserved features, so this portion must be assumed.

III. Assuming the RD estimate is used as the outcome of a diff-in-diff design, the final assumption is parallel trends. This assumption requires that in the absence of treatment, or in this case the mask mandate, the difference between the treatment and control groups is constant over time. While there is not a statistical test to validate this assumption, one method is to graph out both the treatment and control over time before the mandate and visually examine both trends.

Finally, I conducted several heterogeneity analyses using the same RD equation given above for each of the four situations separated by median income on a zip code basis and by star rating. In particular, this consists of examining beta coefficients for each group, and each of these coefficients represent the impact on survival rating for a half star ratings jump within that particular subgroup. The two subgroups of interest for the zip code variable involve restaurants in a location below the given metro area median average household income and restaurants

above this threshold. The test was also conducted using a "donut" approach, or removing the middle 50% of median household incomes of restaurant locations from the data. These subgroups serve as a proxy for low and high traffic areas respectively assuming traffic is correlated with the socioeconomic status of a restaurant location. The subgroups of interest for the star variable involve running the same test using 3.0 and 4.0 stars as the cutoff threshold instead of 3.5. These tests allow for valuable comparisons between the influence on survival rates over time given a half star ratings jump for different socioeconomic areas and a variety of star ratings.

V. Estimation & Results

Table 2 displays the results of the empirical test by city and time period. The reported results stem from the β coefficient of the estimating equation presented on page 7 given this term describes the impact of specifically mask mandates on the incremental effect of a higher Yelp rating on restaurant survival rates. These results imply that given an additional one-half star ratings increase, one would expect a 1.38% increase to survival rate for restaurants in Boston pre-mandate. For pre-mandate Atlanta, one would expect a 3.21% increase in survival rate. These figures make intuitive sense given a higher star rating should lead to a greater chance of discovery and survival, all else equal. However, neither of these estimates are statistically significant at the 0.05 level taking into account standard errors given they have t-values of 0.46 and 1.54 respectively. In terms of post-mandate estimates, one would expect a 2.07% increase in survival rate for restaurants in Boston and a 2.47% decrease in survival rate for restaurants in Atlanta. Once again, neither of these estimates are statistically significant taking into account standard errors given they have t-values of 0.72 and -1.02 respectively. As a result, it cannot be

concluded for either city in either time period that a one-half star Yelp ratings increase plays a role in influencing survival rates of businesses in that subgroup. However, all of the estimates seem to make directional sense in terms of Yelp star ratings contributing to a lift in survival rate, except for the Atlanta post-mandate estimate. This result is surprising because it indicates that an increase in the Yelp star rating actually decreased the survival rate, but once again not in a statistically significant way.

The next step of the analysis involves finding the diff-in-disc estimator using the equation described above in Section 3. To find the standard error for the diff-in-disc estimator, I was able to use the fact that $Var(\beta_1 + \beta_2) = Var(\beta_1) + Var(\beta_2) + 2Cov(\beta_1, \beta_2)$. Additionally, I can assume $Cov(\beta_1, \beta_2) = 0$, or in other words there is no covariance between either of the β estimates in a given city before and after the Massachusetts mandate took place. These results imply that the impact on survival rate for an additional half star ratings increase was 6.37% higher as a result of the implementation of a mask mandate. However, this result is not statistically significant taking into account standard errors given a t value of 1.05. As a result, it cannot be concluded that mask mandates had a statistically significant role in influencing the impact of Yelp ratings on survival rates at the 0.05 level. That being said, this result makes sense given that in Boston it seemed as though the significance of Yelp rating increases grew after the introduction of mask mandates and waned in Atlanta without the mandate. It also makes directional sense given I would expect mandates drive more digitization and give greater importance to online aggregated rating platforms like Yelp, so seeing a positive value for this coefficient does not come as a surprise.

Finally, a heterogeneity analysis was conducted to determine how these results could potentially be influenced by choosing a different star cutoff or segmenting businesses on

estimated foot traffic. To start with, I ran the same analysis above using 3.0 and 4.0 stars instead of the original 3.5 stars as the cutoff point. Other star cutoffs (i.e. 2.0 stars or 5.0 stars) were not chosen because there were not a sufficient number of points on both sides of the cutoff point for each city and time period subgroup to achieve meaningful results. The main purpose of this analysis is to determine if the influence of an increased Yelp rating on survival rate matters more for low-rated or high-rated businesses given the original 3.5 chosen represents an average rated restaurant. Table 4 shows this analysis using a cutoff point of 4.0 stars, and the only difference from this result is that the Boston pre-mandate coefficient of interest became statistically significant. In other words, a half star increase of going from 4.0 to 4.5 stars in Boston before the mandate led to an increased survival rate by on average 8.03%, which is statistically significant at the 0.05 level. This finding is consistent with the Luca paper and implies the marginal rating increase for a lower rated business did not have as much of an influence as it did for a relatively higher rated business in Boston before the mask mandate. The same analysis was conducted using a cutoff point of 3.0 stars, and there was no difference in these results compared to the original 3.5 cutoff point in terms of any directional changes or the significance of coefficients of interest.

The other heterogeneity analysis conducted in this project was segregating businesses by location above and below average median metro area household income in addition to a “donut” approach of removing the middle 50% of incomes for restaurant locations in each city. Businesses were separated into two different buckets of high and low income based on the overall metro average in the normal approach, and if they were in the bottom or top 25th percentile of their respective city in the donut approach. This metric could be thought of as a proxy to better understand whether higher or lower income areas benefit more in terms of

survival rate impact based on a half star ratings increase. Table 6 shows the results from the lower income pool separated by median. This analysis did not yield any differences in results compared to the original test in terms of any directional changes or significance of coefficients of interest. The diff-in-disc coefficient of interest implies the introduction of a mask mandate led to Yelp reviews having a 15.01% higher impact on survival rates than without the mandate, but this is not a statistically significant result given the high associated standard error. In terms of the high income grouping separated by median, a similar result was found in that there were no differences in results compared to the original test in terms of directional changes or significance of coefficients of interest. The diff-in-disc coefficient of interest implies the introduction of a mask mandate led to Yelp reviews having a 2.87% higher impact on survival rates than without the mandate, but this is not a statistically significant result given the high associated standard error. Similar results are found for the donut approach and these results are displayed in Tables 8 and 9. It is also important to note that a confounding factor of income and spending divergence took place during the pandemic between higher and lower income areas unrelated to a Yelp ratings increase. In other words, income fell much faster more broadly in lower income households while spending fell to a greater degree in higher income households, and these changes could have an impact on survival rates unrelated to restaurant Yelp ratings. That being said, it is an interesting result that lower-income restaurants seemed to have a higher jump in survival rates from an incremental Yelp review with a mandate in place than higher-income restaurants, and this could be an interesting direction to look further into.

Finally, Figures 4-5 show examples of RD graphs for the original tests. These graphs display survival rate by average restaurant stars given in addition to a confidence band at each point. The cutoff of interest is going from 3.5 to 4 stars, so a line of best fit is found before and

after this threshold value. Grey dots represent bin averages for each star rating and the bands around those dots are confidence intervals. This same graph type is shown for each scenario.

VI. Conclusion

Using data compiled by Yelp, I was able to run a difference-in-discontinuities test to exploit the fact that Boston/Cambridge introduced a mask mandate several months ahead of Atlanta. This design allowed me to answer the question of what influence mask mandates have on the impact of Yelp rating increases for small business survival rates. Additional heterogeneity tests were undertaken to examine this effect across a variety of star rating jumps and income classifications. Across all of these tests, the only statistically significant result achieved was the influence of a Yelp ratings increase on restaurant survival before the mask mandate was introduced in Boston for a jump from 4.0 to 4.5 stars. While not many estimators proved to be statistically significant, it was interesting to observe that the coefficients of interest directionally showed that the introduction of a mask mandate increased the influence of an additional half star Yelp rating on survival rates, implying Yelp could have potentially become more important for small and medium businesses during the pandemic.

These directional results have the potential to be important findings if proven to be true given they imply Yelp adoption not only increased, but began influencing real life decisions more than usual during the pandemic. Another interesting finding from the project was that the largest diff-in-disc coefficient of interest came from the low-income restaurant population, implying that Yelp influence mattered the most in terms of long run survival after mask mandates to low-income businesses which would otherwise have a difficult time with traffic. From a policy perspective, this is relevant in explaining the effect of mask mandates and other coronavirus

restrictions on technology adoption and small business discovery. Specifically, Yelp could be an important outlet moving forward for lower-income and low-traffic areas to have sufficient advertising exposure without having to pay nearly as much as their larger, well-established competitors.

In terms of future research, the most exciting development from this paper stems from the difference in magnitude surrounding the lower income restaurant subgroup compared to the higher income population in terms of the influence of mandates on the importance of a half star ratings bump. I think conducting a higher powered heterogeneity analysis could yield important results about Yelp as a means to place restaurants on an equal discovery footing, regardless of socioeconomic status. Additionally, it could be interesting to run a similar type of experiment with a different threshold policy than mask mandates. While mandates are generally a good proxy for covid restrictions, it would be interesting to see these same results using lockdowns or coronavirus case number thresholds as policies of interest instead. Having more granular data on the rationale for closing, whether health related or demand induced, could also lead to more interesting results given it could ameliorate the confounding influence of virus related closures on restaurant survival rates unrelated to Yelp. All in all, I think there is a lot of exciting research to be done on the influence of consumer preferences as a result of digitization stemming from coronavirus restrictions, and Yelp is just the tip of the iceberg.

References

Fang, Limin. *The Effects of Online Review Platforms on Restaurant Revenue, Survival Rate, Consumer Learning and Welfare*. U.C. Davis, 26 Jan. 2019,https://www.econ.ucdavis.edu/events/papers/65Fang.pdf.

Grembi, Veronica, et al. *Do Fiscal Rules Matter? A Difference-in-Discontinuities Design*. Catholic University of Milan, July 2011,https://www.bancaditalia.it/pubblicazioni/altri-atti-convegni/2011-pigou-hobbes/4-Grembi-Nannicini-Troiano.pdf.

Li, Hanlin, and Brent Hecht. *3 Stars on Yelp, 4 Stars on Google Maps: A Cross-Platform Examination of Restaurant Ratings*. Northwestern University, Nov. 2020,https://brenthecht.com/publications/cscw2020_restaurantratings.pdf.

Luca, Michael. *Reviews, Reputation, and Revenue: The Case of Yelp*. Harvard Business School, Dec. 2016, https://www.hbs.edu/ris/Publication%20Files/12-016_a7e4a5a2-03f9-490d-b093-8f951238dba2.pdf.

Zhu, Feng, and Xiaoquan Zhang. *Impact of Online Consumer Reviews on Sales: The Moderating Role of Product and Consumer Characteristics*. Journal of Marketing, Mar. 2010,https://www.jstor.org/stable/pdf/20619095.pdf?refreqid=excelsior%3A05d5bd6a9113aad618c57e7978af309.

Appendix

Table 1: Summary Statistics

Yelp Data Summary Statistics

	Total				Atlanta				Boston/Cambridge			
	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max
Star Rating	3.65	0.72	1	5	3.60	0.76	1	5	3.71	0.65	2	5
Review Count	200.01	312.58	21	7,298	172.17	272.90	21	3,948	236.78	355.02	21	7,298
1(Open Jan 2020)	0.951	0.215	0	1	0.959	0.198	0	1	0.941	0.236	0	1
1(Open Feb 2020)	0.868	0.339	0	1	0.895	0.307	0	1	0.833	0.374	0	1
1(Open Mar 2020)	0.788	0.409	0	1	0.833	0.373	0	1	0.729	0.444	0	1
1(Open Apr 2020)	0.781	0.414	0	1	0.824	0.381	0	1	0.725	0.447	0	1
1(Open May 2020)	0.769	0.421	0	1	0.809	0.393	0	1	0.717	0.451	0	1
1(Open Jun 2020)	0.749	0.435	0	1	0.787	0.410	0	1	0.698	0.459	0	1
1(Open Jul 2020)	0.745	0.436	0	1	0.782	0.413	0	1	0.696	0.460	0	1

Note: 6,275 observations: All, 6,275; Atlanta, 3,571; and Boston/Cambridge, 2,704.
Note: Variables further explained in data section with time period extending from 1/1/2020 - 7/8/2020

Table 2: Regression Discontinuity Results by City & Time Period with 3.5 Star Cutoff

Regression Discontinuity Results by City & Time Period

	Boston, Pre-Mandate	Boston, Post-Mandate	Atlanta, Pre-Mandate	Atlanta, Post-Mandate
	(1)	(2)	(3)	(4)
Stars	0.01380	0.02073	0.03210	-0.02468
Standard Error	0.03018	0.01349	0.04443	0.02425
Gamma_1	0.05350	-0.15012	0.21248	-0.01524
Gamma_2	-0.00105	0.03213	-0.03628	0.00575
Delta_1	1.53767	0.47774	1.25092	0.50642
Delta_2	-0.18675	-0.07552	-0.14286	-0.06823
Bandwidth Estimate	0.833	0.761	1.453	1.291
Observations	2,704	1,938	3,571	2,890
Observations Left of Cutoff	588	384	1,003	780
Observations Right of Cutoff	2,116	1,554	2,568	2,110

Note: Mandate took place on May 6, 2020
Note: Degree 2 polynomial fit
Note: Cutoff point of 3.5 stars

Table 3: Difference-in-Discontinuities Estimator and Coefficients of Interest

Difference in Discontinuity Results

	Boston, Pre-Mandate	Boston, Post-Mandate	Atlanta, Pre-Mandate	Atlanta, Post-Mandate	Diff-in-Disc
	(1)	(2)	(3)	(4)	(5)
Stars	0.01380	0.02073	0.03210	-0.02468	-0.06371
Standard Error	0.03018	0.01349	0.04443	0.02425	0.06046
Bandwidth Estimate	0.833	0.761	1.453	1.291	N/A
Observations	2,704	1,938	3,571	2,890	6,275

Note: Mandate took place on May 6, 2020
Note: Degree 2 polynomial fit
Note: Cutoff point of 3.5 stars

Table 4: Regression Discontinuity Results by City & Time Period with 4.0 Star Cutoff

Regression Discontinuity Results by City & Time Period

	Boston, Pre-Mandate	Boston, Post-Mandate	Atlanta, Pre-Mandate	Atlanta, Post-Mandate	Diff-in-Disc
	(1)	(2)	(3)	(4)	(5)
Stars	0.08031	-0.00579	0.03408	0.00506	0.05708
Gamma_1	-0.17038	0.06380	-0.01158	0.01477	N/A
Gamma_2	0.03981	-0.00692	0.00725	-0.00001	N/A
Delta_1	2.96289	1.05026	3.06467	1.03330	N/A
Delta_2	-0.36342	-0.12573	-0.36530	-0.12536	N/A
Standard Error	0.02241	0.00811	0.03389	0.01627	0.04451
Bandwidth Estimate	0.970	0.866	1.283	1.313	N/A
Observations	2,704	1,938	3,571	2,890	6,275
Observations Left of Cutoff	1,297	870	1,867	1,472	N/A
Observations Right of Cutoff	1,407	1,068	1,704	1,418	N/A

Note: Mandate took place on May 6, 2020
Note: Degree 2 polynomial fit
Note: Cutoff point of 4.0 stars

Table 5: Regression Discontinuity Results by City & Time Period with 3.0 Star Cutoff

Regression Discontinuity Results by City & Time Period

	Boston, Pre-Mandate	Boston, Post-Mandate	Atlanta, Pre-Mandate	Atlanta, Post-Mandate	Diff-in-Disc
	(1)	(2)	(3)	(4)	(5)
Stars	0.03033	0.03212	-0.01614	0.02761	0.04196
Gamma_1	-0.32556	-0.42160	0.59147	-0.14206	N/A
Gamma_2	0.07750	0.08861	-0.11991	0.03339	N/A
Delta_1	0.90938	0.72481	0.06380	0.38847	N/A
Delta_2	-0.14495	-0.12908	0.03682	-0.06667	N/A
Standard Error	0.04545	0.02817	0.02932	0.01799	0.06358
Bandwidth Estimate	0.770	0.902	0.739	0.871	N/A
Observations	2,704	1,938	3,571	2,890	6,275
Observations Left of Cutoff	227	142	501	389	N/A
Observations Right of Cutoff	2,477	1,796	3,070	2,501	N/A

Note: Mandate took place on May 6, 2020
Note: Degree 2 polynomial fit
Note: Cutoff point of 3.0 stars

Table 6: Low Income RD Results via Metro Median Cutoff by City & Time Period

Regression Discontinuity Results by City & Time Period for Low-Income Areas

	Boston, Pre-Mandate	Boston, Post-Mandate	Atlanta, Pre-Mandate	Atlanta, Post-Mandate	Diff-in-Disc
	(1)	(2)	(3)	(4)	(5)
Stars	0.02690	0.02209	0.11709	-0.03873	-0.15101
Gamma_1	0.03127	-0.00138	-0.07655	0.11667	N/A
Gamma_2	0.00297	0.00174	0.00978	-0.01887	N/A
Delta_1	1.48223	0.82768	1.46452	0.11460	N/A
Delta_2	-0.18083	-0.10495	-0.17565	-0.01176	N/A
Standard Error	0.05539	0.03143	0.08481	0.03056	0.11037
Bandwidth Estimate	0.815	0.793	1.377	1.227	N/A
Observations	735	551	1,235	986	1,970
Observations Left of Cutoff	182	122	325	248	N/A
Observations Right of Cutoff	553	429	910	738	N/A

Note: Mandate took place on May 6, 2020
Note: Degree 2 polynomial fit
Note: Cutoff point of 3.5 stars

Table 7: High Income RD Results via Metro Median Cutoff by City & Time Period

Regression Discontinuity Results by City & Time Period for High-Income Areas

	Boston, Pre-Mandate	Boston, Post-Mandate	Atlanta, Pre-Mandate	Atlanta, Post-Mandate	Diff-in-Disc
	(1)	(2)	(3)	(4)	(5)
Stars	0.00990	0.02059	-0.00358	-0.02154	-0.02865
Gamma_1	0.22818	-0.07480	0.36008	-0.06484	N/A
Gamma_2	-0.03038	0.01968	-0.06010	0.01512	N/A
Delta_1	1.37947	0.31739	1.18058	0.71639	N/A
Delta_2	-0.15958	-0.05328	-0.13052	-0.09767	N/A
Standard Error	0.03609	0.01395	0.05156	0.03201	0.07197
Bandwidth Estimate	0.839	0.757	1.502	1.352	N/A
Observations	1,968	1,386	2,336	1,904	4,304
Observations Left of Cutoff	405	261	678	532	N/A
Observations Right of Cutoff	1,563	1,125	1,658	1,372	N/A

Note: Mandate took place on May 6, 2020
Note: Degree 2 polynomial fit
Note: Cutoff point of 3.5 stars

Table 8: Low Income RD Results via Donut Construction by City & Time Period

RD Results by City & Time Period for Donut Construction Low-Income Areas

	Boston, Pre	Boston, Post	Atlanta, Pre	Atlanta, Post	Diff-in-Disc
	(1)	(2)	(3)	(4)	(5)
Stars	0.03609	0.04694	0.13373	-0.04976	-0.19434
Gamma_1	0.76466	-0.25998	0.40431	-0.13952	N/A
Gamma_2	-0.13520	0.04956	-0.06923	0.03063	N/A
Delta_1	1.61531	0.40156	1.33111	0.98592	N/A
Delta_1	-0.14592	-0.07296	-0.14437	-0.13970	N/A
Standard Error	0.06493	0.04205	0.13150	0.03457	0.15643
Bandwidth Estimate	0.861	0.823	1.455	1.322	N/A
Observations	561	416	497	383	1,058
Observations Left of Cutoff	122	86	159	114	N/A
Observations Right of Cutoff	439	330	338	269	N/A

Note: Mandate took place on May 6, 2020
Note: Degree 2 polynomial fit
Note: Cutoff point of 3.5 stars

Table 9: High Income RD Results via Donut Construction by City & Time Period

RD Results by City & Time Period for Donut Construction High-Income Areas

	Boston, Pre	Boston, Post	Atlanta, Pre	Atlanta, Post	Diff-in-Disc
	(1)	(2)	(3)	(4)	(5)
Stars	-0.07154	0.02476	0.01499	0.00932	-0.10197
Gamma_1	-0.87089	-0.30450	0.08130	0.20979	N/A
Gamma_2	0.14567	0.06096	-0.01813	-0.03441	N/A
Delta_1	3.48046	-0.09566	0.44543	-0.27906	N/A
Delta_2	-0.46405	-0.00928	-0.03487	0.04197	N/A
Standard Error	0.04987	0.01900	0.07430	0.04779	0.10321
Bandwidth Estimate	0.813	0.790	1.496	1.472	N/A
Observations	917	619	1,294	1,056	2,211
Observations Left of Cutoff	210	140	316	250	N/A
Observations Right of Cutoff	707	479	978	806	N/A

Note: Mandate took place on May 6, 2020
Note: Degree 2 polynomial fit
Note: Cutoff point of 3.5 stars

Figure 1: Frequency of Yelp Star Ratings by City

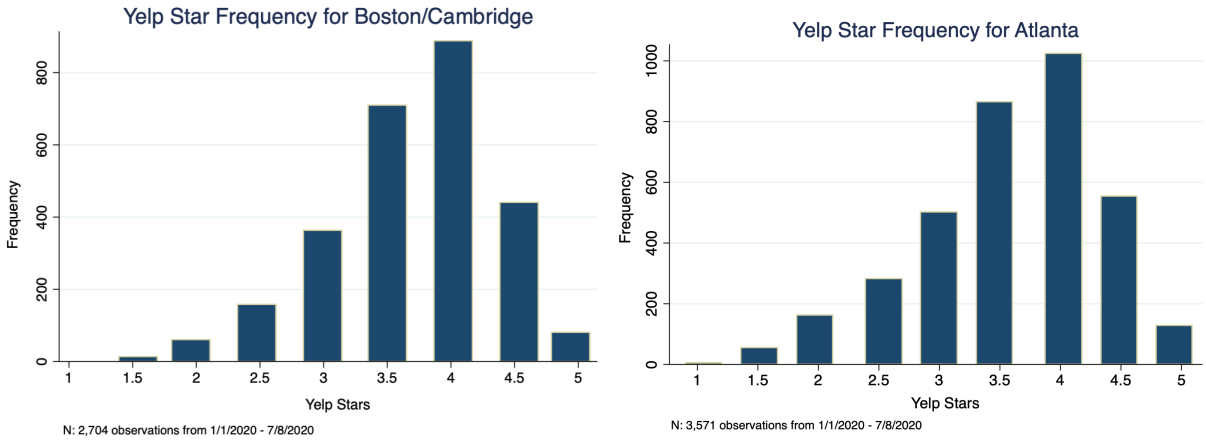


Figure 2: Frequency of Review Counts by City

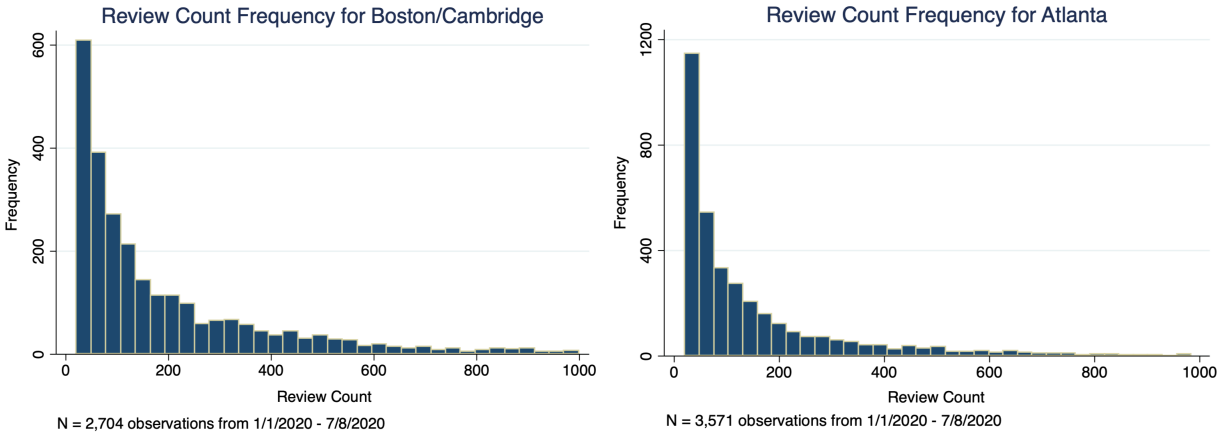


Figure 3: Restaurant Survival Rate January - July 2020 by City

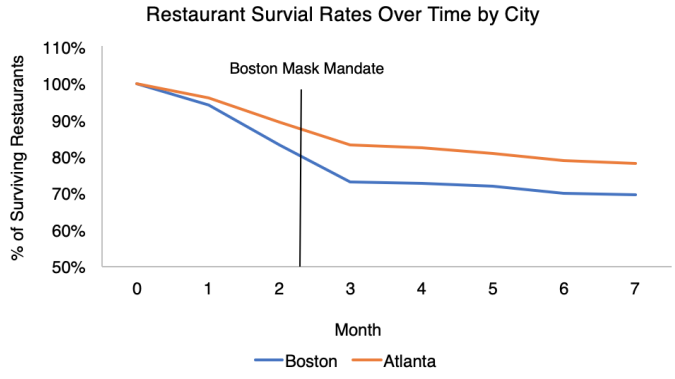


Figure 4: RD Graph of Survival Rate by Star for Atlanta Pre-Mandate with 3.5 Star Cutoff

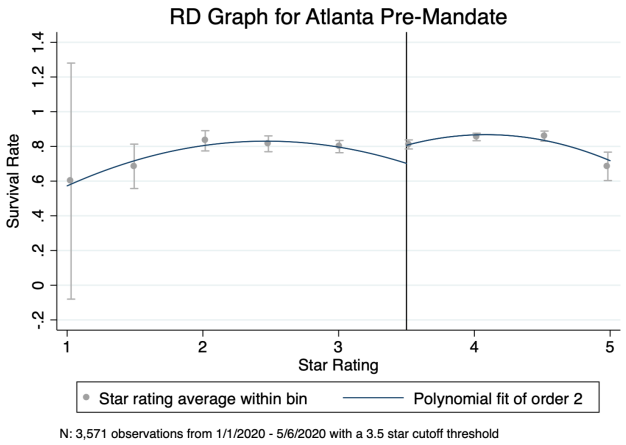


Figure 5: RD Graph of Survival Rate by Star for Atlanta Post-Mandate with 3.5 Cutoff

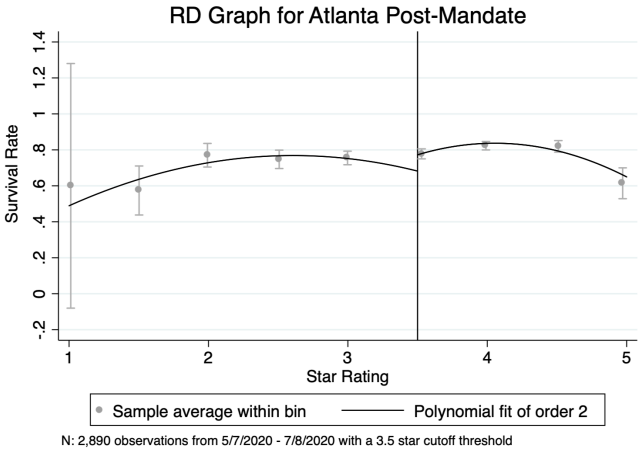
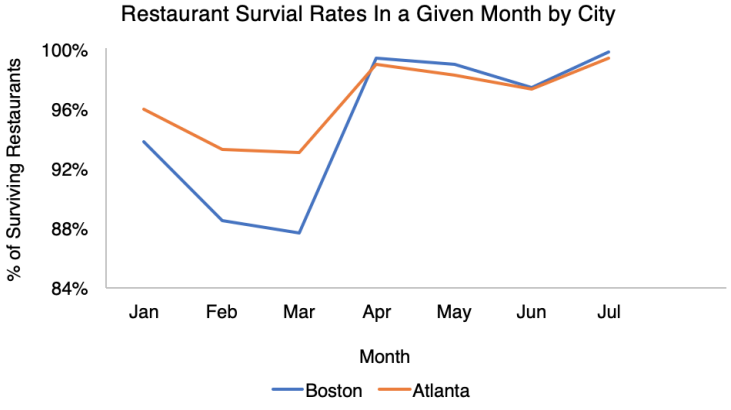


Figure 6: Restaurant Survival Rate in a Given Month January - July 2020



The Effect of National School Lunch Program Eligibility on Test Scores

Miriam Zuo

December 9, 2021

Abstract

Most research on the National School Lunch Program tends to focus on its nutritional content and its impact on student health outcomes. In this paper, I examine the effect of NSLP eligibility, both for free lunch and for reduced-price lunch, on standardized test scores in reading, math, and science. I utilize a regression discontinuity analysis to investigate whether there exists a significant improvement in test scores among students who are barely eligible for NSLP relative to those who are barely ineligible. An 11.5% (6.294 points) increase in science test scores is associated with free lunch eligibility in the same academic year, with a 95% confidence interval from 5.5% (3.001 points) to 17.5% (9.587 points). This result suggests free lunch eligibility has a large, statistically significant effect on science test scores. I also find that free lunch eligibility is associated with improvements in math and science scores in the following year. As a whole, however, proponents of extending NSLP eligibility to more students may need to consider the program's nutritional benefits over its academic ones.

Introduction

The National School Lunch Program (NSLP) has offered free and reduced-price lunches to students since 1946 when it was established by President Harry Truman. Since then, it has become a major source of

nutrition for low-income students all across the US in both public and non-profit private schools. Because it has played a tremendous role in the educational experience and food security of millions of students over the past seven decades, it is important to carefully examine its effects on students' health and academic outcomes.

In this paper, I study the effect of NSLP eligibility on test scores, specifically in reading, math, and science. It seems plausible that NSLP could incentivize students to attend school more often for the sake of a consistent source of nourishment; in turn, having sufficient energy to focus or simply spending more time in class could improve students' test scores. Thus, the principal objective of this paper is to expand our understanding of the effects of NSLP on academic outcomes. However, because the Early Childhood Longitudinal Study (ECLS) data I use does not indicate NSLP enrollment, I focus on NSLP eligibility instead. Unlike much of the existing literature, the findings of this paper could help clarify future discussion on using academic performance as a motivator for expanding NSLP access, e.g. by changing the eligibility requirements: currently, students are eligible for free lunch and reduced-price lunch if their household income is less than 130% or 185% of the federal poverty level, respectively, such that increasing these percentages would enable more students to become eligible for FRP lunch. Alternatively, proponents of NSLP could aim to increase enrollment by encouraging eligible yet unenrolled students to apply through nudges or more aggressive means.

To answer this research question, I utilize a regression discontinuity analysis at the cutoffs for free lunch eligibility and for reduced-price lunch eligibility to determine if students who are barely eligible for free/reduced-price (FRP) lunch demonstrate significantly higher test scores than students who are barely ineligible. I begin with analyzing the effect of NSLP eligibility in year t on test scores (reading, science, and math) in year t. Then, I look at the effect of NSLP eligibility in year $t-1$ or year $t-2$ on test scores in year t to assess whether there may be lasting effects of NSLP eligibility.

I find that NSLP eligibility may have limited effects on student test scores. Though some data limitations may impact the efficacy of regression discontinuity as a tool, my analysis suggests that free lunch eligibility in year t is associated with decreases in reading and math scores, though it is associated with a significant increase in science test scores in the same year. Across the reduced-price eligibility cutoff in year t, I find

statistically significant decreases in year t test scores in all three subjects. Similarly, for FRP lunch eligibility in years $t-1$ and $t-2$, for the most part at a 95% confidence level I cannot conclude that FRP eligibility is associated with significant test score improvements in year t, but it appears that free lunch eligibility in the previous year may improve math and science scores in the current year.

Literature Review

Gordon and Ruffini (2018) looks at the Community Eligibility Provision (CEP) of the Healthy, Hunger-Free Act of 2012. CEP is a program that makes school breakfasts and lunches free for all students, regardless of income, and it is available in schools within districts that have Identified Student Percentages (ISPs) of 40% or higher [1]. The authors utilize the staged rollout of pilot programs in 2012 and find small reductions in suspension rates for elementary and middle school students, though not for high school students, with the largest improvements among the most disadvantaged subpopoulations.

Work by Dunifon and Kowaleski-Jones (2003) does not find significant effects of NSLP on positive behaviors or math or reading scores, i.e. that there are no significant effects on child outcomes after addressing selection through a siblings fixed-effects model. On the other hand, Hinrichs (2010) found a sizable effect of NSLP on long-term educational attainment (though not long-term health outcomes) in the mid-20th century – as Hinrich's data is from before the establishment of School Breakfast Program (SBP), which is potentially a confounding factor in more recent studies examining NSLP (both programs have the same income cutoffs), it isolates for the effects of NSLP in the earliest years following its inception.

Some work has also been done on SBP, with mixed results. Imberman and Kugler (2014) find that math and reading achievement increases are associated with providing school breakfasts in class rather than the cafeteria, with the strongest effects among low-performing, free lunch-eligible, Hispanic, and low body mass index students. Similarly, Frisvold (2012) finds that SBP increases student achievement. Also, Dotter (2013) finds universally free breakfasts increase math and reading test scores, with particularly significant gains among low-performing students attending low-income schools. However, Schanzenbach and Zaki (2014) report sparse evidence on positive effects in nutrition, health, behavior, or achievement. Further, existing

[1] Though CEP pilot programs began in 2012, the nation-wide rollout occurred in 2015, so it should not significantly affect the data collected during grades K-3 for the 2010 cohort (from fall 2010-spring 2014).

literature seems to suggest that effects of NSLP tend to outweigh those of SBP, possibly because more students each lunch at school than breakfast.

Data

The dataset consists of two cohorts of the Early Childhood Longitudinal Study (ECLS). This study is sponsored by the Department of Education, and the data is collected by the National Center for Education Statistics. For the 2010 cohort, the study follows students from kindergarten through fifth grade, but this paper uses data from grades K-3 only as some of the covariates and response variables I use are not available in the public version of the data set for grades 4 and 5. For the 1998 cohort, the study follows students from kindergarten through eighth grade, but data is collected only in kindergarten fall/spring, first grade fall/spring, third grade spring, fifth grade spring, and eighth grade fall/spring (i.e. second, fourth, sixth, and seventh grade are excluded). For both cohorts, data is collected through school, teacher, and parent questionnaires, as well as direct student assessments. Note that there are just two cohorts to date; the next cohort begins with the kindergarten class of 2024.

The main variable I am interested in is whether a student is eligible for the National School Lunch Program (NSLP) – either for free lunch or for reduced-price lunch. As this variable is not available in the data set, for both cohorts I constructed it from data on household income and household size. Since household income and household size are both self-reported in the ECLS-administered questionnaires, not all respondents provided this data. Therefore, in order to generate variables for eligibility, which depends on both household income and household size, I dropped the 119,583 observations (of 236,994) missing either household income or household size. Across the US, free lunch via NSLP is available for students at public and nonprofit private schools whose household incomes are less than 130% of the federal poverty level (which varies based on household size), and reduced-price lunch is available for those with household incomes below 185% of the federal poverty level. Data on the federal poverty level for each year is gathered from the website for the Assistant Secretary for Planning and Evaluation (ASPE). [2]

The dependent variables in the data set are student test scores in math, reading, and science. These test

[2]In the 1998 cohort, income information is not available for when students were in kindergarten, so kindergarten NSLP eligibility cannot be calculated for that cohort.

scores are all available in the spring of grades K-3 for the 2010 cohort. For the 1998 cohort, reading and math test scores are available in grades K, 1, 3, 5, and 8, but science test scores are available just for grades 3, 5, and 8. The public-use manual states that the test scores represent status with respect to achievement on a particular criterion set of assessment items. They are scaled to represent the probabilities of correct answers, summed over all items in the question pools to enable longitudinal comparisons [3]; all reading assessments are out of 212, and math and science are out of 174 and 111, respectively.

Beyond household size and income, other covariates I have included are the child's sex and race, as well as their parent's highest level of education and whether they attend a public or private school. For the 2010 cohort, race was not available in the public version of the data, so I constructed it from parent race variables, which were available.

The cohort data comes in panel form when downloaded from the ECLS website (2010 cohort)/electronic codebook (1998 cohort), but I reshaped it to a long format such that each student can constitute a separate observation in each year (i.e. student X in grades K, 1, 2, and 3 would be 4 observations). The full sample size for both cohorts is 98,172, with 28,794 eligible for free lunch and 40,395 eligible for reduced-price lunch (including those eligible for free lunch), where each observation is one student in a given school year. However, not all variables are available for each observation; observations with missing values (excluding those missing household income and/or household size) have not been dropped from the data set. See Table 1 for summary statistics on the entire sample. Tables 2 and 3 display summary statistics for the free lunch-eligible subsample and RP lunch-eligible subsample, respectively. The outcome variables I will look at are students' reading, math, and science test scores. Figure 1 depicts the distribution of these scores, which are slightly skewed to the right.

A key strength of this dataset is that the cohort data is available in panel form, thereby allowing us to examine the effect of FRP eligibility at a student level rather than at an aggregate level, which may reduce some noise around the RD cutoff. Further, the dataset is large (tens of thousands of students) and longitudinal, enabling analysis of the effect of NSLP eligibility in one year on test scores several years in the future.

[3] The exact procedure for calculating the scaled scores is not given in the public-use manual, but the justification for using scaled scores over raw number-right scoring is that "IRT can compensate for the possibility of a low-ability child guessing several difficult items correctly."

However, there are a number of caveats regarding the data collection procedure that may introduce noise into the RD analyses. First, for both cohorts, it's not entirely clear what the reference group is for household income; parents report annual household income in the spring, but I am not sure whether this refers to income over the past 12 months or expected income for the next year (or something else). Since NSLP eligibility is computed in the fall, I am assuming that most parents had the same eligibility status in the fall as they did in the spring questionnaire.

Second, in the 2010 cohort, household size is collected in the fall, when families can apply to NSLP. In the 1998 cohort, however, household size is available in the spring only, so it is possible that some families increased in size during the fall semester such that they may have become eligible after the fall applications were sent out. While it is technically possible to enroll in NSLP at any point in the school year, most families enroll at the start of the school year.

It's also possible that some students switched schools between semesters and were enrolled in NSLP in the fall but not in the spring. Data on whether students switched schools is only available for the kindergarten year, so the RD does not control for students who stayed at the same school for the entire academic year. Note that students who transfer within the same school district automatically retain NSLP status, but outside of the district, students may need to re-apply.

Furthermore, the federal poverty levels are higher in Alaska and Hawaii than in the continental US, but because the public data does not specify the state each student lives in, I cannot adjust eligibility accordingly for students in those states.

Empirical Methods

The Effect of NSLP Eligibility on Current-Year Test Scores

I conduct a regression discontinuity analysis to assess the effect of NSLP on academic outcomes, specifically on standardized math, science, and reading scaled test scores. I analyze the NSLP effect at two cutoffs: the cutoff for free lunch, and the cutoff for reduced-price lunch. Since it is plausible a student could be eligible for the program one year and ineligible the next, I start by assessing the effect of eligibility in year t

on academic outcomes in year t.

The running variable I use is p_{it}, the ratio of household income to the federal poverty level. Let F_{it} represent free lunch eligibility for student i in year t such that when the student's household income p_{it} exceeds 130% of the federal poverty level that year FPL_{it}. Note that FPL_{it} is based on both the year and the student's household size.

$$F_{it} = \begin{cases} 1 & \text{if } p_{it} < 1.3, \\ 0 & \text{if } p_{it} \geq 1.3 \end{cases} \tag{1}$$

The sharp RD specification of the effect of free lunch eligibility on academic outcomes is as follows, where R_{it}, M_{it} and S_{it} represent reading, math, and science scores for student i in year t, respectively. Coefficients are allowed to differ across the eligibility cutoff via the interaction terms, and a quadratic relationship is permitted in order to grant more flexibility to the relationship between the running variable p_{it} and the dependent variable.

$$R_{it} = \alpha_r + \rho_r F_{it} + \beta_r(p_{it} - 1.3) + \gamma_r(p_{it} - 1.3)^2 + \delta_r F_{it} \cdot (p_{it} - 1.3) + \nu_r F_{it} \cdot (p_{it} - 1.3)^2 + \epsilon_{it} \tag{2}$$

$$M_{it} = \alpha_m + \rho_m F_{it} + \beta_m(p_{it} - 1.3) + \gamma_m(p_{it} - 1.3)^2 + \delta_m F_{it} \cdot (p_{it} - 1.3) + \nu_m F_{it} \cdot (p_{it} - 1.3)^2 + \epsilon_{it} \tag{3}$$

$$S_{it} = \alpha_s + \rho_s F_{it} + \beta_s(p_{it} - 1.3) + \gamma_s(p_{it} - 1.3)^2 + \delta_s F_{it} \cdot (p_{it} - 1.3) + \nu_s F_{it} \cdot (p_{it} - 1.3)^2 + \epsilon_{it} \tag{4}$$

The RD analyses utilize the MSE-optimal bandwidth of p_{it} as most of the existing literature on RD use this metric in bandwidth selection. In regards to observation weights, I use an uniform kernel; since the area around the cutoff is noisy, it seems more reasonable to weigh observations within the selected bandwidth equally.

Similarly, let RP_{it} represent reduced-price lunch eligibility for student i in year t such that

$$RP_{it} = \begin{cases} 1 & \text{if } p_{it} < 1.85 \\ 0 & \text{if } p_{it} \geq 1.85 \end{cases} \tag{5}$$

The RD specification is identical to equations 2-4 above, but F_{it} is replaced by RP_{it}, and the cutoff is

re-centered around the reduced-price lunch cutoff such that p_{it} is subtracted by 1.85 instead of 1.3.

Assumptions for Validity

For the RD analysis, the only assumption for validity required is that all potential outcomes are continuous at the point of discontinuity for NSLP eligibility (both at the free lunch cutoff and the reduced-price lunch cutoff).

As this assumption is fundamentally untestable, I tested for discontinuities in child sex, race, household size, public/private school attendance, and parent education at both cutoffs. Table 4 records the results of these tests. Most covariates do not display a statistically significant discontinuity at the cutoff. There is a marginally significant increase of 5.4% in parent HS graduation rate and an increase of 11.3% in the proportion of Hispanic students across the free lunch cutoff; across the reduced-price lunch cutoff, there are statistically significant drops in household size (0.998 people) and parent college graduation rate (4.5%) and an increase in the proportion of Black students (3.4%). Still, as the magnitude of these effects are relatively small, I will proceed with the RD analyses.

One other potential concern is that families may be eligible for other programs that have the same income cutoffs as NSLP, in which case the effect of NSLP on academic outcomes could also include the effects of those other programs. The School Breakfast Program (SBP), which provides free or reduced-price breakfasts, uses the same guidelines, so the effect of NSLP eligibility could also include some of the effects of SBP eligibility. Beyond SBP, however, I have not found any federal programs that use identical income guidelines.

It is possible that some households could deliberately report a lower income in order to qualify for either free or reduced-price lunches. Using the McCrary test for density right around the RD cutoffs, I aim to check if people are manipulating their responses to be below the cutoff. As shown in Figure 2, the McCrary test yields a discernibly higher density just below the free lunch cutoff; at the reduced-price lunch cutoff, the test indicates that there is a significantly higher density just above the cutoff. Since the p-values from the tests at both cutoffs are 0, both tests suggest that we should reject the null hypothesis that there is no manipulation in reporting household income, which ultimately increases the noise of RD. Still, I proceed with the RD analysis as the magnitude of these density differences, as seen in Figure 2, are relatively small.

Another source of noise is that income is collected in intervals such that it is not possible to determine household income to the nearest thousand. Instead, I have opted to use the midpoint of each interval. This means that some households that actually qualify for FRP lunch may not be eligible in the data and vice versa. In conjunction with the caveats discussed in the Data section, the efficacy of the RD analysis may be limited by the noisiness of the area surrounding the free and RP lunch cutoffs.

The Effect of NSLP Eligibility on Future Test Scores

I also check the effects of eligibility in year $t-1$ on test scores in year t for students in the 2010 cohort. Since data for the 1998 cohort is available for grades 1, 3, 5, and 8 only [4], I would like to check the effect of NSLP in year $t-2$ on academic outcomes in year t for this cohort, i.e. the effect of NSLP eligibility in grade 1 is on test scores in grade 3 and the effect of NSLP eligibility in grade 3 on test scores in grade 5.

For the 2010 cohort, redefine F_{it-1} and RP_{it-1} as follows:

$$F_{it-1} = \begin{cases} 1 & \text{if } p_{it-1} < 1.3 \\ 0 & \text{if } p_{it-1} \geq 1.3 \end{cases} \tag{6}$$

$$RP_{it-1} = \begin{cases} 1 & \text{if } p_{it-1} < 1.85 \\ 0 & \text{if } p_{it-1} \geq 1.85 \end{cases} \tag{7}$$

The RD specifications are entirely analogous to equations 2-4 and 6-8, but p_{it}, F_{it}, and RP_{it} are substituted for p_{it-1}, F_{it-1}, and RP_{it-1}. For the 1998 cohort, F_{it-2} and RP_{it-2} are defined as

$$F_{it-2} = \begin{cases} 1 & \text{if } p_{it-2} < 1.3 \\ 0 & \text{if } p_{it-2} \geq 1.3 \end{cases} \tag{8}$$

$$RP_{it-2} = \begin{cases} 1 & \text{if } p_{it-2} < 1.85 \\ 0 & \text{if } p_{it-2} \geq 1.85 \end{cases} \tag{9}$$

[4] See footnote 1.

Again, the RD specifications are analogous to equations 2-4 and 6-8, but with p_{it-2}, F_{it-2}, and RP_{it-2} used in place of each occurrence of p_{it}, F_{it}, and RP_{it} since I examine the effect of eligibility two years ago on current year test scores.

Results

The Effect of NSLP Eligibility on Current-Year Test Scores

To visualize the relationship between the dependent variable reading scores and free lunch eligibility, Figure 3 is the RD plot of reading scores on free lunch eligibility. It uses a MSE-optimal bandwidth of the same width on either side of the $c = 1.3FPL$ cutoff (i.e. the cutoff for free lunch eligibility is 130% of the federal poverty level) and a uniform kernel, and it allows for a quadratic p_{it} term and interactions between p_{it} and F_{it} or RP_{it} such that slopes can vary on either side of the free lunch eligibility cutoff. Figures 4 and 5 are the RD plots of math and science scores on free lunch eligibility; they also utilize the same cutoff, MSE-optimal bandwidths, uniform kernels, and quadratic and interaction terms.

The effect of free lunch eligibility in year t seems not to be associated with higher reading or math test scores in year t; instead, students who were not eligible for free lunch tended to have significantly higher reading and math scores. As shown in the first row of Table 5, students who were barely non-eligible scored an average of 12.91 points higher on reading and 17.41 points higher on math, which respectively represent increases of 11.5% and 19.3% from the average scores in those subjects. The point estimate for the free lunch eligibility effect on reading scores is marginally significant ($p < 0.01$) and the point estimate for the free lunch eligibility effect on math scores is significant ($p < 0.001$), suggesting that free lunch eligibility is associated with a decrease in reading and math scores.

Conversely, free lunch eligibility may have a statistically significant positive effect on science test scores. Combining the 1998 and 2010 cohorts, it seems like free lunch eligibility is associated with a 6.294 point increase in science test scores, which is 11.5% better than the average score of 54.87.

Figures 6-8 graphically depict the relationship between reading, math, and science scores and RP lunch eligibility. As with the RD analysis of test scores on free lunch eligibility, these plots utilize MSE-optimal

bandwidths, uniform kernels, and quadratic and interaction terms, but the cutoff is shifted to $c = 1.85FPL$ to reflect that the cutoff for reduced-price lunch is 185% of the federal poverty level. As shown in the second row of Table 5, the results for RP lunch eligibility suggest that RP lunch eligibility may have a negative effect on test scores. Students who were barely non-eligible for reduced-price lunch scored an average of 39.22 points (or 34.9%) higher than average on reading, 10.61 points (or 11.8%) higher on math, and 14.90 points (or 27.2%) higher on science. Given that these results are significant at the 1% level, the RD analysis implies that RP lunch eligibility decreases test scores.

The Effect of NSLP Eligibility on Future Test Scores

I run a regression discontinuity analysis on the effect of FRP eligibility in years $t-1$ and $t-2$ on test scores in year t. Table 6 depicts the effect of FRP lunch eligibility in year $t-1$ on test scores in year t; it includes the 2010 cohort only.[5] The first row shows the effects of free lunch eligibility. At 95% confidence, there is no significant difference detected in reading scores between students who were right above and right below the free lunch cutoff in the previous year; the 95% confidence interval suggests the true free lunch eligibility effect is probably somewhere between a decrease of 23.56 points (20.9%) to an increase of 13.25 points (11.8%). For math and science, however, students who were barely eligible for free lunch in the previous year scored on average 38.48 points (42.6%) higher and 41.91 points (76.4%) higher, respectively, than students who were barely ineligible. Both of these effects are significant at $p = 0.05$, suggesting that free lunch eligibility may have a significant positive effect on math and science scores for the following year.

In the second row, reduced-price eligibility in year $t-1$ is not associated with any significant differences in math or science test scores in year t; though both point estimates are slightly positive (suggesting that RP lunch eligibility could be associated with lower test scores), both have 95% confidence intervals that contain 0 such that the effects are statistically insignificant. Students who were barely ineligible for RP lunch in the previous year scored on average 3.78 points (3.4%) higher on reading than those who were eligible, which is a significant increase at the 5% level. The RD coefficients in Table 6 suggest that although reduced-price lunch eligibility in the previous year may not improve test scores, free lunch eligibility may have a significantly positive effect on test scores in the current year.

[5] See footnote 1 for the rationale on limiting to this cohort.

Table 7 shows the effect of FRP lunch eligibility in year $t-2$ on test scores in year t, i.e. the effect of eligibility in grade 1 on test scores in grade 3 or the effect of eligibility in grade 3 on test scores in grade 5. This table limits the sample to the 1998 cohort. In the first row, students who were barely ineligible for free lunch two years ago tend to score 15.40 points (13.7%) higher on reading than peers who were barely ineligible, and they scored an average of 7.77 points (14.2%) higher on science. There is no statistically significant difference in math scores between students who were right above and right below the free lunch cutoff two years ago, as the 95% confidence interval suggests the free lunch eligibility effect on math scores is likely between a 1.28 point (1.4%) decrease and a 9.09 point (10.1%) increase. In the second row, students who were barely ineligible for reduced-price lunch two years ago tended to score 8.70 points (7.7%) higher on reading, 9.73 points (10.8%) higher on math, and 5.78 points (10.5%) higher on science; all of these effects are statistically significant at $p = 0.05$. Essentially, my research suggests that being eligible for NSLP two years ago may not improve reading, math, or science scores in the current year.

Robustness Checks

Following RD analyses at both the free and reduced-price cutoffs, I conduct a placebo test at 75% of the federal poverty level (i.e. below the free lunch cutoff of 130% FPL) to evaluate whether an effect is found; as there is no significance to this randomly selected cutoff, no statistically significant effect should be found, so the presence of one would suggest problems in data cleaning or empirical design.

As shown in Table 8, it appears that being just below this placebo cutoff is associated with significant test score improvements. In reading, math, and science, students who are just above the placebo cutoff have test scores that are significantly higher than those who are not; the magnitude of this effect is large (13.4%, 17.1%, and 19.4% higher than average, respectively) and highly significant ($p < 0.001$). To my knowledge, there are no student programs that use 75% of the federal poverty level as an eligibility cutoff, so there should be no statistically significant difference in test scores of students on either side. As such, the presence of these effects essentially invalidates the placebo test, though I am not sure why they exist.

Recall that household incomes are recorded as intervals in the public dataset; in all preceding analysis, I have opted to use the midpoint value of each interval. As a separate robustness check, I replace the

midpoints with the lower end of each interval and re-run RD analyses of test scores in year t across the free and reduced-price lunch eligibility cutoffs in year t. The results can be seen in Table 9. Compared to the results generated using midpoint values in Table 5, the effects of free lunch eligibility are similar; however, the effect on science scores is no longer positive (i.e. at 95% confidence those who are barely eligible no longer score significantly higher than those who are barely ineligible), and at 5% significance, reading scores are higher (by 2.53 points or 2.3%) among those who are barely eligible for free lunch. By using the lower end of each end point, some students who were previously considered ineligible for RP lunch are now eligible. As seen in the second row of Table 9, RP lunch eligibility is associated with a 10.34 point (9.2%) improvement in reading, a 7.39 point (8.2%) improvement in math, and a 3.06 point (5.6%) improvement in science, all of which are significant at $p = 0.05$.

Conversely, Table 10 depicts the effects of FRP eligibility on test scores when the upper end of each income interval is used to generate the running variable. As opposed to Table 5, free lunch eligibility is no longer associated with positive effects on test scores; in reading, math, and science, students who are ineligible for free lunch score significantly higher (at the 0.01% significance level) than those who are eligible. In Table 10, RP lunch eligibility seems to be associated with significantly better reading and math scores, though not with significantly better science scores. Together, Tables 9 and 10 suggest that the robustness of the RD analysis may be negatively impacted by the interval structure of the household income variable, but the broadest takeaways – there may be improvements due to free lunch eligibility, and the effects of free lunch eligibility likely have a larger magnitude than those of reduced-price lunch eligibility – seem to hold.

Tables 11 and 12 demonstrate the effect of FRP lunch eligibility in year t on test scores in year t using a linear relationship and a cubic relationship, respectively. The sign of each estimate remains the same as those in Table 5 (the effect of FRP lunch eligibility in year t on test scores in year t using a quadratic relationship): reading, math, and science test scores are, on average, lower when students become eligible for FRP lunch except for science test scores, which tend to increase when students become eligible for free lunch (though not when they become eligible for reduced-price lunch). However, the estimates of the effect of free lunch eligibility on science test scores in Tables 11 and 12 differ from that in Table 5 in that 0 is contained in both of their 95% confidence intervals. Note that the standard errors are much larger for the

linear model than for the cubic one (since there is less flexibility in the linear model to fit the RD regression to the data) such that the 95% confidence interval for the free lunch eligibility effect on science test scores is between a 3.35 point (6.1%) decrease and 25.93 point (47.3%) increase in the linear model and between a 4.17 point (7.6%) decrease and 12.23 point (22.3%) increase in the cubic model.

Therefore, though point estimates remain negative, at a 95% confidence level we cannot conclude that free lunch eligibility has a significant positive effect on science test scores when the relationship between p_{it} and S_{it} is constrained to be linear or allowed to be cubic.

Conclusion

The results suggest that free or reduced-price lunch eligibility alone is not a major contributor to improved academic outcomes. If a student is eligible for free lunch in a given year, my analysis indicates she may have an improved science test score, though her math and reading scores will still likely be lower than her peers with slightly higher household incomes. For the other cutoff, it appears that eligibility for reduced-price lunch is not associated with significant improvements in math, reading, or science test scores.

It is plausible that the magnitude of effects would be greater if household income were recorded as dollar values rather than intervals, as it is very likely that some households that are ineligible for FRP lunch are marked as eligible in my RD analysis, and vice versa. Even though the Early Childhood Longitudinal Study's public data does not enable this analysis, it is possible that NSLP enrollment – not just NSLP eligibility – could move the needle on test scores in a more consistently significant manner. Still, this research suggests that expanding free or reduced-price lunch eligibility is likely not the most effective avenue for improving test scores; rather, the motivation for extending NSLP access may primarily come from its nutritional effects.

Future research on the academic effects of the National School Lunch Program could focus on finding a set of panel data that records NSLP enrollment, as eligibility rates are higher than enrollment rates. Alternatively, it would be interesting to consider the effects of universal free school lunch; a small group of schools have made school lunch free for all students without the need to apply, and examining the changes to student health, behavior, and academic performance (whether through grades or standardized testing) could yield insight on the efficacy of these programs, as well as inform continued debate on whether such

programs should be expanded.

Appendix: Tables and Figures

Table 1: Summary statistics (all)

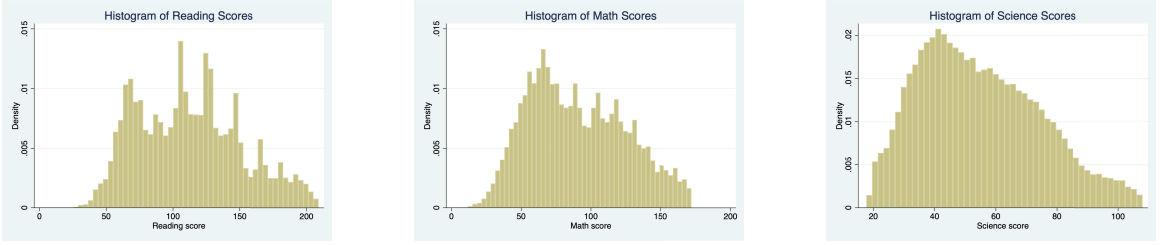
	Mean	SD	Min	Max	Count
Female	0.49	0.50	0	1	98058
Grade	2.55	2.24	0	8	98058
Reading score	112.40	38.84	25	209	92779
Math score	90.29	34.87	12	172	93137
Science score	54.87	19.79	18	108	77983
Household size	4.62	1.39	2	18	98058
Public	0.85	0.36	0	1	94497
Household income	66.33	52.94	5	200	98058
Free lunch eligible	0.29	0.46	0	1	98058
RP lunch eligible	0.41	0.49	0	1	98058
Percent of FPL	3.08	2.51	0	18	98058
White	0.51	0.50	0	1	95891
Black	0.08	0.27	0	1	95891
Hispanic	0.16	0.36	0	1	95891
Asian/Pacific Islander	0.07	0.25	0	1	95891
Less than HS	0.03	0.17	0	1	92231
Some HS	0.06	0.24	0	1	92231
HS	0.22	0.41	0	1	92231
Some college	0.32	0.47	0	1	92231
College	0.37	0.48	0	1	92231
Observations	98058				

Household income reported in $1000s of dollars.

Percent of FPL is the ratio of household income to the federal poverty level.

The last five schooling variables refer to the highest level of parent education.

Data from the 1998 and 2010 cohorts of the Early Childhood Longitudinal Study.



(a) Histogram of reading scores (b) Histogram of math scores (c) Histogram of science scores

Figure 1: Distribution of test scores

Test scores are slightly skewed to the right. Note that grading ranges are not consistent across subjects so x-axis ranges differ.

Table 2: Summary statistics (free lunch eligible)

	Mean	SD	Min	Max	Count
Female	0.49	0.50	0	1	28739
Grade	2.25	2.06	0	8	28739
Reading score	97.91	33.62	26	209	26758
Math score	77.52	31.00	12	171	27025
Science score	45.72	17.13	18	108	23145
Household size	5.07	1.75	2	18	28739
Public	0.96	0.19	0	1	27598
Household income	18.01	8.89	5	88	28739
Free lunch eligible	1.00	0.00	1	1	28739
RP lunch eligible	1.00	0.00	1	1	28739
Percent of FPL	0.75	0.33	0	1	28739
White	0.22	0.41	0	1	27616
Black	0.13	0.33	0	1	27616
Hispanic	0.29	0.45	0	1	27616
Asian/Pacific Islander	0.07	0.25	0	1	27616
Less than HS	0.09	0.28	0	1	26986
Some HS	0.16	0.36	0	1	26986
HS	0.37	0.48	0	1	26986
Some college	0.30	0.46	0	1	26986
College	0.08	0.28	0	1	26986
Observations	28739				

All households in this subsample are eligible for free lunch (< 130% FPL).

See Table 1 for variable definitions and data time frame.

Table 3: Summary statistics (reduced-price lunch eligible)

	Mean	SD	Min	Max	Count
Female	0.49	0.50	0	1	40323
Grade	2.37	2.11	0	8	40323
Reading score	101.45	35.04	26	209	37662
Math score	80.56	32.06	12	172	37974
Science score	47.90	17.84	18	108	32222
Household size	4.97	1.66	2	18	40323
Public	0.95	0.22	0	1	38703
Household income	23.00	11.95	5	88	40323
Free lunch eligible	0.71	0.45	0	1	40323
RP lunch eligible	1.00	0.00	1	1	40323
Percent of FPL	0.99	0.47	0	2	40323
White	0.28	0.45	0	1	38948
Black	0.12	0.32	0	1	38948
Hispanic	0.26	0.44	0	1	38948
Asian/Pacific Islander	0.07	0.25	0	1	38948
Less than HS	0.07	0.25	0	1	38054
Some HS	0.13	0.33	0	1	38054
HS	0.35	0.48	0	1	38054
Some college	0.34	0.47	0	1	38054
College	0.11	0.32	0	1	38054
Observations	40323				

All households in this subsample are eligible for reduced-price lunch (< 185% FPL).

See Table 1 for variable definitions and data time frame.

Table 4: Covariate Discontinuities Across the Free and RP Lunch Cutoffs

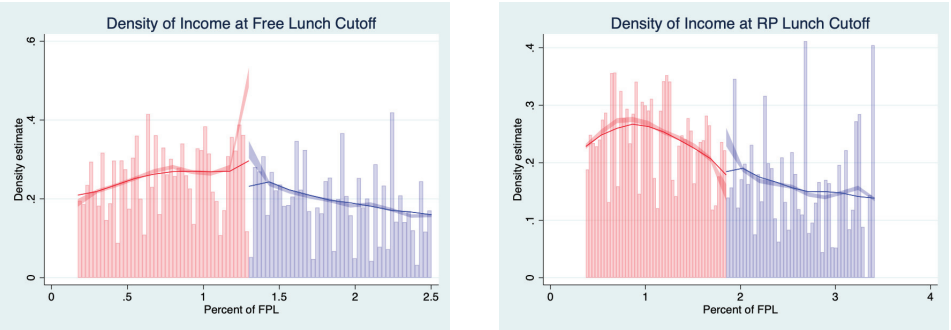
	(1) Household size	(2) Female	(3) White	(4) Black	(5) Hispanic	(6) Public	(7) HS	(8) College
Free lunch eligibility	0.0825	0.0428	-0.0256	0.0303	0.113**	0.0245	0.0541*	-0.0147
	(0.383)	(0.0242)	(0.0353)	(0.0253)	(0.0348)	(0.0143)	(0.0245)	(0.0135)
RP lunch eligibility	-0.998***	0.0228	0.0149	0.0340**	0.00529	-0.00536	-0.00434	-0.0448**
	(0.0921)	(0.0192)	(0.0228)	(0.0132)	(0.0168)	(0.0145)	(0.0195)	(0.0154)
Observations	98058	98058	95891	95891	95891	94497	92231	92231

Standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Coefficients are RD estimates using covariates as response variables. MSE-optimal bandwidth and uniform kernel used.

Quadratic in running variable p_{it} (ratio of household income to federal poverty level).

Interaction between p_{it} and eligibility and p_{it}^2 and eligibility.



(a) (b)

Figure 2: McCrary density tests at the free and RP lunch cutoffs

Manipulation tests of the running variable (ratio of HH income to FPL) use a local quadratic approximation, unrestricted density estimation, and a uniform kernel. Vertical lines are at the free (130% FPL) and RP (185% FPL) lunch cutoffs, respectively. Discontinuity magnitudes represent differences in income density. Magnitudes are statistically significant but relatively small.

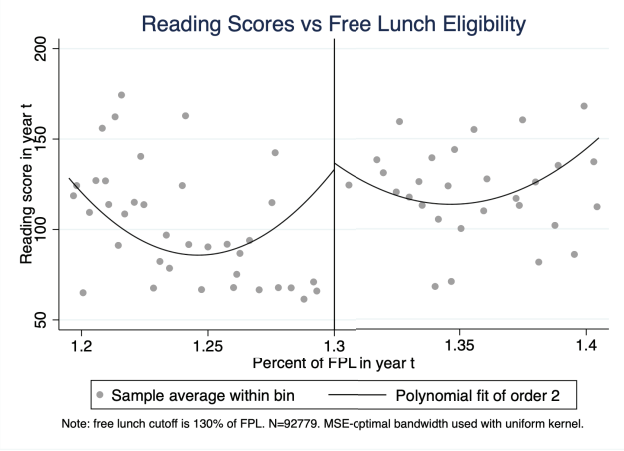


Figure 3: Graphical RD analysis of reading scores at free lunch cutoff (130% FPL)

Points represent averages within bins. Fitted line is quadratic in the running variable (ratio of HH income to FPL), with interactions between the running variable and eligibility. Vertical line represents free lunch cutoff. MSE-optimal bandwidth and uniform kernel used. Free lunch eligibility is associated with a 12.91 point (11.5%) decrease in reading scores.

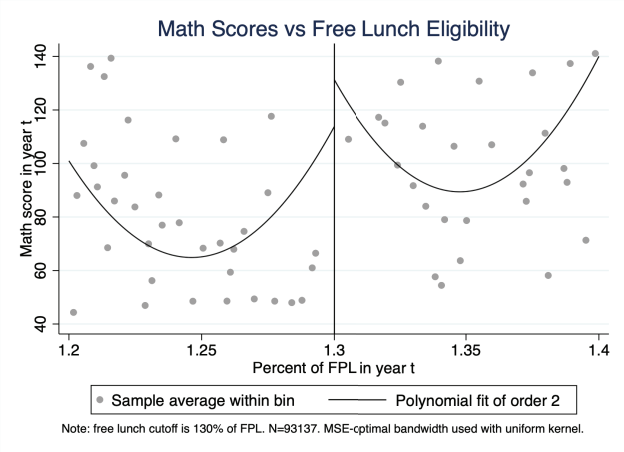


Figure 4: Graphical RD analysis of math scores at free lunch cutoff (130% FPL)

Points represent averages within bins. Fitted line is quadratic in the running variable (ratio of HH income to FPL), with interactions between the running variable and eligibility. Vertical line represents free lunch cutoff. MSE-optimal bandwidth and uniform kernel used. Free lunch eligibility is associated with a 17.41 point (19.3%) decrease in math scores.

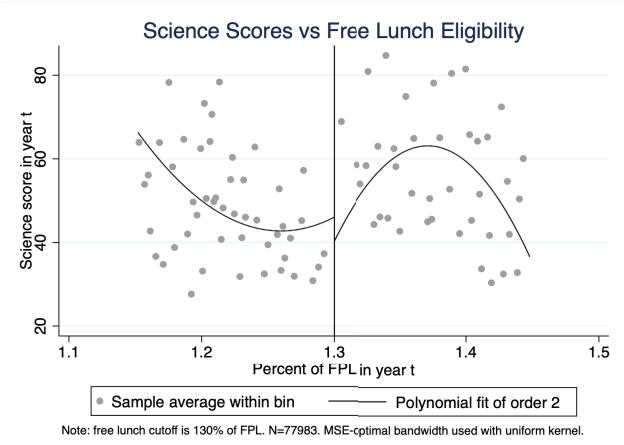


Figure 5: Graphical RD analysis of science scores at free lunch cutoff (130% FPL)

Points represent averages within bins. Fitted line is quadratic in the running variable (ratio of HH income to FPL), with interactions between the running variable and eligibility. Vertical line represents free lunch cutoff. MSE-optimal bandwidth and uniform kernel used. Free lunch eligibility is associated with a 6.294 point (11.5%) increase in science scores.

Table 5: Effect of FRP Lunch Eligibility in Year t on Test Scores in Year t

	(1) Reading score	(2) Math score	(3) Science score
Free lunch eligibility	12.91**	17.41***	-6.294***
	(4.164)	(4.041)	(1.680)
RP lunch eligibility	39.22***	10.61***	14.40***
	(1.753)	(1.238)	(1.131)
Observations	92779	93137	77983

Standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Coefficients are RD estimates. MSE-optimal bandwidth and uniform kernel used.
Quadratic in running variable p_{it} (ratio of household income to FPL).
Interaction between p_{it} and eligibility and p_{it}^2 and eligibility.

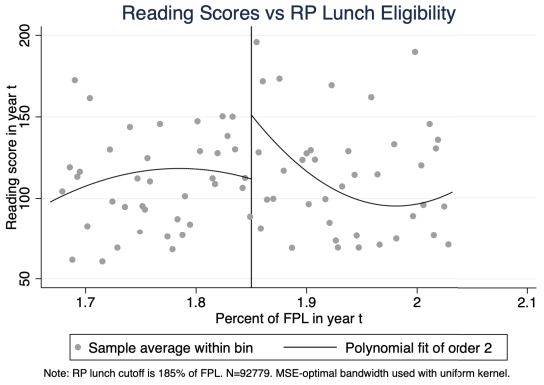


Figure 6: Graphical RD analysis of reading scores at RP lunch cutoff (185% FPL)

Points represent averages within bins. Fitted line is quadratic in the running variable (ratio of HH income to FPL), with interactions between the running variable and eligibility. Vertical line represents RP lunch cutoff. MSE-optimal bandwidth and uniform kernel used. RP lunch eligibility is associated with a 39.22 point (34.9%) decrease in reading scores.

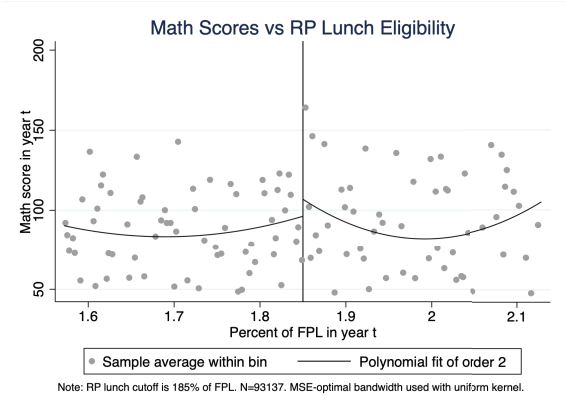


Figure 7: Graphical RD analysis of math scores at RP lunch cutoff (185% FPL)

Points represent averages within bins. Fitted line is quadratic in the running variable (ratio of HH income to FPL), with interactions between the running variable and eligibility. Vertical line represents RP lunch cutoff. MSE-optimal bandwidth and uniform kernel used. RP lunch eligibility is associated with a 10.61 point (11.8%) decrease in math scores.

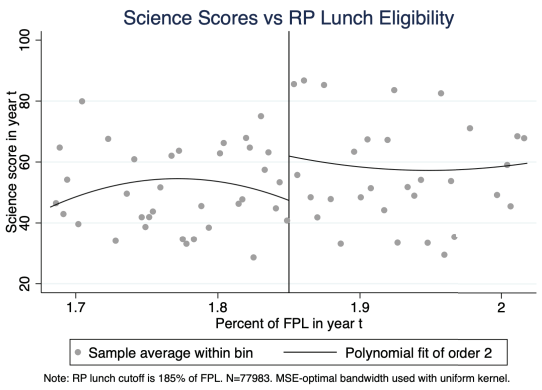


Figure 8: Graphical RD analysis of science scores at RP lunch cutoff (185% FPL)

Points represent averages within bins. Fitted line is quadratic in the running variable (ratio of HH income to FPL), with interactions between the running variable and eligibility. Vertical line represents RP lunch cutoff. MSE-optimal bandwidth and uniform kernel used. RP lunch eligibility is associated with a 14.40 point (27.2%) decrease in science scores.

Table 6: Effect of FRP Lunch Eligibility in Year t-1 on Test Scores in Year t

	(1) Reading score	(2) Math score	(3) Science score
Free lunch eligibility	5.162	-38.48**	-41.91*
	(9.393)	(12.36)	(16.57)
RP lunch eligibility	3.776*	0.811	1.965
	(1.648)	(1.781)	(1.037)
Observations	30340	30334	30301

Standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

2010 cohort only. Coefficients are RD estimates.

MSE-optimal bandwidth and uniform kernel used.

Quadratic in running variable p_{it} (ratio of household income to FPL).

Interaction between p_{it} and eligibility and p_{it}^2 and eligibility.

Table 7: Effect of FRP Lunch Eligibility in Year t-2 on Test Scores in Year t

	(1) Reading score	(2) Math score	(3) Science score
Free lunch eligibility	15.40**	-3.903	7.766**
	(4.951)	(2.646)	(2.696)
RP lunch eligibility	8.701**	9.725***	5.778**
	(3.114)	(2.901)	(1.844)
Observations	21337	21413	21403

Standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

1998 cohort only. Coefficients are RD estimates.

MSE-optimal bandwidth and uniform kernel used.

Quadratic in running variable p_{it} (ratio of household income to FPL).

Interaction between p_{it} and eligibility and p_{it}^2 and eligibility.

Table 8: Placebo Test

	(1) Reading score	(2) Math score	(3) Science score
Placebo cutoff	-15.05***	-15.42***	-10.65***
	(1.230)	(1.096)	(0.664)
Observations	92779	93137	77983

Standard error in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
75% of FPL chosen as a placebo cutoff. Coefficients are RD estimates.
MSE-optimal bandwidth and uniform kernel used.
Quadratic in running variable p_{it} (ratio of household income to FPL).
Interaction between p_{it} and eligibility and p_{it}^2 and eligibility.

Table 9: Effect of FRP Lunch Eligibility in Year t on Test Scores in Year t Using Lower End of Each Income Bin

	(1) Reading score	(2) Math score	(3) Science score
Free lunch eligibility	-2.530*	27.00***	10.77***
	(1.263)	(1.977)	(1.242)
RP lunch eligibility	-10.34***	-7.386***	-3.057***
	(2.266)	(1.964)	(0.801)
Observations	92779	93137	77983

Standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Coefficients are RD estimates. MSE-optimal bandwidth and uniform kernel used.
Quadratic in running variable p_{it} (ratio of household income to FPL).
Interaction between p_{it} and eligibility and p_{it}^2 and eligibility.
Lower end of each income bin used.

Table 10: Effect of FRP Lunch Eligibility in Year t on Test Scores in Year t Using Upper End of Each Income Bin

	(1) Reading score	(2) Math score	(3) Science score
Free lunch eligibility	24.85***	19.75***	9.410***
	(2.013)	(1.946)	(1.134)
RP lunch eligibility	-42.05***	-8.678***	1.900**
	(3.142)	(2.345)	(0.688)
Observations	92779	93137	77983

Standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Coefficients are RD estimates. MSE-optimal bandwidth and uniform kernel used.
Quadratic in running variable p_{it} (ratio of household income to FPL).
Interaction between p_{it} and eligibility and p_{it}^2 and eligibility.
Upper end of each income bin used.

Table 11: Effect of FRP Lunch Eligibility in Year t on Test Scores in Year t Using Linear Relationship

	(1) Reading score	(2) Math score	(3) Science score
Free lunch eligibility	19.65***	33.42***	-11.29
	(5.821)	(3.329)	(7.469)
RP lunch eligibility	25.57***	15.16***	9.565***
	(1.332)	(1.116)	(0.726)
Observations	92779	93137	77983

Standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Coefficients are RD estimates. MSE-optimal bandwidth and uniform kernel used.
Linear in running variable p_{it} (ratio of household income to FPL).
Interaction between p_{it} and eligibility.

Table 12: Effect of FRP Lunch Eligibility in Year t on Test Scores in Year t Using Cubic Relationship

	(1) Reading score	(2) Math score	(3) Science score
Free lunch eligibility	79.53***	89.15***	-4.032
	(7.762)	(6.618)	(4.185)
RP lunch eligibility	39.69***	44.51***	3.015**
	(1.715)	(1.859)	(0.923)
Observations	92779	93137	77983

Standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Coefficients are RD estimates. MSE-optimal bandwidth and uniform kernel used.
Cubic in running variable p_{it} (ratio of household income to FPL).
Interaction between p_{it} and eligibility, p_{it}^2 and eligibility, and p_{it}^3 and eligibility.

References

Calonico, S., Cattaneo, M., and Farrell, M (2017). "rdrobust: Software for regression-discontinuity designs." The Stata Journal 17(2): 372-404.

Dotter, D. (2013). "Breakfast at the Desk: The Impact of Universal Breakfast Programs on Academic Performance." Mathematica Policy Research.

Dunifon, R., and Kowaleski-Jones, L. (2003). "Associations between Participation in the National School Lunch Program, Food Insecurity, and Child Well-Being." Social Service Review 77(1): 72–92.

Frisvold, D.E. (2015). "Nutrition and Cognitive Achievement: An Evaluation of the School Breakfast Program." Journal of Public Economics, 124: 91-104.

Gordon, N. and Ruffini, K. (2018). "School Nutrition and Student Discipline: Effects of Schoolwide Free Meals." NBER Working Paper 24986. September.

Hinrichs, P. (2010). "The Effects of the National School Lunch Program on Education and Health," Journal of Policy Analysis and Management, 29(3): 479-505.

Imberman, S.A. and Kugler, A. (2014). "The Effect of Providing Breakfast on Achievement and Attendance: Evidence from an In-Class Breakfast Program." Journal of Policy Analysis and Management, 33(3): 669-699.

McCrary, J (2008). "Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test." Journal of Econometrics, 142(2): 698-714.

Schanzenbach, D.W. and Zaki, M. (2014). "Expanding the School Breakfast Program: Impacts on Children's Consumption, Nutrition, and Health." NBER Working Paper 20308. July.

The Effect of Conservation Reserve Program Grants on Organic Farming

Julia Caravias [1]
December 9, 2021

Massachusetts Institute of Technology

Abstract

The USDA's Conservation Reserve Program (CRP) offers payments to farmers to remove environmentally sensitive land from agricultural production and instead implement restorative practices. This program has been successful in improving soil health and environmental quality, but after these land grants expire it is important to understand if farmers continue to use regenerative practices on their land, which I estimate by exploring the prevalence of organic farming. In this paper I explore the causal effect of CRP grants on the prevalence of certified organic farms. Organic farms are required to use more environmentally friendly practices, which makes organic certification a good indicator of regenerative farming. I use a two-way fixed effects model to study the effect of having a high share of farmland awarded CRP grants on the share of organic farmland at the state level. My estimate for this effect is that having a high share of CRP land is associated with an increase of 0.0264 in the share of organic farmland. However, the large standard error of this estimate makes the result too imprecise to draw conclusions about the causal effect of CRP land grants on organic farming.

[1] Thank you to Professor Dave Donaldson and Kelsey Moran for their guidance on this paper.

Introduction

Soil health is critically important for the livelihood of farmlands and the environment. Existing research on the CRP provides strong evidence that the program has immediate positive benefits for the environment. Because the CRP is such a critical program for improving soil health in the US, the impact of CRP grants on farmers' use of the land after enrollment is of great interest. Organic farming is an appropriate indicator of environmentally friendly land use since organic farms tend to implement regenerative practices and have better soil outcomes. If there is a causal effect of enrolling land in the CRP on converting that land to organic farms, this would suggest that the program positively impacts farmers' post-enrollment decisions to use more regenerative land use practices.

This paper researches the causal effect of acreage covered by CRP grants on the acreage of certified organic farms. Using data on the acreage of land covered by CRP grants and acreage of certified organic farmland at the state level, I apply a difference-in-differences approach with a two-way fixed effects model to estimate the effect of increased CRP land on organic farming.

There is extensive research on the impact of CRP grants on environmental restoration which demonstrates that land enrolled in the CRP has the highest amount of environmental degradation and is thus most likely to improve from the program (Fleming 2004). In addition, there is a large measured effect of the CRP on the environmental health of cropland and organic soil carbon is higher after CRP enrollment (Gebhart, Johnson, Mayeux, Polley 1994) .

Previous research has explored land use practices after CRP contracts expire, however the effect of CRP enrollment specifically on organic farming has yet to be studied. Research using mail surveys have investigated the land use decisions of farmers after CRP enrollment and found that although there is a high interest among farmers to re-enroll their land in the CRP after

grant expiration, many farmers reported being unable to do so (Barnes et al 2020). This study finds that 28.3% of landowners reported converting the field to dryland crops and 61.9% reported leaving the majority of the field in grasslands while only 4.7% reported re-enrolling the field into another conservation program. (Morefield 2016) uses geospatial data from the USDA Farm Service Agency to determine that for years 2010-2013 almost 30% of expiring CRP land returned to the production of corn, soy, winter and spring wheat, and sorghum in a 12-state, Midwestern region of the U.S. This research also finds that specifically designated wildlife habitats, grasslands, and wetland areas were the top post-CRP retirement uses. Only about 3% of expiring land shifted into similar conservation programs.

In this paper, I first provide context about the policy relevance of the CRP, the process by which land is selected for CRP grants, and how organic farming relates to the CRP and soil health. I provide a detailed description of the data used in my analysis from the USDA Farm Service Agency (FSA) and the USDA Economic Research Service. Finally, I describe the difference-in-differences approach used to estimate the effect over time on states that were awarded a high share of CRP grants and discuss my findings.

Policy Background

Enacted in 1985, the CRP is the USDA's oldest and largest environmental protection program. Contracts for the CRP are awarded for 10-15 years periods and awards are selected based on demonstrated environmental sensitivity. During the award period, farmers are paid to remove environmentally sensitive land from agricultural production and instead plant species that improve environmental health and quality.

Currently, the CRP protects over 20 million acres of agricultural land and the enrolled land mitigates more than 12 million metric tons of carbon dioxide equivalents (USDA FSA). The program has been highly successful in establishing land cover to improve water quality, prevent soil erosion and reduce loss of wildlife habitat (Swan, Easter, Paustian, 2018).

Soil health is a critically important factor for the long term livelihood of farmlands and the surrounding environment. However, practices such as tilling, pesticide use, and nitrogen fertilizer use to keep up with demand for high yields has led to the degradation of soil health. Improved soil health has many benefits, including improved crop yields, better crop disease resilience, and ecosystem improvements for local wildlife populations. Moreover, farmland soil has the capacity to sequester atmospheric carbon, which is why sustainable agriculture has gained attention as a promising climate solution.

This paper explores the effect of participation in the CRP on the use of sustainable agricultural practices, measured through the prevalence of organic farming. This is important because the use of extractive farming methods can reverse the soil health benefits gained from the CRP. Thus, it is an interesting research question to explore if the CRP influences post-enrollment land use decisions.

In this paper, I use organic farm certification to identify farms that have adopted sustainable agricultural measures. Certified organic farms by the USDA are required to promote ecological balance, conserve biodiversity and foster the cycling of resources (EPA, 2021). This can entail the use of cover crops, crop rotations, green manures, elimination of synthetic pesticides and fertilizers, and soil and water conservation. Organic farming significantly improves soil health outcomes and has great environmental benefits (EPA, 2021).

There is anecdotal evidence suggesting that incentives exist for CRP land to become certified as an organic farm post-enrollment. Organic farm certification requires that land is removed from conventional farming methods and is not treated with synthetic fertilizer or herbicide for a three year period (SARE Outreach, 2003). Since costs related to this transition period are a major barrier to organic adoption, expiring lands from the CRP are often strong candidates for organic certification since farmers are provided financial support in the processes of removing land from agricultural production (North Dakota State University, 2016). This paper looks specifically at the relationship between states with high CRP participation and organic farm certification, since states with higher amounts of CRP enrollment presumably have higher amounts of land with CRP contracts expiring.

Data

For my explanatory variable, I use data from the USDA Farm Service Agency (FSA) on the number of acres of land that have received CRP grants from 1997-2011 for each state. The FSA awards and oversees CRP land grants and publishes historic data on acres covered by CRP grants starting in 1986. CRP grants are awarded to various types of land including grasslands, croplands, and ranches and the data used reflects all types of grants awarded.

For my response variable, I use data from the USDA Economic Research Service collected from USDA-accredited state and private certification groups to measure organic farmland acreage. This data gives the acres of certified organic farmland for each state for 1997, 2000-2008, 2010-2011. Note that there is no data available for 1998, 1999, and 2009 because this data is collected from organic farmland surveys that are not conducted on a regular basis.

I use data from the 2017 USDA agricultural census on the total acres of farmland by state to get the share of CRP land and the share of organic farmland by state. I divide the total acres of

CRP land and organic farmland by total acres of farmland by state for each year to calculate the CRP share and organic share variables used in my analysis. Using a single measure of total acres of farmland for all time periods is an appropriate approximation since there is little variation in the acres of farmland by state in this time window used in my analysis. By using the share of land rather than quantities in acres, I am better able to compare small and large states which have varying amounts of farmland available.

Table 1 presents summary statistics on the percent of farmland covered by CRP grants and the percent of farmland that is certified as organic broken down by region. These statistics are averaged across all states for all years used in my analysis (1997, 2000-2008, 2011). This table demonstrates that there is notable variability in the share of CRP land and organic farms across regions. However, across the time period within a given state, there is generally low variability.

Figure 1 and Figure 2 show the geographic average share of land covered by CRP grants and share of organic farms at the state level, respectively. Figure 1 and Table 1 demonstrate that within regions the average share of CRP land is generally homogenous. Figure 2 demonstrates that most states have a relatively low average share of organic farms and a few select outliers with higher share including Alaska, Vermont, Maine, New York, California and Connecticut.

Empirical Methods

Using a difference-in-differences approach, I study the effect of having a high share of farmland covered by CRP grants on the share of organic farmland at the state level. Specifically, I apply a two-way fixed effect model with the below regression:

$$Y_{s,t} = \gamma_s + \delta_t + \beta X_{s,t} + \varepsilon_{s,t}$$

$Y_{s,t}$= share of organic farm land in state *s* for year *t*

$X_{s,t}$= indicator variable that for having a high share of CRP grant land in year *t*

γ_s= location effect for state *s*

δ_t= time effect for year *t*

β= effect of CRP grant awards on organic farmland

$\varepsilon_{s,t}$= error term in state *s* for year *t*

In this model, I define an indicator variable $X_{s,t}$ that is equal to 1 if the share of farmland covered by the CRP in state *s* for year *t* is greater than 0.03 and is equal to 0 otherwise. That is, $X_{s,t}$ indicates that there is a high share of CRP land for a given observation. The threshold of 0.03 is approximately close to the nationwide average of 0.0286. Since the variation of the share of CRP grants is low within states, this chosen threshold does well in providing a cutoff for a sizable increase in CRP land coverage. Here, the coefficient β is the causal effect of interest and can be interpreted as the effect of going from a low to high share of CRP land acreage on the share of organic farms.

To control for variation across location and time, I also include γ_s, the state fixed effect, and δ_t, the year fixed effect. I cluster standard errors at the state level in order to be able to make a broader conclusion about the effect of CRP land grants in the U.S. since my data on CRP land share and organic farming land share is sampled at the state level.

An ideal experiment to study the effect of CRP grants on organic farming would randomly assign CRP grants to a representative sample of farms while maintaining a control group of farms that do not receive grants. Because this experiment is not feasible, I instead use

the difference-in-difference method to compare the change in prevalence of organic farming as the amount of land covered by the CRP varies within the state across time.

Difference-in-differences is an approach for quasi-experimental studies that uses longitudinal data to approximate the counterfactual of treatment and estimate a causal effect. Difference-in-differences compares outcomes over time between a population that receives treatment to a control population to estimate the effect of an intervention. This approach is commonly used to evaluate policies and its application is appropriate to study the effect of the CRP.

Difference-in-differences analysis relies on the parallel trends assumption. That is, I assume that treated states would have a similar amount of organic farms to the control states in the absence of treatment. Here the treatment is receiving a high share of CRP grants. I use an event study analysis shown in Figure 4 to determine the strength of the parallel trends assumption. To generate this figure, I regressed an indicator for having a high share of CRP land on the share of organic farms, lagged by the year in which the threshold of a high CRP land share was reached. Figure 4 plots the resulting coefficients for each year prior to and after a state has reached the threshold of a high share of CRP land. The coefficients can be interpreted as the incremental change in organic farmland share as CRP share goes from a low to high amount. The regression equation for the event study is given below:

$$Y_{s,t} = \gamma_s + \delta_t + \sum_{j=-6}^{6} \beta_j d_{i,t-j} + \varepsilon_{s,t}$$

$Y_{s,t}$= share of organic farm land in state *s* for year *t*

γ_s= location effect for state *s*

δ_t= time effect for year *t*

β_j= incremental effect of change in CRP grant awards for year *t* on organic farmland (casual effect of interest)

$d_{s,t}$= indicator variable that equals 1 in the year of treatment and 0 otherwise

$\varepsilon_{s,t}$= error term in state *s* for year *t*

If the parallel trends assumption holds, the resulting event study would have beta values approximately close to zero for years prior to treatment and a distinct change in the beta values after treatment. This would indicate that there is no continuous trend occurring within the data prior to treatment.

Figure 4 shows that there is no clear trend in the beta coefficients and the pre-treatment coefficients appear to be approximately zero. Examining the pre-treatment betas in Figure 4 verifies the expected pattern for the parallel trends assumption. That is, the coefficients have a notable change after the treatment year and do not follow a continuous trend in the pre- and post-treatment years. This suggests that treatment and control states would have experienced the same change in organic farming in the absence of treatment. However, the estimates for these coefficients are not precise enough to indicate that they are statistically different from zero. As such, it is not possible to conclude that the parallel trends assumption holds.

This application of a difference-in-differences approach has some clear limitations. It would be ideal to study the effect of CRP grants at the individual farm level in order to track which specific land is converted to organic farmland. However, since data is only available at the state level, I assume that changes in CRP farmland and organic farmland are due to the same farms converting to organic farms, rather than, for instance, new organic farms being established. This seems like a reasonable assumption since the proportion of farmland across this time window remains relatively constant according to historical census data.

Results

The result of my two-way event study estimates that the effect of the share of CRP land on the share of organic farmland is centered on 0.0264 and lies between -0.0281 and 0.0809 with 95% confidence. As such, I cannot reject the null hypothesis that beta is zero at standard levels because the results are too imprecise to draw conclusions about the causal effect. To reiterate, the beta coefficient of 0.0264 is interpreted as the effect of having a high share of CRP acreage on the share of organic farmland by 0.0264. The results of this regression are shown in Table 2, which highlights the large standard errors of the estimate.

Given that the average share of organic farmland for the entire country is 0.0162, the magnitude of the beta coefficient has notable economic significance. This implies that at the state level, having a high share of CRP land acreage could increase the share of organic farm acreage by almost 60% of the mean. It is also notable that with 95% confidence we can rule out that the effect of CRP grants on organic farming is larger than 0.0809. This is surprising since there is evidence that after enrollment in the CRP, farmers' have incentives to convert their land to

organic farms so one might expect that the effect of having a high share of CRP land on the prevalence of organic farms would be quite large.

The validity of this estimate is uncertain due to the failure to meet the parallel trends assumption as shown in Figure 4 and discussed in the previous section. Without evidence that the parallel trends assumption holds, the estimate of the effect of CRP land on organic farming from the two-way fixed effects model cannot be assumed to be causal. That is, it is possible that the observed trend in organic farming may have occurred in the absence of increased CRP land enrollment. In addition, some key assumptions used in the analysis make this measure less accurate. For instance, I estimate the share of CRP and organic land by dividing the amount of this land for a given state and year by the total amount of farmland as measured in the 2017 agriculture census. This approximation seems reasonable since the total amount of farmland does not vary greatly across years, but using the exact amount of total farmland for that given year and state would be a more accurate estimate.

In addition, it is important to note that this analysis cannot be generalized to draw conclusions about the effect of CRP grants on the likelihood of that farm to become organic certified since the analysis is done at the state level.

Conclusion

Using a difference-in-differences approach, this paper studies the effect of having a high share of CRP land grants on the acreage of organic farms at the state level. By comparing states with a low share of CRP land grants by acre to states with a high share across a time window of 1997, 2000-2008, 2011, I find that the effect of having a high share of CRP grants is correlated

with an increase of 0.0264 in organic acres. However, due to the large standard errors of this estimate, it is not possible to reject the null hypothesis that there is no effect.

Future analysis on this topic could be improved by using direct measures of soil health and reported use of regenerative farming practices, such as conservation tillage, crop rotations or limited fertilizer use as the response variable. Analysis using data on direct soil outcomes and regenerative practices rather than organic farming would give more representative information about the effect of CRP grants on land outcomes and farmers' behavior. Although organic farming is a good indicator for regenerative farming methods, there are many barriers to becoming officially certified as an organic farm and not all farmers who use regenerative practices are certified organic. Thus, it is likely that many farmers may transition to more regenerative practices after enrollment in the CRP without becoming organic certified, which is not accounted for in this paper.

There are many improvements that could be made to strengthen the assumptions used in this paper. It can take 2-3 years after expiration of CRP grants to become organic certified, and the data used in this paper only measures the quantity of land currently enrolled in the CRP, not the land with expiring contracts, so my analysis might miss a delayed effect of expiring contracts on organic farms. As such, my analysis does not directly measure whether specific farms that received CRP grants are converted to organic farms. Instead, this paper only studies the immediate association between the quantity of CRP grants and organic farms, assuming a negligible amount of new organic farms are being added or non-CRP farms are converted to organic.

The effect of CRP land grants on organic farming has important policy implications. Although the CRP has impressive immediate benefits for the soil health and environmental outcomes, the post-enrollment use of CRP land is uncertain. If there is evidence to suggest that CRP land grants influence farmers to adopt environmentally-friendly practices, this would provide additional reason to support the CRP.

Links to Data Sources:

- USDA Farm Service Agency (FSA): CRP Enrollment and Rental Payments by State, 1986-2019
  https://www.fsa.usda.gov/Assets/USDA-FSA-Public/usdafiles/Conservation/Excel/HistoryState86-19.xlsx
- USDA Economic Research Service: Table 2. U.S. certified organic farmland acreage, livestock numbers, and farm operations
  https://www.ers.usda.gov/webdocs/DataFiles/52407/Farmlandlivestockandfarm.xls?v=9862.4
- U.S. Census of Agriculture: Organic Agriculture Survey
  https://www.nass.usda.gov/Surveys/Guide_to_NASS_Surveys/Organic_Production/index.php

Works Cited:

Barnes JC, Sketch M, Gramza AR, Sorice MG, Iovanna R, Dayer AA (2020). *Land use decisions after the Conservation Reserve Program: Re-enrollment, reversion, and persistence in the southern Great Plains*. Conservation Science and Practice. https://doi.org/10.1111/ csp2.254

D.L. Gebhart, H.B. Johnson, H.S. Mayeux, H.W. Polley (1994) *The CRP increases soil organic carbon,* Journal of Soil and Water Conservation, 49 (5) 488-492 https://www.jswconline.org/content/49/5/488

Fleming, R.A., (2004), *An econometric analysis of the environmental benefits provided by the Conservation Reserve Program*: Journal of Agricultural and Applied Economics, v. 36, no. 2, p. 399–413. https://ageconsearch.umn.edu/record/43388

Organic Farming, (2021), US Environmental Protection Agency, https://www.epa.gov/agriculture/organic-farming

SARE Outreach, (2003), *What Is Organic Farming?* https://www.sare.org/publications/transitioning-to-organic-production/what-is-organic-farming/

Swan, Easter, Paustian (2018), *Quantifying changes in soil carbon and greenhouse gas emissions from adoption of CRP*, USDA Farm Service Agency https://www.fsa.usda.gov/Assets/USDA-FSA-Public/usdafiles/EPAS/natural-resouces-analysis/Multiple-Benefits/pdfs/CRP_Report_revised_7May2018.pdf

USDA Farm Service Agency, *About the Conservation Reserve Program* https://www.fsa.usda.gov/programs-and-services/conservation-programs/conservation-reserve-program/

North Dakota State University, (2016), *Bringing Land in the Conservation Reserve Program Back Into Crop Production or Grazing* https://www.ag.ndsu.edu/publications/crops/bringing-land-in-the-conservation-reserve-program-back-into-crop-production-or-grazing#section-17

Philip E Morefield et al, (2016), *Grasslands, wetlands, and agriculture: the fate of land expiring from the Conservation Reserve Program in the Midwestern United States.* Environ. Res. Lett. 11 094005 https://doi.org/10.1088/1748-9326/11/9/094005

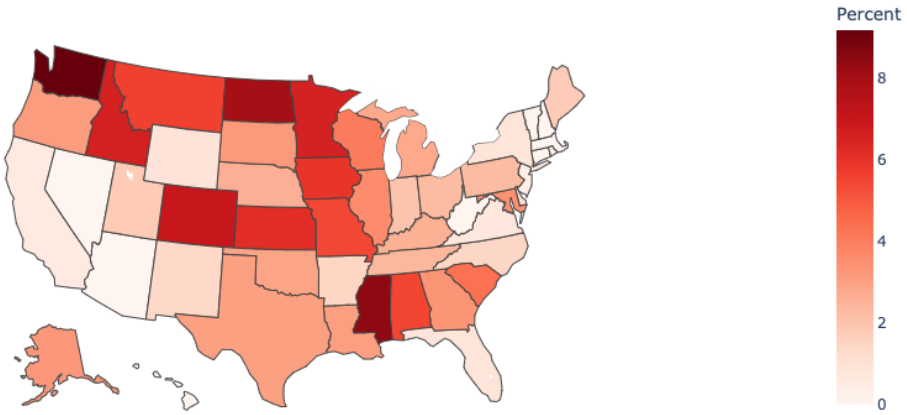
Appendix:

Table 1: Summary Statistics

	mean	med	std dev	min	max
Panel A: CRP land (% of total farmland)					
Total	2.86	2.49	2.49	0.00	9.18
Northeast	0.67	0.21	0.83	0.02	2.27
Midwest	4.40	3.83	1.96	2.02	8.00
South	2.92	2.95	2.10	0.08	8.46
West	3.04	1.80	3.10	0.00	9.18
Panel B: Organic Farms (% of total farmland)					
Total	1.62	0.32	7.93	0.01	56.32
Northeast	1.28	0.57	1.43	0.02	4.70
Midwest	0.31	0.24	0.25	0.08	0.97
South	0.07	0.03	0.09	0.01	0.29
West	4.86	0.44	15.47	0.07	56.32
N = 600					

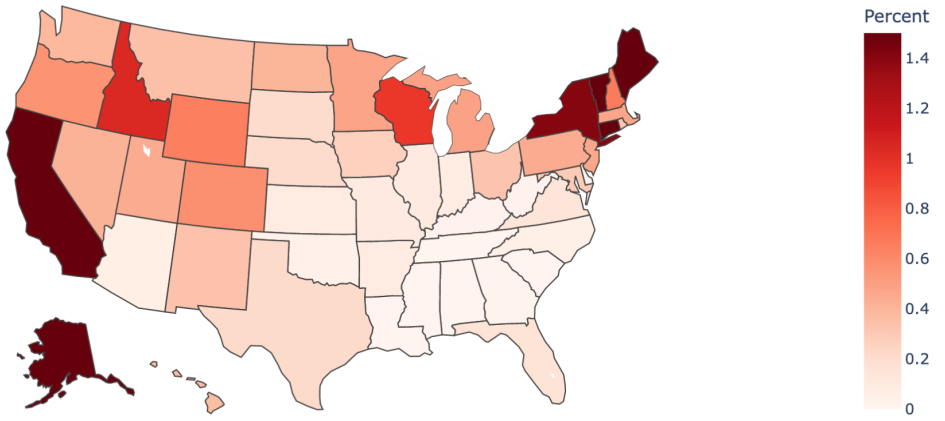
Notes:
Includes data from 1997, 2000-2008, 2010-11
Regional definitions:
Northeast: CT, ME, MA, NH, RI, VT, NJ, NY, DE, PA
Midwest: IL, IN, MI, OH, WI, NE, ND, SD
South: FL, GA, NC, SC, VA, WV, AL, KY, MS, TN, AR, LA, OK, TX
West: AZ, CA, ID, MT, NV, NM, UT, WY, AK, CO, HI, OR, WA

Figure 1: Average Percent of Farmland Awarded CRP Grants



Notes: State-level percentages calculated as average acres of land awarded CRP grants over study window (1997, 2000-2008, 2011) divided by the total acres of farmland in 2017.

Figure 2: Average Percent of Farmland as Organic Farms



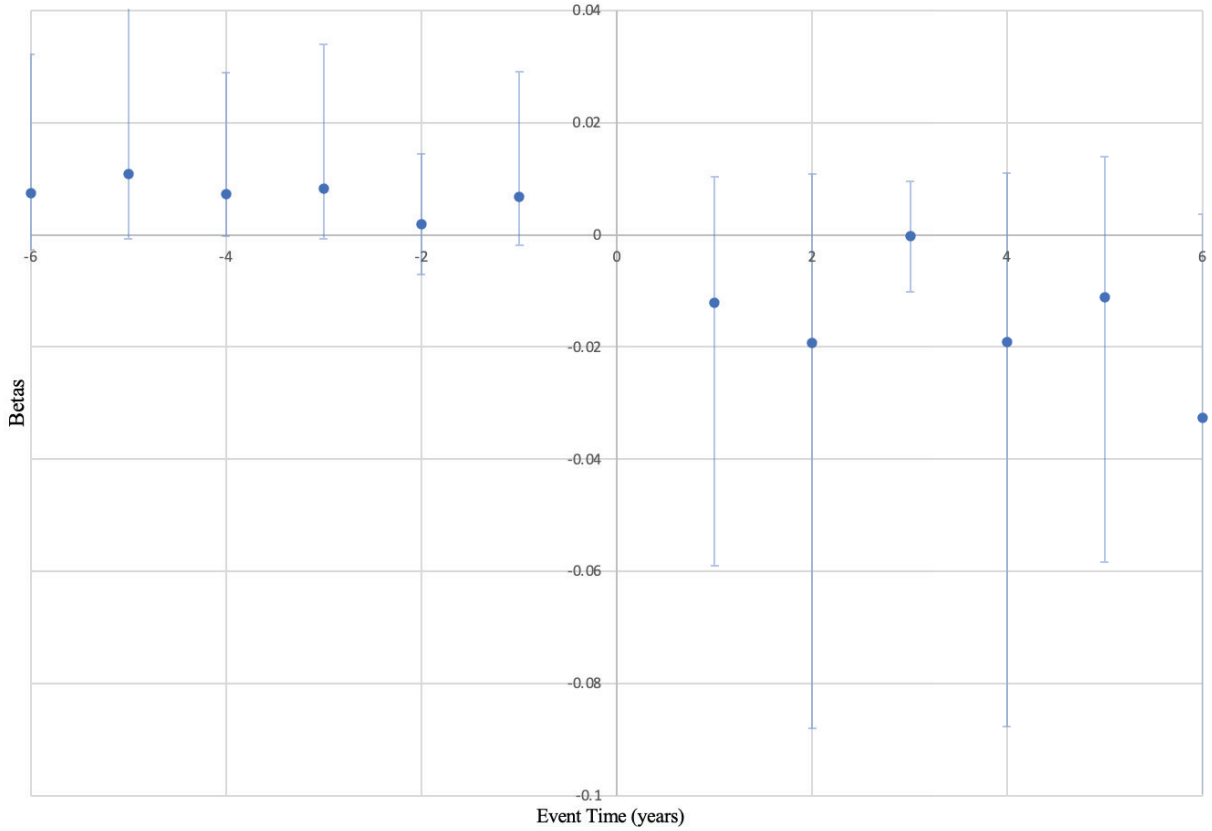
Notes: State-level percentages calculated as average acres of organic farmland over study window (1997, 2000-2008, 2011) divided by the total acres of farmland in 2017. Alaska has 56.32% of farmland as organic farms, not shown to scale in Figure 2

Table 2: Two-Way Fixed Effects Model

	(1)
High CRP Share	0.0264
	(0.0271)
State Fixed Effects	X
Year Fixed Effects	X
	N = 600
	R^2 = 0.379

Notes: State-level observations. Robust standard errors clustered at the state level shown in parentheses. “High CRP Share” is an indicator variable that equal to 1 when the share of farmland enrolled in the CRP is greater than

Figure 4: Event Study



Notes: 95% confidence interval shown in blue bars. Year of treatment normalized to x = 0.

The Impact of Medicaid Expansion on Risky Behavior in Low Income Individuals

Luke Stewart

MIT Undergraduate Economics Journal 2022

Abstract

I utilize variation in Medicaid eligibility requirements at the state level in differences-in-differences and triple differences frameworks to estimate the effect of Medicaid expansion on rates of high risk health behaviors such as smoking and drinking. Using micro-level survey data, I find null effects in the aggregate population. However, I find that young and middle-aged people had better behavioral outcomes and older people had worse behavioral outcomes, despite young people visiting their physicians less and older people visiting their physicians more.

Introduction

In the debate surrounding health insurance in the US and around the world, a hotly contested question is how health insurance affects the extent to which individuals behave in ways that will increase their demand for healthcare. Policy interventions that reduce the financial impact on individuals of risky health behaviors can lead to increased healthcare utilization and spending. This type of effect, which occurs because the negative outcome of a risky

behavior is insured against, removing some disincentive of the behavior, is known generally as moral hazard.

Due to this effect, it is possible that granting health insurance to individuals may drive them toward riskier health behaviors (or drive them away from quitting said behaviors), like smoking or binge drinking. However, increased primary care utilization may provide a countervailing effect. If individuals utilize more primary care, they may be more likely to reduce their risky health behaviors (or never begin them in the first place) at the advice of their primary care physician or other healthcare professional.

Therefore, the effect of health insurance on the tendency of individuals to engage in risky behavior is theoretically ambiguous, leaving it up to empirical study. This paper exploits the Medicaid eligibility requirement policy change from the Affordable Care Act and utilizes differences-in-differences (DD) and triple differences (DDD) methods to evaluate the policy effect on rates of smoking and binge drinking. In the general population, I find no significant effects on rates of smoking and binge drinking from the policy intervention. In that sense, I mostly agree with the existing literature. However, I also find that the effects are very heterogeneous across age. The rest of the paper is organized as follows: Section 1 provides a background of the relevant literature, Section 2 explains the data and methodologies used and discusses some assumptions of the models, Section 3 showcases results of the model estimation and checks modelling assumption, and Section 4 concludes.

1 Background

Over the last two decades, questions about how health insurance impacts outcomes such as healthcare utilization and high-risk health behaviors have been investigated many times. This literature dates back to the RAND Health Insurance Experiment of the 1980s. This experiment, reviewed in Aron-Dine et al. (2013), aimed to determine the impacts of health insurance on both healthcare spending and health outcomes utilizing random assignment

of health care plans. The most relevant result of the RAND study to the research here is the finding that overall healthcare spending increased with a decrease in out of pocket costs, indicating the downward sloping demand curve of healthcare markets. This is relevant to the question of moral hazard, because the increased healthcare spending associated with reduced costs can stem from both a downward sloping demand curve and also potential moral hazard.

In 2008, another randomized trial concerning health insurance occurred in Oregon. The Oregon Health Insurance Experiment, discussed in Finkelstein et al. (2012), Taubman et al. (2014), and many others, allowed certain low income individuals to lottery into the state's Medicaid program. Exploiting this random assignment, Finkelstein et al. (2012) find that after Medicaid enrollment, individuals have significantly higher healthcare utilization, lower out-of-pocket costs, and better health outcomes. Those enrolled have higher probabilities of hospital admission and emergency room utilization, which could be driven in part by a higher degree of risky behavior. Taubman et al. (2014) also finds an increase in the rate of emergency room utilization among the insured. While this increased emergency use could be driven by an increase in risky health behaviors, it is also possible that individuals did not undertake any additional risky behavior, and instead went to the hospital when they otherwise would not have used any healthcare had they not been insured.

Other studies have used identification strategies that involve policy shocks as opposed to random assignment. For instance, Dave & Kaestner (2009) uses the exogenous assignment of Medicare to individuals aging into the program to examine the causal effect of health insurance on behaviors and finds that obtaining health insurance reduces preventative behaviors (e.g. a healthy diet and regular exercise) and increases unhealthy behaviors among men over 65, after controlling for contact with medical professionals.

Many studies have utilized the same exogenous variation that I plan to use to identify effects on behavior caused by an increase in health insurance coverage. Dave et al. (2019) use vital statistics and employ a DD strategy based on Medicaid expansion to examine the impacts of health care on prenatal care behaviors. They find that Medicaid expansions led

to an increase in prenatal smoking and weight gain for low-educated mothers. Cotti et al. (2019) use data on consumer purchases in a DDD strategy following Medicaid expansion to find that the introduction of healthcare had large negative effects on the purchase of tobacco products and large positive effects on the purchase of smoking cessation products.

Courtemanche et al. (2018) and Simon et al. (2017) both utilize the BRFSS state-level data and Medicaid expansion to attempt to measure moral hazard. The former uses a DDD strategy and the latter utilizes DD. Neither study finds significant effects on behaviors, despite some evidence of increased care utilization. One potential drawback of each of these studies is the relatively short period of post-treatment data. If effects are initially small and grow over a number of years, neither study would identify those effects.

2 Data and Methodology

BRFSS Data

For this project, I will primarily take advantage of individual-level survey data from the Behavioral Risk Factor Surveillance System. This is a survey that captures data from across states, conducted yearly via telephone by the CDC (*BRFSS* 2021).

This survey collects data on a wide variety of topics, including demographic info, self-reported health conditions, healthcare utilization, tobacco and alcohol use, and diet. The survey has been carried out in some form since 1984, and was made nationwide in the 1990s. In most recent years, the sample sizes of BRFSS are over 400,000 individuals across the 50 states and DC.

One potential issue arising from using telephone survey data dating back decades is that the use of landlines and cell phones have evolved over time, and survey formats must evolve to respond to such changes. Prior to 2011, the BRFSS used only landlines to conduct the survey, making the data systematically different from 2011 onwards. This could cause biases in basic DD frameworks, but when using survey microdata, we can exploit demographic

variables to control for any composition effects this would have on our estimates.

The primary fields I take from these data are demographic controls and self-reported behavioral variables focused on tobacco and alcohol use. I choose these variables because others, such as those related to diet, can take years to develop and change for a given individual following a policy shock. We would not expect them to be responsive in the years immediately following a major change in rates of insurance or Medicaid participation.

For information on Medicaid eligibility expansion, I use data distributed by the Kaiser Family Foundation, which includes the date of Medicaid expansion in a given state (KFF 2021).

Table 1 displays summary statistics for quantitative variables in my primary estimation sample. For most of my models, I restrict the sample to individuals between 18 and 64 years old, who have household income below $20,000 per year. Because income is a grouped variable, it is not possible to adjust properly for inflation, so the $20,000 ceiling is effectively higher (in real terms) in the early years of the survey compared to the most recent years. However, this is still a relatively low household income throughout the sample period, so biases from this should be second order. Figures 1 and 2 summarize demographic trends in my data over time for the states in which Medicaid was not expanded (the control group) and the states in which Medicaid was expanded (the treatment group). Figure 3 provides a first look at rates of smoking and drinking (our outcomes of interest) over time in the treatment group and donor pool.

An individual is counted as smoking in this sample if they either "smoke every day" or "smoke some days" according to their responses to a number of smoking-related questions.[1] The relevant outcome variable for drinking behavior is the number of occasions in the month prior to the interview that the interviewee reported having 5 or more alcoholic beverages in a single sitting.

It should be noted that over time, the sample's average age increases slightly, which

[1]The dummy variable for smoking is calculated by the CDC based on the interviewee's responses to other questions. See here for a list of calculated variables in the 2020 survey.

Table 1: Summary statistics of imputed quantitative variables, by treatment group

(a) **Panel A: Control group**

	Mean	SD	Min	Max	*N*
1993					
General Health	2.63	1.13	1	5	4,373
Imputed age	38.11	13.48	21	62	4,383
Imputed income	11,471.59	5,239.49	5,000	17,500	4,383
2000					
General Health	2.89	1.18	1	5	4,982
Imputed age	39.99	13.68	21	62	4,992
Imputed income	12,685.30	5,127.19	5,000	17,500	4,992
2010					
General Health	3.34	1.17	1	5	10,613
Imputed age	48.46	11.78	21	62	10,659
Imputed income	11,768.22	5,241.33	5,000	17,500	10,659
2020					
General Health	3.08	1.18	1	5	5,811
Imputed age	45.09	13.80	21	62	5,835
Imputed income	12,241.65	5,276.33	5,000	17,500	5,835
Total					
General Health	3.07	1.20	1	5	25,779
Imputed age	44.31	13.58	21	62	25,869
Imputed income	12,001.72	5,243.38	5,000	17,500	25,869

could be related to non-response bias in younger individuals over time. This is a relevant factor in our full sample analysis, but not one that I can address fully due to data limitations. However, this does motivate the use for age controls in any regression run.

It is important to note here that although it would be optimal for our sample to only include those that are Medicaid eligible, it is typically very difficult to determine Medicaid eligibility from survey responses. To compound this difficulty, key variables like income are encoded in the survey in groups (i.e. group 1 corresponding to income below $10,000, etc.). Therefore, to the extent this study will identify effects at all, the aggregate effects will be on a population consisting of unaffected individuals as well as those directly affected by Medicaid expansion.

Figure 4 attempts to quantify the discrepancy between how my sample restrictions

(b) **Panel B: Treatment group**

	Mean	SD	Min	Max	*N*
1993					
General Health	2.57	1.12	1	5	14,153
Imputed age	38.08	13.54	21	62	14,188
Imputed income	11,414.05	5,271.24	5,000	17,500	14,188
2000					
General Health	2.87	1.19	1	5	13,744
Imputed age	40.24	13.53	21	62	13,773
Imputed income	12,355.70	5,212.90	5,000	17,500	13,773
2010					
General Health	3.26	1.18	1	5	23,314
Imputed age	48.40	11.82	21	62	23,403
Imputed income	11,880.63	5,251.90	5,000	17,500	23,403
2020					
General Health	2.98	1.19	1	5	17,285
Imputed age	44.86	14.04	21	62	17,341
Imputed income	12,241.36	5,277.97	5,000	17,500	17,341
Total					
General Health	2.97	1.20	1	5	68,496
Imputed age	43.74	13.73	21	62	68,705
Imputed income	11,970.56	5,265.52	5,000	17,500	68,705

predict Medicaid enrollment and actual Medicaid enrollment, by comparing the proportion of data kept after filtering my sample to the proportion of individuals in the US covered by Medicaid.[2] This figure reveals how potentially worrisome the lack of direct Medicaid coverage representation in the data really is. Generally, as time goes on, my sample restrictions predict fewer people being placed on Medicaid, whereas the actual trend is increasing. I propose that this is likely due to how income is treated in my data and methodology. Because I don't account for inflation or generally increasing incomes, there are individuals in my full sample that are not included after filtering that are within the income requirements of the Medicaid expansion. As a result, the number of individuals remaining after filtering is an underestimate. Therefore, the interpretation of any results we find should be limited to the sample itself, and any extrapolation beyond that should be taken with a grain of salt.

[2]This data has been pulled from Statista, originally released by the US Census (US Census Bureau & Yang 2021)

(c) **Panel C: Full sample**

	Mean	SD	Min	Max	N
1993					
General Health	2.59	1.12	1	5	18,526
Imputed age	38.09	13.53	21	62	18,571
Imputed income	11,427.63	5,263.68	5,000	17,500	18,571
2000					
General Health	2.88	1.19	1	5	18,726
Imputed age	40.18	13.57	21	62	18,765
Imputed income	12,443.38	5,192.14	5,000	17,500	18,765
2010					
General Health	3.29	1.18	1	5	33,927
Imputed age	48.41	11.81	21	62	34,062
Imputed income	11,845.46	5,248.78	5,000	17,500	34,062
2020					
General Health	3.00	1.19	1	5	23,096
Imputed age	44.92	13.98	21	62	23,176
Imputed income	12,241.44	5,277.45	5,000	17,500	23,176
Total					
General Health	3.00	1.20	1	5	94,275
Imputed age	43.89	13.69	21	62	94,574
Imputed income	11,979.09	5,259.46	5,000	17,500	94,574

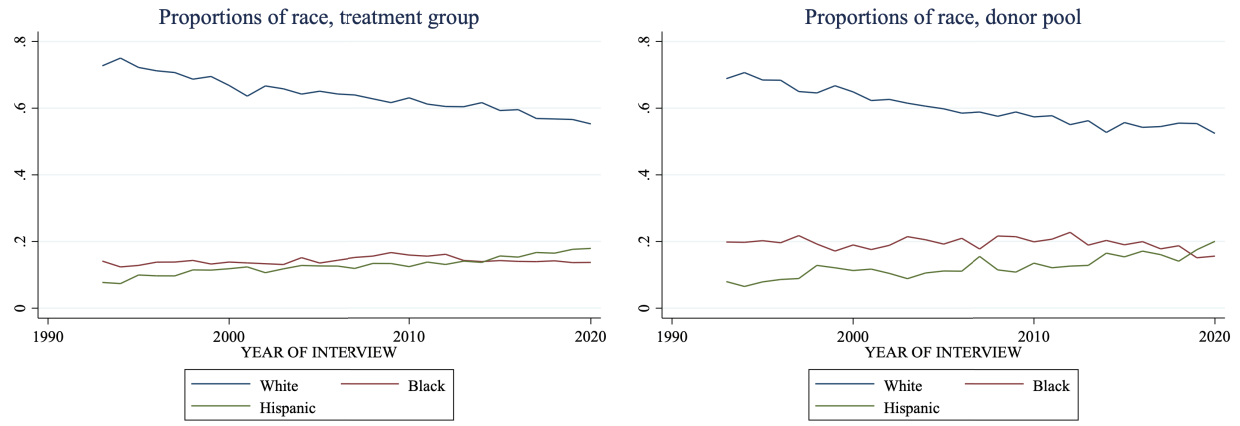
Data gathered from the BRFSS survey at the individual level from 1993 to 2020.
General health given as an index from 1 to 5 with 1 being the best.
Age and income were imputed from grouped variables by taking the
simple average of the bounds of each group.

Differences-in-differences and triple differences

To identify aggregate state-level effects of Medicaid expansion on rates of certain risky behaviors, I will use a handful of different methods. I will use DD estimates as my baseline with which I will compare DDD estimates. To do this, I will estimate a regression of the following form

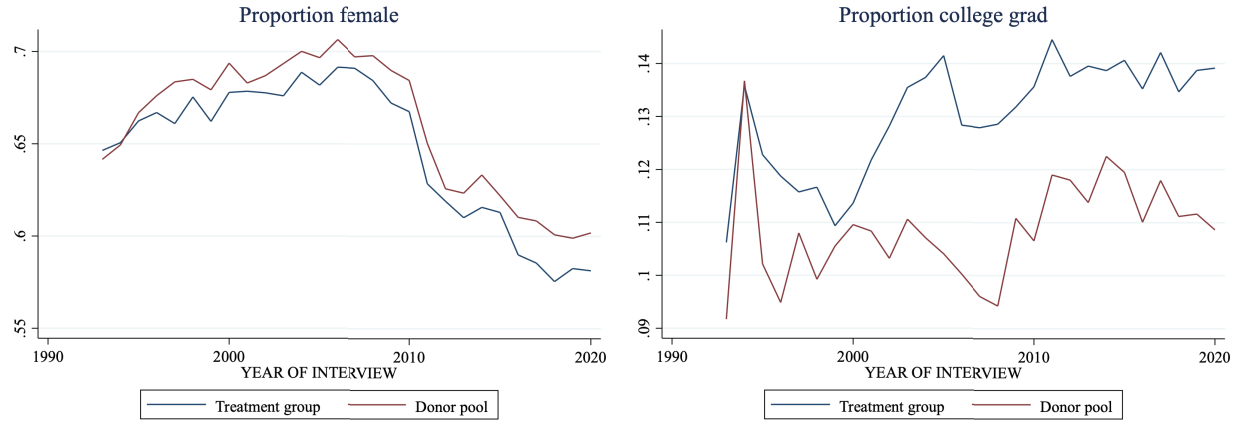
$$Y_{it} = \gamma_i + \delta_t + \beta TREAT_{it} + X_{it} + \varepsilon_{it} \tag{1}$$

where γ_i and δ_t are state and year fixed effects respectively, X_{it} are demographic controls such as race and education, $TREAT_{it}$ is a dummy for being in a state that has expanded



(a) (b)

Figure 1: Proportion of individuals in sample of different races

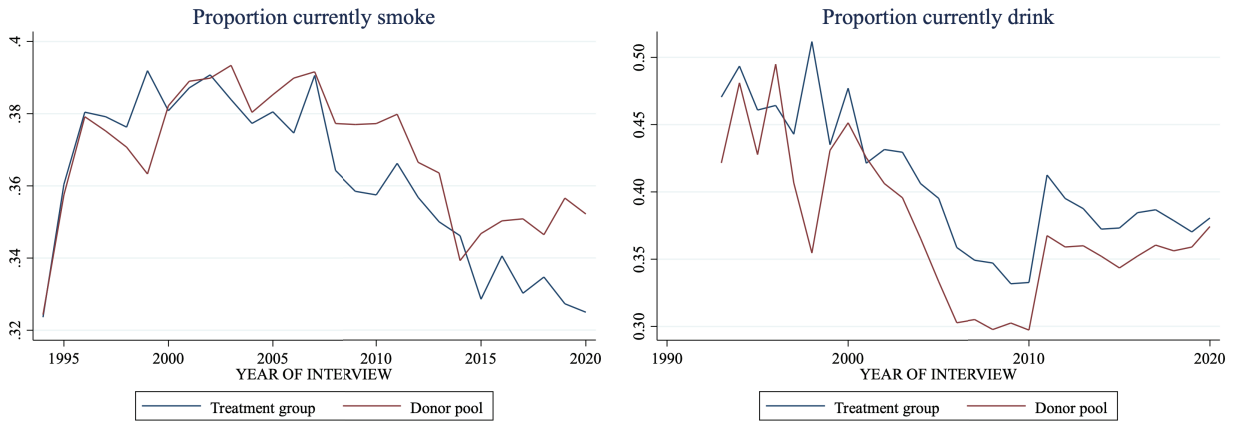


(a) Proportion of female respondents in sample (b) Proportion of college graduates in sample

Figure 2

Medicaid eligibility in year t, and ε_{it} is an error term. Under the parallel trends assumption, β is identified as a causal effect on the population in aggregate. Note that we're studying the impact of state-level policy shocks, but the individual-level survey data allow us to take advantage of demographic controls to improve precision.

In my DDD analysis, I will compare the difference in outcomes for higher income and lower income individuals pre- and post-policy change, by state. This methodology reflects the difference in treatment intensity that should occur between higher and lower income individuals in this context, due to the expanded Medicaid eligibility requirements applying



(a) (b)

Figure 3: Proportion of individuals engaging in smoking and drinking behavior

to low income individuals only. To do this, I will estimate a regression of the form

$$Y_{it} = \gamma_i + \delta_t + \rho L_{it} + \beta L_{it} * TREAT_{it} + \mu TREAT_{it} + \varepsilon_{it} \tag{2}$$

where L_{it} is a dummy variable for being low income and the other variables are defined as above. In this case, β captures the relative causal effect of being treated on low income populations, as compared to high income populations. This model attempts to control for the possibility that there might be other policies or programs implemented in treated states throughout the relevant period that affect the whole population, not just low income individuals. If no such programs or policies existed, and it was as if high income individuals received no treatment throughout the relevant period, we would expect $\mu = 0$. In these models only low income groups (with income below \$20,000 per year) and high income groups (with income above \$50,000 per year) are included.

In an extension of the basic model, I also estimate treatment effects for different age groups. I define three age groups: young people (age 18-25), middle-aged people (26-50), and older individuals who are not yet eligible to receive Medicare benefits (51-64). I then

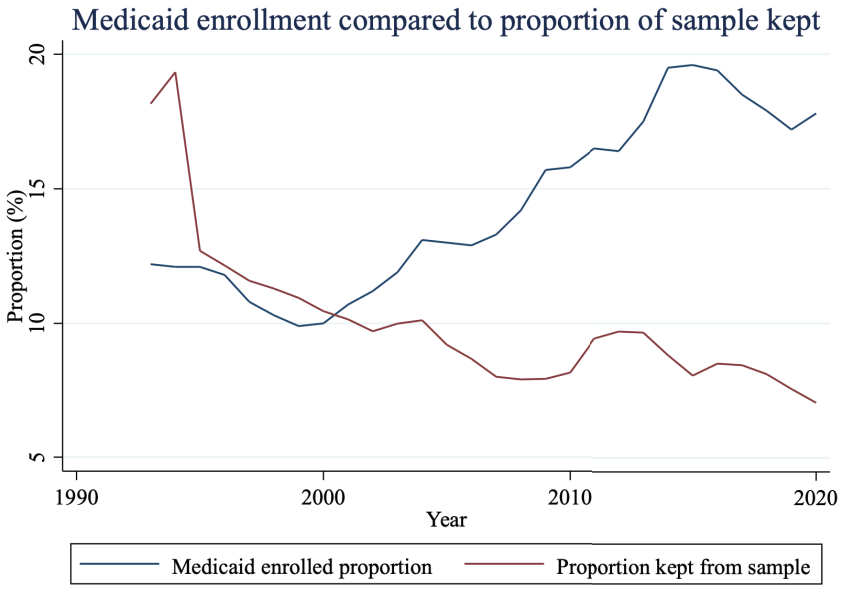


Figure 4: Examining sample filter accuracy

estimate a model of the following form (in the DD case)

$$Y_{it} = \gamma_i + \delta_t + \nu_1 B_{it} + \nu_2 C_{it} + \beta_0 TREAT_{it} + \beta_1 TREAT_{it} B_i + \beta_2 TREAT_{it} C_{it} + X_{it} + \varepsilon_{it} \tag{3}$$

where B_{it} is a dummy indicating that the individual is aged 18 to 25 and C_{it} is a dummy indicating that the individual is aged 26 to 50. Therefore, the causal effect of treatment for individuals in the 18-25 age range is $\beta_0 + \beta_1$, for individuals in the 26-50 age range is $\beta_0 + \beta_2$, and for individuals in the 51-64 age range is β_0. I run a similar model under a DDD specification to examine relative treatment effects between low and high income, young individuals, low and high income middle-aged individuals, and low and high income older individuals.

3 Results

Primary estimation

I estimate different specifications of Equations (1) and (2) to produce DD and DDD estimates of the effect of Medicaid expansion on smoking and binge drinking rates. Table 2 contains the DD and DDD estimates from two models: a model without controls and a model with basic demographic controls (including age, sex, race, education, and employment status). Models with expanded controls (including medical outcomes such as a general health index and vaccination status) were considered, but were not estimated to reduce risk of bias stemming from controlling on outcomes. Models 1 and 3 are basic DD/DDD models, and models 2 and 4 add the controls. I find no statistically significant effects with these model specifications in the full sample on either smoking or binge drinking for low income individuals.

For smoking, the treatment effect on high income individuals in the DDD model is very near 0 and precisely estimated, validating our DD estimates. For drinking, we see a moderate and marginally significant decrease in rate of binge drinking occurrences among high income individuals, with no significant relative effect between low and high income individuals. This suggests that their may have been other factors affecting binge drinking rates in the treated states in this time period (e.g. policies against drunk driving or changing liquor laws).

Parallel trends

The DD and DDD estimates obtained above are only valid under the assumptions described previously, namely parallel trends. To some extent, we can verify the parallel trends assumption graphically. To do so, we estimate

$$\tilde{Y}_{it} = \gamma_i + X_{it} + \nu_{it} \tag{4}$$

Table 2: DD/DDD regression estimates

(a) **Panel A: Smoking rate**

	(1) DD	(2) DD w/ controls	(3) DDD	(4) DDD w/ controls
Treatment indicator	-0.007	-0.007	-0.003	-0.001
	(0.006)	(0.004)	(0.003)	(0.003)
Low Income			0.227	0.119
			(0.009)	(0.008)
Treated*Low Income			-0.001	-0.002
			(0.010)	(0.008)
Female		-0.044		-0.022
		(0.005)		(0.002)
Black		-0.106		-0.067
		(0.012)		(0.009)
Hispanic		-0.206		-0.109
		(0.012)		(0.007)
Other Race		-0.017		0.010
		(0.013)		(0.007)
HS Grad.		-0.080		-0.080
		(0.007)		(0.007)
Col. Grad.		-0.220		-0.196
		(0.010)		(0.009)
25 - 50 y.o.		0.041		0.018
		(0.006)		(0.006)
50+ y.o.		-0.048		-0.021
		(0.009)		(0.006)
Other controls	No	Yes	No	Yes
Observations	753408	747686	2834176	2821578

where γ_i are state fixed effects and X_{it} is a vector of demographic controls. We take ν_{it}, the residuals from this regression, and plot it against time, separately for treatment and control states. For this to be well defined as a time series, we must restrict our analysis to units that were treated at the same time, so I limit my analysis here to those states that expanded Medicaid beginning January 1, 2014 and assume that this extends to the full sample.[3] Because state fixed effects were included, we should see two very similar series, practically moving together, if parallel trends is to be satisfied.

Figure 6 shows the plots of the ν_{it} series for both the smoking rate and binge drinking

[3]Most of the states that expanded Medicaid did so on January 1, 2014, so this assumption is reasonable.

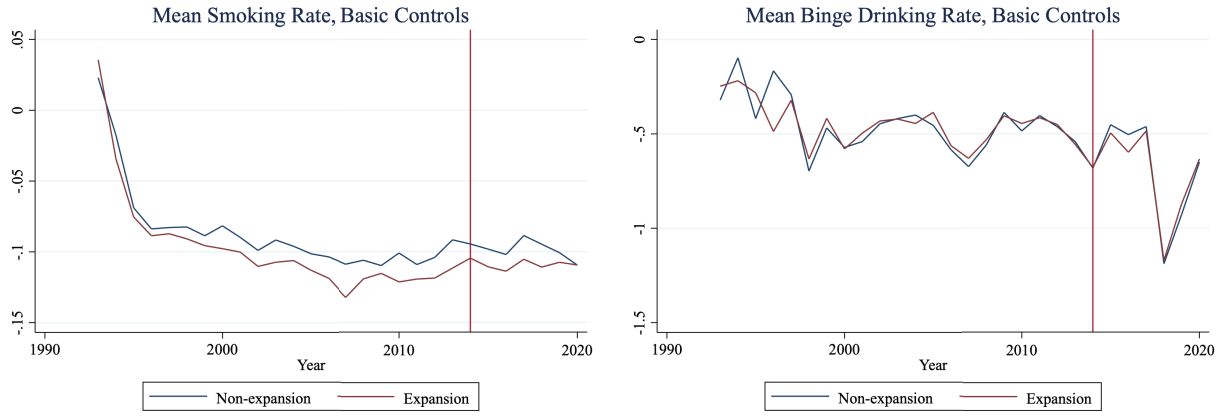
(b) **Panel B: # of binge drinking occasions**

	(1) DD	(2) DD w/ controls	(3) DDD	(4) DDD w/ controls
Treatment indicator	-0.035	-0.049	-0.062	-0.046
	(0.058)	(0.056)	(0.025)	(0.023)
Low Income			0.882	0.393
			(0.030)	(0.031)
Treated*Low Income			0.037	0.013
			(0.044)	(0.043)
Female		-1.569		-0.899
		(0.061)		(0.033)
Black		-0.599		-0.511
		(0.056)		(0.034)
Hispanic		-0.696		-0.354
		(0.087)		(0.046)
Other Race		-0.009		-0.040
		(0.069)		(0.041)
HS Grad.		-0.501		-0.695
		(0.054)		(0.048)
Col. Grad.		-1.073		-1.250
		(0.059)		(0.056)
25 - 50 y.o.		-0.012		-0.217
		(0.043)		(0.036)
50+ y.o.		-0.195		-0.555
		(0.056)		(0.034)
Other controls	No	Yes	No	Yes
Observations	257996	256283	1597152	1591482

Notes: Standard errors shown in parentheses. All models are differences-in-differences or triple differences, with state and year fixed effects, estimated using OLS. In the DD models, the variable of interest is a dummy variable equal to one in Medicaid expansion states post-expansion. In the DDD models, the variable of interest is a dummy variable equal to one for low-income individuals in Medicaid expansion states post-expansion. DDD models capture the relative effect of Medicaid expansion on low income populations. The other controls include dummy variables for marital status, employment, and income (broken down into subcategories) variables. Standard errors are clustered at the state and year level.

rate outcome, separate by control and treatment units. We see that for smoking rates, and to an even greater extent for binge drinking rates, the series overlap considerably pre-treatment. This gives reassurance that parallel trends is plausibly satisfied after controlling for demographic variables.

Figure 5: Checking parallel trends



Estimating by age

Because most individuals begin smoking at a young age[4], it might be informative to examine the effect on smoking rates of Medicaid expansion on a subsample of young individuals. In fact, there may be other heterogeneous effects by age for both smoking and drinking.

Table 3 shows results similar to those in Table 2, but now using Equation (3) as our model of choice. In Panel A, effects on smoking rates are shown. Under the DD model, Young individuals see a significant reduction in smoking rates of 6.8% (SE = 1.3%) when controlling for demographic factors.[5] Middle-aged individuals see a marginally significant reduction of smoking rates of 1.9% (SE = 0.87%) when using controls. However, older individuals see a significant increase in smoking rates of 1.9% (SE = 0.6%) under the same treatment.

Using a DDD model, both young and middle aged low income people see significant negative effects on smoking rates relative to their high income counterparts as suggested by the DD model. Low income older individuals continue to see significant positive effects on smoking rates relative to their high income peers.

We see a similar story in the binge drinking results, shown in Panel B of Table 3. Young

[4]See here for more info.

[5]This is calculated by means of an F-test on the sum of the coefficient on the treatment indicator and the treatment indicator*young interaction.

Table 3: DD/DDD regression estimates - age interacted

(a) **Panel A: Smoking rate**

	(1) DD	(2) DD w/ controls	(3) DDD	(4) DDD w/ controls
Treatment indicator	0.024	0.019	-0.004	-0.001
	(0.007)	(0.006)	(0.003)	(0.003)
Less than 25 y.o.	-0.026	0.063	0.034	0.035
	(0.010)	(0.009)	(0.004)	(0.004)
25 - 50 y.o.	0.060	0.096	0.014	0.027
	(0.009)	(0.006)	(0.002)	(0.002)
Treated*Young	-0.098	-0.086	-0.020	-0.037
	(0.018)	(0.016)	(0.006)	(0.008)
Treated*Middle-aged	-0.044	-0.033	0.005	0.005
	(0.009)	(0.007)	(0.004)	(0.003)
Low Income			0.224	0.093
			(0.007)	(0.006)
Low Income*Young			-0.069	0.007
			(0.008)	(0.007)
Low Income*Middle-aged			0.039	0.056
			(0.009)	(0.008)
Treated*Low Income			0.030	0.026
			(0.009)	(0.008)
Treatment*Low Income*Young			-0.072	-0.046
			(0.014)	(0.010)
Treatment*Low Income*Middle-aged			-0.045	-0.040
			(0.012)	(0.011)
Female		-0.044		-0.022
		(0.005)		(0.002)
Black		-0.106		-0.067
		(0.011)		(0.009)
Hispanic		-0.206		-0.111
		(0.012)		(0.008)
Other Race		-0.016		0.010
		(0.013)		(0.007)
HS Grad.		-0.080		-0.080
		(0.007)		(0.008)
Col. Grad.		-0.221		-0.196
		(0.010)		(0.009)
Other controls	No	Yes	No	Yes
Observations	753408	747686	2834176	2821578

people experienced a decrease in reported monthly binge drinking occurrences of 0.275 (SE = 0.091), which equates to roughly 3.3 fewer occurrences per year. Middle aged individuals and older individuals saw no significant effect on binge drinking occurrences in a DD model.

In a DDD model, we see similar relative effects for young people as we did in the DD model. For middle aged individuals, we see a significant relative effect of 0.290 (SE = 0.065) fewer occurrences per month, which equates to roughly 3.5 fewer occurrences per year. As in the case of smoking, low income older individuals saw a small but significant increase in number binge drinking occurrences relative to their high income peers.

Frequency of yearly check-ups

In short, it seems that younger and middle-aged individuals tended to have better behavioral health outcomes after Medicaid expansion than older individuals. One possible explanation for this is that younger individuals had more exposure to their primary care physician or other healthcare professionals, thus receiving more information about the hazards of risky behaviors.

I use another survey question from the BRFSS that asks if the interviewee has received an annual check-up from their doctor in the 12 months prior to their interview.

Table 4 shows results of DD and DDD models that are essentially identical as those in Table 3, just using the indicator variable of having received a yearly check-up in the 12 months prior to the interview as an outcome variable.

Did young people do better because they saw their physicians more often? No! In the DD model, young people saw no significant changes in the proportion of those who had received a yearly check-up. In the DDD model, low income young individuals saw a significantly more negative treatment effect than their high income counterparts.

In the DD specification, low income middle-aged individuals were 4.6% (SE = 1.1%) more likely to have received a check-up, but this effect disappears relative to their high-income counterparts in the DDD specification.

Older individuals were the only group more likely to have received a check-up in both model specifications.

4 Conclusion

In this study, my full sample results mostly agree with those found in Courtemanche et al. (2018) and Simon et al. (2017). For the most part, in this sample, there are no significant effects of Medicaid expansion on high risk behavior for low income populations. Therefore, in aggregate, the moral hazard effect does not outweigh other effects to a concerning degree.

The more intriguing results are those that are separate by age. We see young and middle aged individuals had significantly better behavioral health outcomes after Medicaid expansion, indicating that the moral hazard effect does not dominate. Furthermore, I find that older individuals had significantly worse behavioral health outcomes after Medicaid expansion, which is in line with the results of Dave & Kaestner (2009), discussed above.

However, it is not clear that exposure to healthcare professionals is responsible for the better behaviors in the young and middle-aged cohorts, or that lack of exposure is responsible for the worse behaviors of older cohorts.

It is possible that the proportion of individuals receiving yearly check-ups does not sufficiently account for exposure to healthcare professionals. After all, yearly check-ups are relatively routine, and may represent "low-impact" exposure for many individuals. Future work should focus on properly identifying differences in both quantity and intensity of healthcare exposure for different age groups to explain these findings.

Based on the results in Table 2 (the full sample DD/DDD analysis), we can rule out ranges of results. In this policy context, where a potential positive effect on smoking rates would counteract the purpose of coverage-expanding policies, a *lack* of a large positive effect can be just as important as a large negative effect. The confidence interval for our results in all models rules out effects as large as 1 percentage point on smoking rates, and 1.5 binge

drinking occasions per month. These provides a cap on the potential net moral hazard effect of Medicaid expansion.

(b) **Panel B: # of binge drinking occasions**

	(1) DD	(2) DD w/ controls	(3) DDD	(4) DDD w/ controls
Treatment indicator	0.107	0.091	-0.125	-0.076
	(0.076)	(0.071)	(0.030)	(0.028)
Less than 25 y.o.	0.076	0.266	0.866	0.731
	(0.065)	(0.056)	(0.049)	(0.037)
25 - 50 y.o.	0.048	0.219	0.284	0.343
	(0.043)	(0.037)	(0.021)	(0.021)
Treated*Young	-0.336	-0.366	0.117	-0.133
	(0.117)	(0.106)	(0.060)	(0.052)
Treated*Middle-aged	-0.187	-0.161	0.129	0.107
	(0.063)	(0.061)	(0.028)	(0.023)
Low Income			1.071	0.624
			(0.044)	(0.044)
Low Income*Young			-0.803	-0.576
			(0.066)	(0.049)
Low Income*Middle-aged			-0.248	-0.225
			(0.043)	(0.042)
Treated*Low Income			0.234	0.173
			(0.075)	(0.070)
Treatment*Low Income*Young			-0.443	-0.212
			(0.118)	(0.106)
Treatment*Low Income*Middle-aged			-0.311	-0.290
			(0.070)	(0.065)
Female		-1.570		-0.894
		(0.062)		(0.034)
Black		-0.600		-0.513
		(0.057)		(0.034)
Hispanic		-0.693		-0.346
		(0.086)		(0.046)
Other Race		-0.005		-0.041
		(0.069)		(0.041)
HS Grad.		-0.500		-0.689
		(0.054)		(0.049)
Col. Grad.		-1.073		-1.244
		(0.059)		(0.057)
Other controls	No	Yes	No	Yes
Observations	257996	256283	1597152	1591482

Notes: Standard errors shown in parentheses. All models are differences-in-differences or triple differences, with state and year fixed effects. In the DD models, the variable of interest is a dummy variable equal to one in Medicaid expansion states post-expansion. In the DDD models, the variable of interest is a dummy variable equal to one for low-income individuals in Medicaid expansion states post-expansion. DDD models capture the relative effect of Medicaid expansion on low income populations. The other controls include dummy variables for marital status, employment, and income (broken down into subcategories) variables. Standard errors are clustered at the state and year level.

Table 4: DD/DDD regression estimates on medical checkups

	(1) DD	(2) DD w/ controls	(3) DDD	(4) DDD w/ controls
Treatment indicator	0.036	0.036	-0.014	-0.016
	(0.009)	(0.009)	(0.007)	(0.007)
Less than 25 y.o.	-0.121	-0.055	-0.150	-0.141
	(0.012)	(0.008)	(0.007)	(0.006)
25 - 50 y.o.	-0.112	-0.071	-0.114	-0.108
	(0.006)	(0.004)	(0.002)	(0.002)
Treated*Young	-0.055	-0.042	-0.010	0.005
	(0.012)	(0.009)	(0.008)	(0.005)
Treated*Middle-aged	0.007	0.010	0.009	0.012
	(0.007)	(0.007)	(0.003)	(0.003)
Low Income			-0.080	-0.149
			(0.007)	(0.006)
Low Income*Young			0.035	0.082
			(0.008)	(0.007)
Low Income*Middle-aged			0.006	0.032
			(0.006)	(0.005)
Treated*Low Income			0.076	0.076
			(0.009)	(0.008)
Treatment*Low Income*Young			-0.051	-0.056
			(0.010)	(0.010)
Treatment*Low Income*Middle-aged			-0.006	-0.006
			(0.009)	(0.009)
Female		0.102		0.107
		(0.006)		(0.007)
Black		0.139		0.124
		(0.005)		(0.005)
Hispanic		0.041		0.033
		(0.007)		(0.005)
Other Race		0.038		0.024
		(0.006)		(0.003)
HS Grad.		0.014		0.019
		(0.003)		(0.003)
Col. Grad.		-0.003		0.026
		(0.005)		(0.004)
Other controls	No	Yes	No	Yes
Observations	653540	648370	2501826	2490376

Notes: Standard errors shown in parentheses. All models are differences-in-differences or triple differences, with state and year fixed effects. In the DD models, the variable of interest is a dummy variable equal to one in Medicaid expansion states post-expansion. In the DDD models, the variable of interest is a dummy variable equal to one for low-income individuals in Medicaid expansion states post-expansion. DDD models capture the relative effect of Medicaid expansion on low income populations. The other controls include dummy variables for marital status, employment, and income (broken down into subcategories) variables. Standard errors are clustered at the state and year level.

References

Aron-Dine, A., Einav, L. & Finkelstein, A. (2013), 'The RAND Health Insurance Experiment, Three Decades Later', *Journal of Economic Perspectives* **27**(1), 197–222.
URL: *https://pubs.aeaweb.org/doi/10.1257/jep.27.1.197*

BRFSS (2021).
URL: *https://www.cdc.gov/brfss/index.html*

Cotti, C., Nesson, E. & Tefft, N. (2019), 'Impacts of the ACA Medicaid expansion on health behaviors: Evidence from household panel data', *Health Economics* **28**(2), 219–244.
URL: *https://onlinelibrary.wiley.com/doi/10.1002/hec.3838*

Courtemanche, C., Marton, J., Ukert, B., Yelowitz, A. & Zapata, D. (2018), 'Early Effects of the Affordable Care Act on Health Care Access, Risky Health Behaviors, and Self-Assessed Health: Early Effects of the Affordable Care Act', *Southern Economic Journal* **84**(3), 660–691.
URL: *https://onlinelibrary.wiley.com/doi/10.1002/soej.12245*

Dave, D. & Kaestner, R. (2009), 'Health insurance and ex ante moral hazard: evidence from Medicare', *International Journal of Health Care Finance and Economics* **9**(4), 367–390.
URL: *http://link.springer.com/10.1007/s10754-009-9056-4*

Dave, D. M., Kaestner, R. & Wehby, G. L. (2019), 'Does public insurance coverage for pregnant women affect prenatal health behaviors?', *Journal of Population Economics* **32**(2), 419–453.
URL: *http://link.springer.com/10.1007/s00148-018-0714-z*

Finkelstein, A., Taubman, S., Wright, B., Bernstein, M., Gruber, J., Newhouse, J. P., Allen, H., Baicker, K. & Oregon Health Study Group (2012), 'The Oregon Health Insurance Experiment: Evidence from the First Year*', *The Quarterly Journal of Economics* **127**(3), 1057–1106.
URL: *https://academic.oup.com/qje/article/127/3/1057/1923446*

KFF (2021), 'Status of State Medicaid Expansion Decisions: Interactive Map'.
URL: *https://www.kff.org/medicaid/issue-brief/status-of-state-medicaid-expansion-decisions-interactive-map/*

Simon, K., Soni, A. & Cawley, J. (2017), 'The Impact of Health Insurance on Preventive Care and Health Behaviors: Evidence from the First Two Years of the ACA Medicaid Expansions: Impact of Health Insurance on Preventive Care and Health Behaviors', *Journal of Policy Analysis and Management* **36**(2), 390–417.
URL: *https://onlinelibrary.wiley.com/doi/10.1002/pam.21972*

Taubman, S. L., Allen, H. L., Wright, B. J., Baicker, K. & Finkelstein, A. N. (2014), 'Medicaid Increases Emergency-Department Use: Evidence from Oregon's Health Insurance Experiment', *Science* **343**(6168), 263–268.
URL: *https://www.science.org/doi/10.1126/science.1246183*

US Census Bureau & Yang, J. (2021), 'Percentage of people covered by Medicaid in the United States from 1990 to 2020'.
URL: *https://www.statista.com/statistics/200960/percentage-of-americans-covered-by-medicaid/*

The Effect of State Mandates of College-level Curriculum in Secondary Schools on Students' SAT Scores

Tianyuan Zheng

March 2, 2022

Abstract

There are many long-standing equity issues in secondary school education in the United States. According to previous studies, there are at least 6 times more schools that offer 4 or less AP courses than there are schools offering more than 18, and the north tend to have a higher percentage of advanced high schools than the south. To ensure that high schools provide adequate academic rigor and opportunities to students, some states enact a curriculum mandate on secondary education institutes, requiring them to offer college-level curricula, the most common being AP (Advanced Placement) courses. This paper attempts to analyze the effect of such state mandates on student achievement measured by SAT scores and explores how different groups of students are affected differently by the mandate. In a somewhat counter-intuitive light, this paper finds state-level enforcement of college-level curriculum to negatively correlate with students' SAT scores - notably, for states with high SAT participation rates ($> 75\%$), having an active mandate significantly lowers students' SAT composite score average by 23 and 18 points for males and females respectively. Additionally, when grouped by GPA, both high-achieving and low-achieving students suffered greater magnitudes of point loss than A-/B students.

In the United States, high-achieving secondary school students may elect to take college-level courses to enhance their academic experiences and, in most cases, attempt to earn credits that are both recorded in their high school transcripts and transferable to their college of matriculation. Most students take college-level courses through widely acknowledged programs such as the Advanced Placement (AP) program by the College Board, and a globally recognized, non-profit program known as the International Baccalaureate (IB), while some take part in dual-enrollment programs usually run by their high school and a local college in conjoinment.

It is a common fact that the quality and policies regarding education vary substantially across all states in the United States. The significant imbalance between high schools is immediately clear as one observes some high schools offering some of the world's most advanced courses and equipment, while many more high schools suffer from a lack of graduation requirements and low academic performance. As a result, students from under-privileged high schools do not have advanced academic resources to match with their more affluent counterparts. Indeed, a report published by the Center for American Progress indicates that there are more than 6 times as many schools in the United States that offer less than 4 AP courses than schools that offer 18-37 AP courses. Specifically, it points out that while 73% of Virginia's high school students attend a school offering more than 11 AP courses, only 14% of high school students in the less-developed state of Mississippi enjoy the same AP course opportunities.

Following the observation above and my own experience in spending my high school years in a U.S. state that is considered less proficient in education, I wondered if student achievement would be raised if high schools were required to offer opportunities for students to take college-level courses. In other words, this paper is interested in examining the question below:

Does state mandate on schools/districts to provide AP classes cause students' academic achievement, measured by SAT scores, to improve?

One interesting observation revealed by the data collection process is that states with low participation rates (i.e. "group 1" states, as defined above) yielded much higher average scores throughout categories than states with wide participation in the SAT (i.e. "group 2" states). This phenomenon is not surprising, though, as states with low participation rates in the SAT often have close to 100% participation rate in the ACT (American College Testing), another

process in this paper, students who took the SAT in group 1 states have an average GPA that is close to half a letter grade (10%) higher than their group-2-state peers; it is speculated that their decision to take the SAT were mainly incentivized by their motivation to apply to selective colleges and talent search programs.

Thus, following the general question, another interest of study is raised:

Are there significant differences in the extent to which state mandate on schools/districts to provide AP classes have on SAT achievements of students from group 1 states and group 2 states?

This paper attempts to answer those questions by performing differences-in-differences regressions for all observations and separately for each group of states. Following an analysis run on simple averages of SAT scores, this paper also explores variants of the two aforementioned questions with respect to different gender, ethnicity, academic achievement, and nationality groups.

2 Context

Literature Review

Education policies have long been one of the key driving forces of student achievement, yet sometimes it is not immediately clear which, if any, cohorts will be significantly benefited by them. To better construct our assumptions for this paper and gain insights into specific cohorts that policies may benefit to different extents, I find the following papers of interest in understanding the fundamental questions of this paper.

In terms of basic demographics, the grouping design in this paper is justified by the result from the Education Data Initiative that 69.1% of the high school students who graduate from high school will go on to attend college. As one may assume that most students who take the SAT plan on going to college, this threshold has helped determine the percentage participation cutoff for group 2 states in my paper to be above 70%. Furthermore, according to the Center for American Progress, even when schools have similar offerings of college classes, underrepresented ethnicity groups are still generally found under-performing in AP exams. This finding

There has been several studies focusing on AP students' achievements compared to their non-AP counterparts. Some have found statistically significant results - indeed, a paper (Warne, 2017) references that AP students often score better on standardized tests than "average" students (Ewing, Camara, & Millsap, 2006; Mattern, Shaw, & Xiong, 2009; McKillip & Rawls, 2013). Another recent study by College Board (2021) indicates that AP students have "significantly stronger college outcomes" even when they have scored 1 or 2 on AP exams.

Furthermore, a study by the Mid-Atlantic Regional Educational Laboratory at Mathematica discovered that high schools that mandate students to take at least 1 AP course generally had higher AP exam passing rates than schools that did not have this mandate, implying that college-level curriculum can enhance a secondary school student's academic achievement. However, some (Warne) voice concern over the fact that many aspects of the AP program are "poorly understood" due to the differences between the 34 AP courses offered; specifically, it is difficult to draw causal inferences between the rates of AP participation and achievements because I do not have a reliable way of quantifying student interactions with the AP program, including but not limited to the participation rate of AP exams and the way AP classes are taught.

This paper aims to explore if state mandates on college-level curriculum in high school are indeed effective, whether by parts or in whole - and, if so, which types of student it would benefit the most academically. It attempts to diversify the pool of existing literature (which are mostly based on assumptions of high schools offering AP classes) by adding the dimension of policy into the analysis and taking the effect of policy enforcement - or lack thereof - into the account.

Policy Background

This paper particularly concerns the existence and variation of statewide mandates on secondary schools offering college-level courses. According to a table presented by Education Commission of the States (ECS), as of 2016, 28 states do not require high schools to offer any type of college-level courses; eight states (Arkansas, Connecticut, Indiana, Iowa, Louisiana, Mississippi, South Carolina, West Virginia) and the District of Columbia required each high school to offer at least one AP course, and fourteen states follow a looser version of the mandate

as International Baccalaureate (IB) programs or Cambridge courses.

Curiously, ECS recently provided an update of the same table as this paper was in progress containing information accurate until October 2021. At the time of writing, only 3 states (Arkansas, Indiana, South Carolina) maintained this mandate; West Virginia and Mississippi have loosened the mandate to the district level, and all other states are silent on this issue. One speculates that this may be one of the many abrupt changes and reforms in policy due to the COVID-19 pandemic. Since AP course instructions and assessments were forced to take place in less controlled settings (i.e. online), the AP program has somewhat suffered a loss of credibility as colleges reportedly adjusted policies regarding acceptance of AP credits in place of introductory-level courses in 2020 and 2021. Therefore, Advanced Placement - along with similar programs - may be seen as less effective by states. To some extent, this interesting event aligns with the result of my paper - many states may have given up their mandates as they found compulsory AP-course offers in high schools to have a negative effect on students' achievements.

Assumptions Required for Validity of the Proposed Empirical Test

For the difference-in-difference inference to perform as expected, this paper assumes two pre-treatment parallel trends: First, the change in average SAT scores before and after the introduction of the state mandate would have been the same (or similar) for both treatment states and control states had the mandate not been in effect. Second, the states in the same group have parallel trend in average SAT scores and the percentage of students in each graduating class who took the SAT. This paper also assumes that there have been no other changes in education law that is relevant to high school curriculum or SAT participation rates for graduating high school classes in 2006-2015.

Problems Anticipated

The most concerning problem by far is the lack of individualized data on SAT scores. Since all available data come in pooled form, I am only able to view each state per year as a singular observation, leading to insufficient observations and wide margins of standard errors. For this reason, concluding with statistical significance may have been more difficult with existing data

Another issue emerges as PDF reports provided by the College Board display data in inconsistent formats. Perhaps the most striking example is the lack of data on the standard deviation of SAT scores - reports of later years have provided information on standard deviation of average scores, but those of former years have not. Therefore, it is difficult to counteract the heteroskedasticity of average scores across years.

To exemplify this scenario, one observes that some averages in provided PDFs, such as the one for Puerto Rican males in class of 2009 taking the test in Wisconsin, is determined by 9 students; on the contrary, the same average for Massachusetts is determined by 503 students. As a result, assuming the average of 9 scores is more likely to be affected by random chance than that of 503 scores.

Finally, as mentioned in the previous section, there may be more policies concerning student academic achievement, which may affect SAT scores more than the state mandate on college-level curricula. For instance, one such policy is a state mandate for all students to take the SAT. (For similar reasons, states that saw a significant increase in SAT participation rates between the years 2006-2015, such as Delaware, was excluded from the data set.) However, there may be more subtle, specific educational policies tailored to the individual states that are not immediately visible from the data collected for this paper.

3 Data

This paper observes SAT scores of students who graduated high school over the years 2006-2015 and took the SAT test at least once physically the 50 United States and District of Columbia. To complete the intended analysis, it also builds several other types of data, including state laws on college-level curricula offered in secondary schools and the percentage of total students taking the SAT in a given class year. The procedures for data collection and considerations for exclusion criteria are detailed below.

First, to identify which states have enacted the mandate in the desired time frame for this paper, I consulted a summarized table on state mandates made available by the National Center for Educational Statistics. From there on, education laws of states of interest are consulted to determine which year the mandates are individually put into place. This information allows me

to college-level class requirements in secondary education, and value 1 if the state places *any* restrictions on schools or districts offering Advanced Placement (AP) courses, International Baccalaureate (IB) programs, or any other type of certified college level courses in high schools.

Along with AP and IB programs, high schools across the United States occasionally participate in dual-enrollment programs, in which a high school student can be concurrently enrolled in both their high school and a local higher institution, such as a college or university. Scenarios surrounding dual-enrollment can be ambiguous - students may or may not receive credit at their high school or the higher institution. Therefore, for the purpose of this study, I excluded any state policies involving dual-enrollment from our consideration of "college-level curriculum mandate".

This paper also utilizes data on the percentage of enrolled high school students taking the SAT in a given state and year for two purposes: state grouping and pre-treatment parallel-trend comparison. This data is viable for collection through the Education Commission of the States. By concatenating 7 forms, I was able to obtain data on trends of student participation in the SAT from years 2006-2015.

The collection of data on SAT scores is restricted to students who graduated in the years 2006-2015 and took the 2006-updated version of the SAT test with reading, math, and writing sub-scores in 6 selected states (Alabama, Arkansas, Wisconsin, Connecticut, Massachusetts, and New Jersey). The states are grouped into 2 categories by SAT participation rate: group 1 (AL, AR, WI) had consistent low statewide SAT participation ($< 10\%$), and group 2 (CT, MA, NJ) had consistent high participation ($> 75\%$) in years 2006-2015. Here, "consistent" is defined as having 20% or less fluctuation in participation percentage for group 1, and 10% or less fluctuation for group 2. Those states are selected so that both groups have one treatment state that instituted a mandate between year 2006 to 2015: respectively, Arkansas' mandate was first in effect in 2010, and Connecticut's in 2012. The rest of the states are controls; in addition, each state must satisfy the consistent participation trend. It was surprisingly difficult to find states with high participation rates that satisfied the criteria for this paper - this is exemplified by New York, whose participation rate dropped from 90% in 2006 to 76% in 2015, and Delaware, whose participation rate rose from 73% to 100%, curiously even without a statewide SAT mandate.

On the other hand, participation rate varied less dramatically in group 1 states; however, small variations (such as from 7% to 4%, in the case of Nevada) indeed mark larger changes.

In years 2006-2015, the SAT exam tested students on 3 subjects: reading, math, and writing, each scored on a scale from 200 to 800. The minimum composite score on the SAT is thus 600, and maximum 2400. Each student is only counted once no matter how many times they participated in the SAT; if they participated more than once, their latest score and survey responses are used. Ideally, anonymous individualized data on SAT scores would be tremendously helpful to this paper. Unfortunately, those data are not available upon request. Instead, the SAT score data has been collected by hand from exhaustive PDF reports by state and year of graduation from Archive SAT Suite Data provided by the College Board. In addition to total average SAT scores, pooled average SAT scores by gender, race, grade point average, and nationality are collected, along with average GPA, total number of students, and students taking their last SAT in each grade level as controls. Table 1 in Appendix I as well as Figure 1, 2 in Appendix II provide a brief summary of the data set.

4 Empirical Methodology

Differences-in-differences model

This analysis follows the diff-in-diff regression model below and perform a different test for each group of states:

$$Y_{s,t} = \alpha + \delta_{DD} COL_{s,t} + \sum_{\text{k = state}} \beta_k STATE_{s,t} + \sum_{j}^{\text{years}} \gamma_j YEAR_{j,t} + e_{s,t}$$

where $Y_{s,t}$ is the average SAT score, depending on subcategory of interest; $\alpha, \delta_{DD}, \beta_k, \gamma_j$ are estimators, and $e_{s,t}$ is the error term with respect to state s, year t; $COL_{s,t}$ indicates whether state s has mandated college-level curriculum in year t, which is our main regressor at interest; and $STATE_{s,t}$ and $YEAR_{j,t}$ are indicator variables that take canonical values 0/1.

This regression is first run on all observations to estimate the general effect of state mandates on student achievement, then run separately for group 1 and group 2 states as one of the possibly ways to explore the second question outlined above. Standard errors will be clustered

at the state level.

Methods for Testing Assumptions

To ensure the validity of this paper, one needs to test for the parallel pre-trend assumption. One method to do so is by visual inspection; Figures 1 and 2 in Appendix II do so by graphing general SAT score and SAT participation observations.

Figure 1 shows average composite SAT score by state in years 2006-2015; the first dotted line (at $x = 2010$) denotes the first year following Arkansas' state mandate, and the second dotted line (at $x = 2012$) denotes the first year following Connecticut's state mandate. We see that group 1 SAT score trends generally align before 2010, and group 2 SAT score trends align before 2012.

Figure 2 depicts the percentage of students in graduating class of 2006-2015 that took the SAT test in each of the 6 states. Group 1 observations are highlighted in red lines, while group 2 observations are in blue lines. One wishes for the trends of SAT participation to be as close to constant as possible, so to minimize the effect of significant SAT participation changes on SAT scores. The 6 states chosen satisfy our criteria because they (1) have similar trends within group with little fluctuation, and (2) stay in their defined range of participation percentage criteria (i.e. less than 10% and more than 75%, respectively.)

Pre-treatment trends can also be tested by using an event study model detailed with the equation below:

$$Y_{s,t} = \alpha + \mu_s + \delta_t + \sum_{j=-3, j\neq -1}^{3} \beta_j c_{s,t}^j + e_{s,t}$$

Where $Y_{s,t}$ is the observed average SAT score at state s and year t, μ_s is the state indicator, δ_t is the year indicator, and $e_{s,t}$ is the error term. We define $c_{s,t}^j$ below:

$$c_{s,t}^j = \begin{cases} \sum_{r=-\infty}^{-3} \Delta x_{s,t-r} & \text{, if } j < -3 \\ \Delta x_{s,t-j} & \text{, if } -3 < j < -3 \\ \sum_{r=3}^{\infty} \Delta x_{s,t-r} & \text{, if } j > 3 \end{cases}$$

where $x_{s,t}$ denotes the treatment indicator. The regression result of the event study is plotted as shown in Figure 3. We see that the lag effect of mandates is not significantly different from 0,

and thus the parallel trend assumption is not rejected.

5 Results

Due to low number of observations, most results from regressions are statistically insignificant. However, one can still deduce several intriguing implications from inspecting the score trends shown in tables. Specifically, when inspecting average scores controlled by different subgroups (i.e. gender, ethnicity, nationality, and grade point average), the results become more sophisticated and surprising.

Table 2 describes the effect of having a college-level curriculum state mandate on students' average SAT composite scores (i.e. the sum of scores in 3 subjects: reading, mathematics, and writing). The regression models included year and state effects for year 2006-2015 and all 6 states. One observes that the year effect on SAT score is mostly statistically insignificant. Results in column (1) indicate that across all states, students who took the SAT at a time the mandate was in effect scored about 9 points less on the SAT than their non-mandate-bound counterparts, but this result is insignificant. However, the results in columns (2) and (3) seem more interesting - students from Group 1 (i.e. low-participation) states are 95% likely to gain roughly -14 to 35 points from the mandate with standard deviation 12.35,making the effect not significantly different from 0. However, students from Group 2 (i.e. high-participation) states are found to score significantly lower (by around 20 points) when the mandate is in effect (se=4.245).

Table 3 similarly displays effects of mandates on SAT composite scores by state grouping: it was produced using the same regression model as Table 2, only with gender-specific parameters. From columns (3) and (6), for both males and females in Group 2 states, the mandate seems to lower student average score with statistical significance, by 23.21 and 18.37 points respectively. Furthermore, though statistically insignificant, we see that when categorized by state groups, females benefit slightly more than males from the mandate - in Group 1, the effect of mandate was around 9 points higher for female than male, and around 5 points higher for those in Group 2.

Table 4, 5, and 6 resulted from regression analysis that focused on a specific type of categorization. In Table 4, students who identify as U.S. citizens, permanent residents, and inter-

national are compared. Specifically, in each section of the table, the effect of mandates (with respect to all subjects) on average scores are displayed. Again, the regressions themselves included state and year effects, but they are not displayed due to the sheer volume of numbers. Interestingly, international students scored 69.36 points lower in mathematics when affected by mandate with statistical significance (se=18.17).

In Table 5, the effect of mandates on different race groups (as defined by the College Board) are displayed by gender. Interestingly, as suggested by columns (4), (5), and (6), while SAT scores of males in every race category are not significantly different from 0, the effect of the mandate on scores of females are almost always negative, with statistical significance in slightly more race categories, with the exception of Puerto Ricans. It is notable that the negative effect of the mandate was especially significant on Asian Females, as their composite, reading, and math scores show a significant decrease with p-value < 0.05 in general; the composite score of group 2 Asian females are significantly reduced by 41.21 points with standard error 7.094.

Results from columns (4), (5), and (6), albeit mostly statistically insignificant, are nevertheless striking, as they seem to contradict that of Table 2. However, this naive interpretation could be deceiving as it assumes the same weight for each observation from each state and year, yet in fact these observations are averages of significantly different number of students across observations. For example, the class of 2006 saw 1434 white females taking the SAT in Alabama, but 21709 in Massachusetts, yet their averages are each counted as one observation - thus, the individual scores of the 1434 students in Alabama each bears a weight that is 10 times higher than the scores of the 21709 students in Massachusetts. Indeed, the inclusion of columns (2) and (3) confirms Table 2's result: one observes from column (3) that when controlled by state groups, the effect of state mandates become more statistically significant for Asians, African Americans, various Hispanic groups, and white male.

Lastly, the effect of mandates on students grouped by their grade point average (GPA) is reported in Table 6. A graphical representation of Table 5's result is also reported in Figure 4 in Appendix II. Similarly, the effects were measured to be more statistically significant for Group 2 observations - the negative effect of having a state mandate on student scores seems to be most consequential and significant on both ends of the GPA spectrum, i.e. for students who score A+ and those who score C and below. On the other hand, the effects on scores from group 1 states

across all GPA groups were slightly positive and insignificant. It is also notable that within group 1 state observations, those representing higher GPA groups were averages of significantly larger populations than those representing lower GPA groups due to the self-selectivity of students who take the SAT from group 1 states. In fact, there were so few students with D and below GPA taking the SAT that several group 1 states did not report their averages. Therefore, the results from column (2) may be less useful.

6 Conclusion

Summary and Implications

This study performed multiple regression analyses under near-identical models as described in the empirical methodology section. Specifically, the $Y_{s,t}$ term in the regression stood for total score and each subject score for each subcategory defined by race, GPA, and nativity. This regression design is a direct result of the aggregated form that the SAT score data was in. Hypothetically, if the SAT data are on the individual level, then the $Y_{s,t}$ term would be student score (composite and by each subject), while subcategory information is included as controls.

Overall, the regression analyses mostly yielded statistically insignificant results, partly due to having only 60 or less observations and the imbalance of representation in each observation. Nevertheless, the results widely indicated that having a state mandate requiring high schools to offer college-level courses can be harmful rather than helpful to students' academic achievements; across all nationality groups and most ethnicity and GPA groups, the existence of a state mandate negatively associated with students' SAT scores.

It was perhaps counter-intuitive to find that having a college-curriculum mandate seem to mostly "harm" students rather than improving their scores. One interpretation may be that as schools start including required AP courses (or similarly acknowledged college-level courses), they may have dedicated the most qualified instructors to teach them, leaving the majority of the student population - who takes regular classes - to even less ideal education resources, resulting in a drop in SAT performance. The fact that having a mandate is associated with more than 50 points of decrease in SAT score for less-than-average students from group 2 states aligns with this interpretation.

On the other hand, the statistically significant decrease in score for high-achieving students from group 2 states invites a more sophisticated interpretation. Since advanced students are more intrinsically motivated to find more academic resources, they were likely already excelling in at least one academic subject, and an additional AP course in their respective disciplines of expertise may be less helpful in increasing their scores. On the other hand, those high-achieving students may not be equally excelling in every subject - a report from College Board provides evidence that approximately 92.8% of AP students took less than 4 AP exams in 2009. Which implies that most high-achieving students still took at least one regular core course each year. Therefore, an additional AP course offered in one of their weaker subjects may analogously have a slightly detrimental effect as instructors of regular level courses in such subjects are highly possibly weaker than AP instructors.

Alternatively, one may interpret this result by noticing that students often spend more time preparing for assessments in college-level courses than regular high school courses, which are less challenging and time-consuming. Therefore, the students that elect to take advanced courses (who are mostly above-average) may be allotting less time to prepare for the SAT, resulting in a lower score.

Additional Issues Encountered

Alongside problems anticipated at pre-analysis stage, this paper has somewhat suffered from the incompleteness of data in Archive SAT suite reports in addition to the already sparse data set. Especially for group 1 states, some observations concerning average scores of minority groups (such as Puerto Ricans) were not reported due to privacy concerns as the numbers of test takers in such categories were often extremely low.

During the analysis stage, I also realized that since the numbers of students in each subcategory were vastly different (ranging from 1 to tens of thousands), the average scores are possibly a skewed representation of the SAT participants. Again, since standard deviation of each observation of score average was not reported completely, a better solution to this problem remains to be found.

Potential Future Improvements

This study would be made significantly stronger from having access to anonymized, individualized data, as it currently suffers from the fact that data are aggregated averages. In the less ideal situation that individualized data cannot be retrieved, this paper can still benefit from adding more observations from more states and years, which this paper was also limited from due to time restraints. One may alternatively improve or renew the analyses of this paper by using a different data set - such as the Stanford Education Data Archive (SEDA) - to measure student achievement.

Appendix I: Tables

Table 1: Average SAT Composite Scores for High School Class of 2006-2015 in Select States

	All Observations		Group 1 States		Group 2 States	
	mean	sd	mean	sd	mean	sd
Avg. Composite Score	1616.82	(96.75)	1702.93	(58.20)	1530.70	(17.70)
Avg. Reading	538.98	(35.50)	571.67	(17.24)	506.30	(7.544)
Avg. Math	545.58	(32.89)	572.70	(25.08)	518.47	(7.104)
Avg. Writing	532.25	(29.30)	558.57	(16.52)	505.93	(6.422)
Avg. Female Total Score	1600.57	(94.58)	1682.40	(63.73)	1518.73	(16.75)
Avg. Male Total Score	1636.27	(100.4)	1728.37	(51.02)	1544.17	(18.69)
Avg. Composite Score by GPA Groups						
GPA: A+	1886.50	(45.39)	1903.27	(55.01)	1869.73	(24.17)
GPA: A	1790.97	(45.67)	1801.87	(54.40)	1780.07	(32.22)
GPA: A-	1682.12	(41.40)	1674.13	(45.17)	1690.10	(36.24)
GPA: B	1493.15	(53.14)	1518.70	(54.92)	1467.60	(37.12)
GPA: C	1291.98	(65.51)	1324.37	(74.39)	1259.60	(32.07)
GPA: D or below	1249.63	(114.2)	1437.60	(208.4)	1218.30	(46.80)
Avg. Composite Score by Gender and Race/Ethnicity						
Nat. Ind. or Alaskan (M)	1552.39	(123.6)	1652.85	(111.4)	1465.33	(37.27)
Nat. Ind. or Alaskan (F)	1519.70	(119.8)	1611.37	(113.0)	1437.20	(37.68)
Asian/As. Am. (M)	1712.90	(86.53)	1754.87	(101.7)	1670.93	(35.27)
Asian/As. Am. (F)	1692.78	(55.24)	1721.77	(53.84)	1663.80	(39.63)
Black/ Af. Am. (M)	1346.12	(96.60)	1429.53	(64.98)	1262.70	(19.15)
Black/ Af. Am. (F)	1353.57	(93.97)	1433.70	(66.22)	1273.43	(17.11)
Mexican (M)	1528.32	(116.1)	1618.40	(70.41)	1438.23	(75.20)
Mexican (F)	1492.13	(126.0)	1591.33	(77.36)	1392.93	(77.29)
Puerto Rico (M)	1415.86	(168.8)	1641.36	(105.8)	1310.63	(34.92)
Puerto Rico (F)	1396.65	(170.8)	1609.25	(107.0)	1283.27	(33.47)
Other Lat. Am. (M)	1510.95	(159.0)	1652.67	(97.19)	1369.23	(20.28)
Other Lat. Am. (F)	1465.87	(150.7)	1603.87	(80.74)	1327.87	(17.38)
White (M)	1685.57	(89.18)	1768.70	(41.17)	1602.43	(13.69)
White (F)	1642.85	(97.51)	1705.40	(105.4)	1580.30	(12.31)
Other/NA (M)	1612.43	(113.5)	1714.37	(64.21)	1510.50	(24.31)
Other/NA (F)	1581.98	(122.8)	1690.30	(77.11)	1473.67	(21.93)
Avg. Composite Score by Nativity						
Citizen	1626.07	(97.86)	1711.63	(63.01)	1540.50	(19.08)
Permanent Resident	1578.40	(177.3)	1744.23	(56.71)	1412.57	(62.19)
International	1625.17	(499.3)	1582.50	(58.94)	1667.83	(707.1)
Observations	60		30		30	

mean coefficients; sd in parentheses clustered by state.

Group 1 states include AL, WI, AR; Group 2 states include MA, NJ, CT.

Table 2: Effect of College curriculum mandate on Average SAT Composite Score by State Grouping

	All	Group 1	Group 2
Avg. SAT composite score	-9.273	10.71	-20.17***
	(10.64)	(12.35)	(4.245)
Year effects included?	Yes	Yes	Yes
State effects included?	Yes	Yes	Yes
group * mandate	2.618		
	(6.058)		
Observations	60	30	30
R^2	0.981	0.932	0.960

Standard errors in parentheses, clustered at state level

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

Table 3: Effect of College curriculum mandate on Average SAT Composite Score by State Grouping and Gender, Controlled on Year and State

	Male			**Female**		
	All	Group 1	Group 2	All	Group 1	Group 2
	(1)	(2)	(3)	(4)	(5)	(6)
mandate	-9.846	5.125	-23.21***	-2.327	14.38	-18.37***
	(6.623)	(13.14)	(4.908)	(6.520)	(12.58)	(3.865)
Year effect included?	Yes	Yes	Yes	Yes	Yes	Yes
State effect included?	Yes	Yes	Yes	Yes	Yes	Yes
Constant	1679.7***	1688.1***	1549.2***	1625.4***	1639.2***	1527.3***
	(13.10)	(19.33)	(1.598)	(16.31)	(22.73)	(2.491)
Observations	60	30	30	60	30	30
R^2	0.985	0.911	0.957	0.977	0.940	0.960

Standard errors in parentheses, clustered by state

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

Table 4: Diff-in-diff Regression of Average SAT score, Composite and by Subject, by Nativity Categories

	Composite	Reading	Math	Writing
All	-9.273	-5.588	5.969	-9.654
	(-0.87)	(-1.44)	(1.57)	(-1.77)
Avg. Composite Score	1649.4***	554.4***	547.5***	547.4***
	(108.08)	(134.12)	(106.89)	(88.49)
Citizen	-0.575	-0.709	7.176	-7.042
	(12.15)	(6.550)	(4.362)	(4.769)
Avg. Composite Score	1660.2***	563.1***	547.1***	550.0***
	(19.06)	(7.016)	(5.713)	(7.016)
Permenant Resident	-106.4	-46.74	-19.70	-39.95
	(87.75)	(32.68)	(32.61)	(31.03)
Avg. Composite Score	1694.1***	540.1***	601.6***	552.5***
	(16.49)	(6.219)	(6.684)	(7.140)
International	78.83	209.7	-69.36***	-61.54*
	(474.1)	(461.9)	(18.17)	(23.52)
Avg. Composite Score	1549.1***	451.2***	601.3***	496.5***
	(113.3)	(115.3)	(8.268)	(13.77)
Observations	60	60	60	60

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

Standard errors in parentheses clustered by state; year and state controls were included in regression but omitted in this table. Each section (All, Citizen, Perm. Res. and International) are separate regressions.

Table 5: Effect of Mandate on SAT Score, Break-down on Gender and Ethnicity Groups

	Composite	Group 1	Group 2	Reading	Math	Writing
	(1)	(2)	(3)	(4)	(5)	(6)
Nat. Am/Alaskan(M)	-8.296	105.1	-59.71	-3.722	-8.243	3.669
	(40.64)	(77.03)	(35.65)	(13.30)	(14.91)	(15.26)
Nat. Am/Alaskan(F)	-86.61	-163.8	-23.25	-26.90	-35.16*	-24.55
	(46.54)	(111.1)	(26.71)	(16.34)	(17.23)	(16.33)
Asian(M)	3.788	12.00	-15.92	-8.683	-5.885	18.36
	(32.35)	(51.00)	(11.56)	(6.203)	(5.054)	(28.22)
Asian(F)	-37.01*	-38.79	-41.21***	-13.85*	-12.31*	-10.86
	(16.25)	(30.41)	(7.094)	(6.044)	(5.465)	(5.907)
African American(M)	24.96	78.50*	-27.75***	6.538	7.827	10.60
	(16.61)	(33.35)	(5.909)	(5.552)	(5.990)	(6.210)
African American(F)	-0.327	17.96	-17.04**	4.635	-4.365	-0.596
	(12.47)	(27.72)	(5.512)	(4.588)	(4.403)	(4.994)
Mexican(M)	-9.471	1.417	-40.37	-2.788	-4.279	-2.404
	(29.48)	(50.48)	(31.58)	(10.74)	(10.98)	(10.27)
Mexican(F)	-19.13	-20.29	-15.50	-10.81	-5.067	-3.260
	(36.35)	(77.80)	(14.46)	(13.36)	(10.96)	(13.98)
Puerto Rican(M)	-34.08	Omitted	-37.79**	-11.62	-19.72	-2.738
	(31.66)	(.)	(10.31)	(9.423)	(13.61)	(10.92)
Puerto Rican(F)	9.488	Omitted	-27.92**	3.077	-2.494	8.905
	(28.37)	(.)	(9.011)	(10.28)	(10.95)	(9.462)
Other Hispanic(M)	-25.60	-62.58	-26.87**	-12.40	-3.067	-10.12
	(42.85)	(76.49)	(9.229)	(14.88)	(15.22)	(14.44)
Other Hispanic(F)	-48.35	-91.33	-15.29**	-14.38	-18.82*	-15.14
	(24.34)	(47.45)	(4.552)	(10.47)	(7.703)	(9.227)
White(M)	-7.346	-3.792	-6.542*	-4.346	-0.183	-2.817
	(6.652)	(10.90)	(2.896)	(2.361)	(1.745)	(3.280)
White(F)	-58.21	-71.04	-4.292	-1.635	-55.86	-0.721
	(52.80)	(75.61)	(3.370)	(2.214)	(51.14)	(2.752)
Other(M)	-5.798	7.333	-19.67	-3.019	-6.808	4.029
	(22.95)	(50.41)	(15.75)	(11.20)	(7.351)	(7.905)
Other(F)	-9.942	-14.42	1.917	-9.356	0.250	-0.837
	(32.97)	(62.09)	(15.20)	(12.23)	(10.45)	(12.23)
Observations	60	30	30	60	60	60

Standard errors in parentheses clustered by state. Scores for Puerto Rican not reported in select group 1 states by the College Board.

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

Table 6: Effect of College curriculum mandate on Average SAT Composite Score by State Grouping and GPA

	All	Group 1	Group 2
	(1)	(2)	(3)
GPA= A+	-14.95	7.750	-40.33***
	(8.973)	(11.39)	(6.725)
GPA = A	-8.923	1.625	-21.50*
	(7.315)	(11.69)	(9.046)
GPA = A-	2.163	14.13	-9.625
	(8.638)	(16.40)	(5.534)
GPA = B	-4.154	11.38	-16.87**
	(8.482)	(14.76)	(4.508)
GPA = C	-6.625	9.000	-26.62***
	(11.78)	(24.29)	(6.224)
GPA = D or below	-25.42	0	-53.42*
	(33.19)	(.)	(20.43)

Standard errors in parentheses clustered by state. Group 1 states did not report averages of students with GPA D or below due to the small number of test takers.

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

Appendix II: Figures

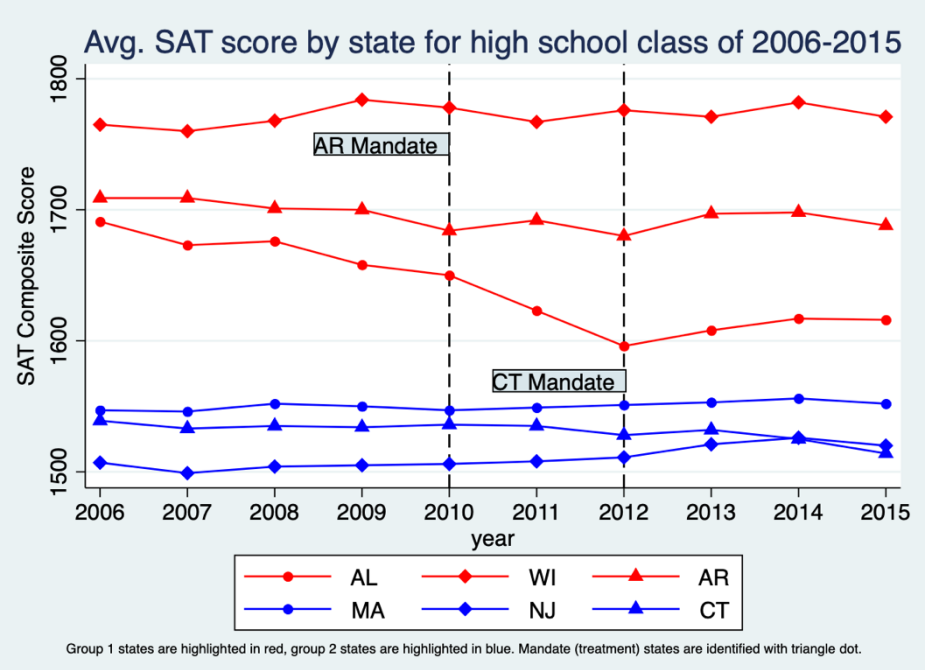


Figure 1: Average SAT scores by state for high school class of 2006-2015 in select states

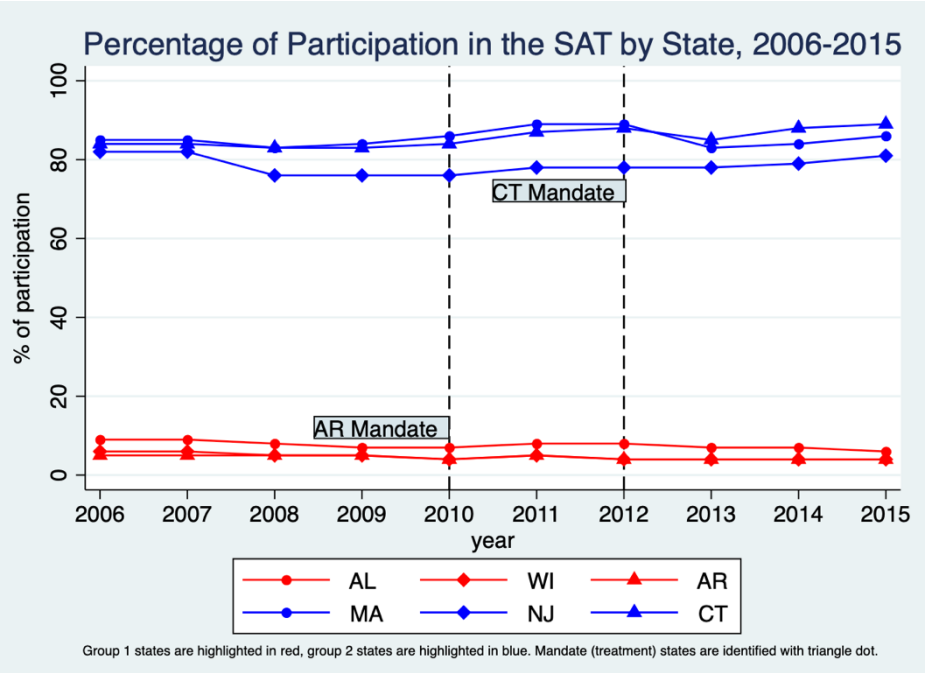


Figure 2: Percentage of Participation in the SAT by State, 2006-2015

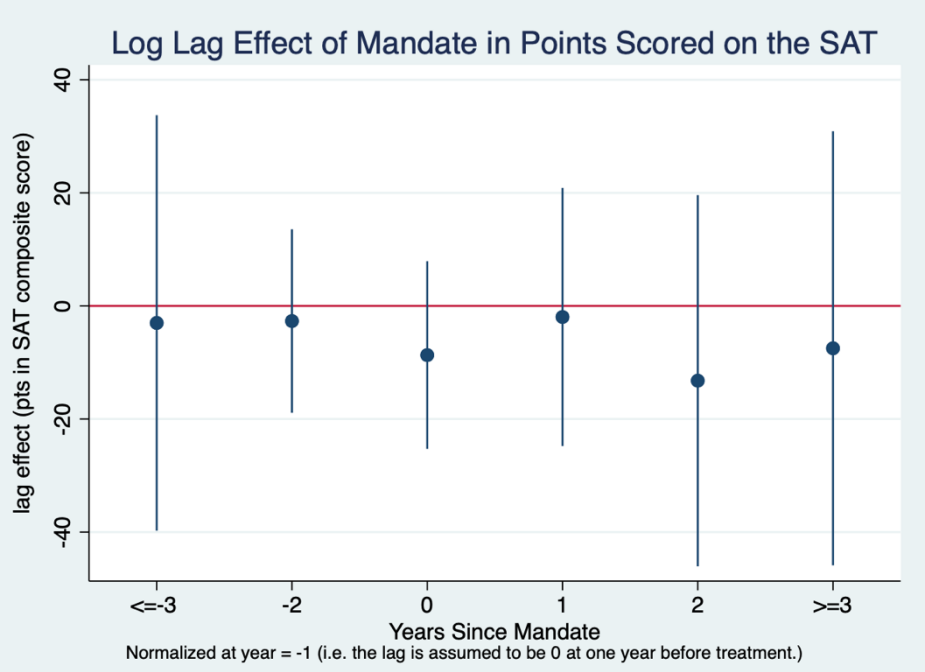


Figure 3: Log Lag Effect of Mandate in Points Scored on the SAT

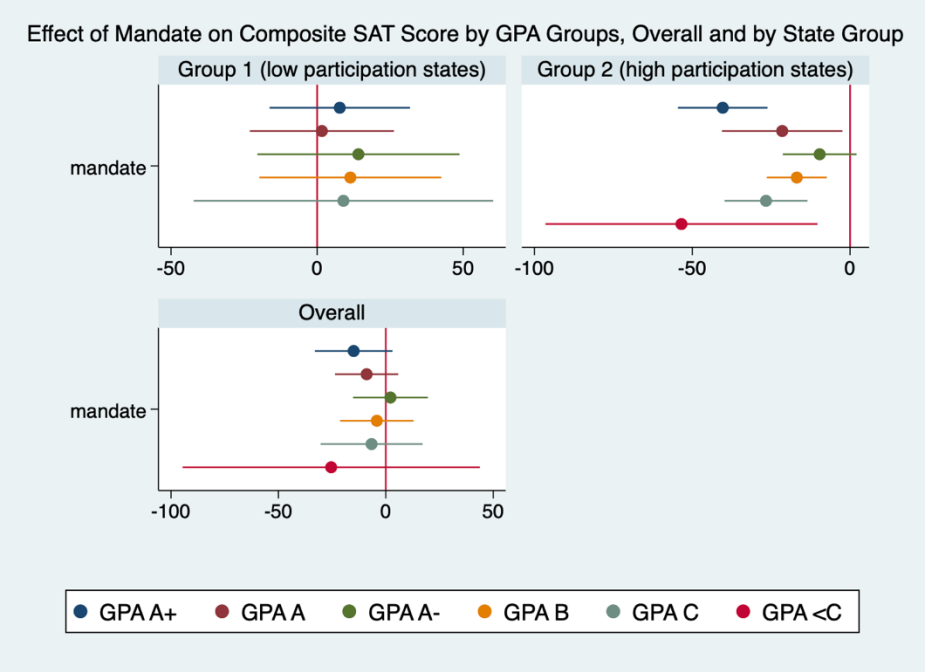


Figure 4: Effect of Mandate on Composite SAT Score by GPA Groups, Overall and by State Group

References

(2021). (rep.). *Digest of Education Statistics*. Retrieved from https://nces.ed.gov/programs/digest/.

(2021). (rep.). *Expanding Access to Advanced Placement Courses*.

Chatterji, R., Quirk, A., & Campbell, N. (2021, June 30). *Closing advanced coursework equity gaps for all students*. Retrieved November 17, 2021, from https://cdn.americanprogress.org/content/uploads/2021/06/29113737/AdvancedCourseWork-report.pdf.

College Board. (2018, March 5). *Archive SAT suite data – research – college board*. Research. Retrieved November 17, 2021, from https://research.collegeboard.org/programs/sat/data/archived.

College enrollment statistics [2021]: Total + by demographic. Education Data Initiative. (2021, August 7). Retrieved November 17, 2021, from https://educationdata.org/college-enrollment-statistics.

Dixon-Roman, E. J., Everson, H. T., & Mcardle, J. J. (2013). Race, poverty and SAT scores: Modeling the influences of family income on black and white high school students' sat performance. *Teachers College Record: The Voice of Scholarship in Education*, *115*(4), 1–33. https://doi.org/10.1177/016146811311500406

Education Commission of the States. (2021). *Advanced Placement: State Mandates AP Course Offerings*. Advanced placement: State mandates AP course offerings. Retrieved November 17, 2021, from https://ecs.secure.force.com/mbdata/MBQuestSNR?Rep=AP01.

New analyses find students who earn a 2 on an AP exam are prepared for the rigor of college courses. New Analyses Find Students Who Earn a 2 on an AP Exam Are Prepared for the Rigor of College Courses – All Access | College Board. (n.d.). Retrieved November 17, 2021, from https://allaccess.collegeboard.org/new-analyses-find-students-who-earn-2-ap-exam-are-prepared-rigor-college-courses.

Number of AP examinations per candidate - College Board. (n.d.). Retrieved November 28, 2021, from https://secure-media.collegeboard.org/digitalServices/pdf/research/2019/Number-of-Exams-Per-Student.doc.

Warne, R. T. (2017). Research on the academic benefits of the Advanced Placement Program. *SAGE Open*, *7*(1), 215824401668299. https://doi.org/10.1177/2158244016682996

Tackling Nutritional Inequality: The Effect of Food Financing Initiatives

Jenny Zhu

This paper analyzes the effect of statewide food financing initiatives in improving access to nutritious and affordable foods in disadvantaged communities. Utilizing comprehensive data from the Census Population Survey Food Security Supplement, I employ a differences-in-differences model as well as a triple difference specification to find that Pennsylvania's 2004 Fresh Food Financing Initiative resulted in an expected 300,000+ decrease in the number of food insecure individuals, with an even larger effect on households with a higher risk of food insecurity (estimated from their baseline characteristics). Moreover, I find that the policy appears to have no effect on households with the most severe conditions of food insecurity. My results not only contribute to existing research on the effectiveness of supply-side food financing policies, but also highlight the potential limitations of the mechanisms through which food financing initiatives have an impact.

1 Introduction

Recently, there has been increasing literature on food deserts, which are areas with limited access to supermarkets. Papers have largely examined disparities in access to healthy food in the United States (Walker et al., 2009), the relationship between food deserts and health (Stack, 2015), and food deserts and the causes of nutritional inequality (Allcott et al., 2017).

Food deserts are also connected to supermarket redlining, the phenomenon in which large chain grocery stores and supermarkets are disinclined to locate in disadvantaged neighborhoods, particularly low-income urban neighborhoods. Similarly, there are stark disparities — around a 10% gap — in access to supermarkets between predominantly black neighborhoods and white neighborhoods (The Reinvestment Fund). As a result of supermarkets closing down in marginalized communities and relocating to suburbs, these communities are disproportionately affected by more limited access to fresh

and nutritious food. Existing literature examines the impact of changes in urban retail food availability on the health status of the urban poor (Eisenhauer, 2001) and the effect of supermarket redlining on neighborhood vulnerability in Hartford, Connecticut (Zhang and Debarchanam, 2016). There is little research, however, on initiatives aimed at dismantling barriers to fresh food and whether or not they are effective.

In this paper, I uncover new analysis regarding the impact of statewide food financing initiatives, which are designed to increase the availability of healthy, affordable food in underserved communities, by assessing the 2004 Pennsylvania Fresh Food Financing Initiative (PA FFFI). Impact assessments have been conducted to evaluate the initiative's effects on real estate, economic activity, and supermarket operating costs, but there currently lacks empirical evidence of the effect of food financing on household food security, one of the most critical measures of food sufficiency. To investigate the causal impact of Pennsylvania's food financing initiative on food insecurity, I employ a difference-in-differences model using household-level data from the Current Population Survey Food Security Supplement. I find that the implementation of the 2004 PA FFFI is associated with an expected 300,000 + decrease in the number of individuals being food insecure; this effect is even larger for households below the 185% poverty level, but is not significant for households which are predominantly black or unemployed. Using a triple difference specification, I further corroborate my estimates by showing that the policy had a disproportionate effect on the households most likely to be food insecure, with the probability of the initiative reducing food insecurity being 5.1 percentage points higher among households which are predicted to be most at risk of food insecurity.

My findings contribute to current literature by identifying the causal impact of fresh food access on food insecurity levels, reinforcing existing research which show a relationship between accessibility of fresh food and health outcomes, factors closely tied to food insecurity. Additionally, my results indicating the ineffectiveness of the initiative on households with the most severe degree of food insecurity build

upon potential explanations for the limitations of supply-side food financing policies, including financial constraints among the most food insecure and differences in demand for healthy food.

The remainder of the paper proceeds as follows. Section 2 provides background on the Pennsylvania Fresh Food Financing Initiative, which ran from 2004-2010. Section 3 describes the dataset, overview of sample restrictions, and the construction of a food insecurity indicator. Section 4 describes the empirical methodology. Section 5 presents several robustness checks and quantifies the magnitude of the main results. Section 6 concludes.

2 Background

Pennsylvania's Fresh Food Financing Initiative (PA FFFI) was a statewide financing program that ran from 2004 to 2010. Established as part of the state's 2004 economic stimulus package, the PA FFFI was the only statewide food financing initiative implemented prior to 2010. Additionally, the PA FFFI did not overlap with any similar initiatives at the national level as the federal Healthy Food Financing Initiative (HFFI), which provides grants to a batch of community development organizations every year, only started awarding funds in 2011.

A partnership between the Commonwealth of Pennsylvania, Reinvestment Fund, The Food Trust, and the Urban Affairs Coalition, the FFFI's main objectives included developing and improving supermarkets in underserved communities to provide a more secure means of obtaining healthy food. In particular, one of the channels through which the FFFI increased access to fresh food was helping grocery stores in areas of need overcome financing barriers. The FFFI provided loans and grants to supermarkets and fresh food retailers for costs such as equipment, acquisition, construction, and various other improvements. To qualify for a FFFI grant, a supermarket must have been "located in a low-moderate income census tract and in a trade area that is underserved" (The Reinvestment Fund, 2011). Supermarkets must have also had a full selection of fresh fruits and vegetables or used their funds to increase their selection of fresh food (Center TRT, 2013).

By 2010, 88 projects were financed, with more than $73.2 million in loans and $12.1 million in grants approved, 400,000+ Pennsylvanians with improved access to healthy food, 1.67 million square feet of commercial space developed, and 5,000 jobs created or preserved (The Food Trust, 2015). One study reveals a possible link between the program and health outcomes, finding that from 2006-2010, a time period in which 20 of Philadelphia's grocery stores received FFFI funding, the city witnessed a 5% decline in rates of childhood obesity (Robbins, 2012).

This paper seeks to understand the impact of the initiative on food insecurity, along with potential channels through which there may have been an impact. First, the initiative may have impacted food insecurity rates through the means in which the funds were deployed. A closer look at specific case studies indicates that many supermarkets which received funding purchased equipment to increase the availability of fresh foods. For example, the FFFI provided $15,175 in grants to 29 corner stores in Philadelphia to purchase space-efficient refrigeration units. This channel of impact is thus driven by the supply-side availability of fresh food and related resources. Second, the jobs and increase in economic activity which resulted may have improved distressed households' abilities to afford sufficient and nutritious food. This effect is driven by households' income and overall economic well-being.

3 Data

3.1 Data

To analyze the impact of the 2004 PA FFFI, I use data from the 2000-2010 Current Population Survey: Food Security Supplement (CPS-FSS) datasets (ICPSR). The Food Security Supplement (FSS) is an annual supplement to the nationally representative, monthly CPS survey conducted with ~50,000 households in the US. The FSS includes household-level data on 476 variables, such as food sufficiency, food security status, amount spent on food, and use of federal food assistance programs. An advantage to this dataset is that it provides granular household level data covering a comprehensive scope of food insecurity measures. A limitation, however, is that the dataset is not panel-level, disallowing my model to

control for household fixed effects. Thus, it is important to control for variables which may be correlated with or even impact the main outcome of interest: food security status.

For my analysis, I cover the time period 2000-2010, using Pennsylvania (PA) as my treatment state and New Jersey (NJ) as my control state for the following reasons. Primarily, Pennsylvania was the first and only state to implement a statewide healthy food financing program aimed at supermarket development prior to 2010. To further, upon evaluation of state legislation on hunger, I find that there were no statewide legislative changes that happened in or after 2004 in PA and NJ that could reasonably have a major impact on food insecurity (National Conference of State Legislatures). Third, the federal Healthy Food Financing Initiative only started rewarding grants in 2011, ensuring that the control state, NJ, did not receive disproportionate federal funding.

To perform my analysis, I combine the FSS datasets over each of the years in the time period 2000-2010 and impose sample restrictions. Most importantly, I restrict the sample to include only PA and NJ since these are the two states being compared. Second, I restrict the sample to only include households without a missing or unknown entry for "Food Security Status," which is the main outcome of interest. Last, I identify fourteen variables in the sample to keep for the purposes of my analysis; many of these variables may have also been correlated with food security status, including race, income, food stamp/SNAP beneficiary status, and whether or not the household was below the 185% poverty level.

The final dataset[1] includes 68,654 observations at the household level, with approximately 6,000-7,000 observations for each year from 2000-2010 across both NJ and PA. The main outcome variable of interest is "Food Security Status", from which I construct four new binary variables to indicate if a household is *food secure, food insecure, very food insecure*, and either food insecure or very food insecure (to indicate some level of *overall food insecurity)*. As defined by USDA, a household is recorded to be *food insecure* if it reports three or more conditions on the CPS-FSS that indicate food insecurity.

[1] Additional details on data construction and sample statistics can be found in Section A of the Appendix.

Some conditions include if respondents were worried their food would run out before they got money to buy more or if respondents could not afford to eat balanced meals. A household is recorded to be *very food insecure* if, in addition to the three conditions listed for *food insecure*, the household must also report that adults in the household ate less than they felt they should and that adults cut the size of meals or skipped meals in 3 or more months. Figure 1, which displays food insecurity trends in NJ and PA from 2000-2010, suggests a drop in food insecurity rates in PA at the time the 2004 PA FFFI was implemented.

3.2 Balance Tests

To better understand whether the sample is statistically balanced between PA and NJ, I conduct a difference in means test between NJ and PA for 2002 and 2003. Results are shown in Table 1 below. As there do appear to be statistically significant differences in annual family income (and related, the percentage of households below the 185% poverty level), race, and food stamps/SNAP beneficiary status, I control for these variables in the difference-in-difference analysis. On the other hand, there does not seem to be a statistically significant difference in employment status, and Table 2 shows that adding controls for employment status to the model has no significant effect on the estimates either.

3.3 Constructing a Food Insecurity Risk Indicator

To explicitly zoom into populations more at risk of food insecurity (and thus would have been more impacted by the initiative), I construct a food insecurity risk indicator so I can better attribute any differences in the outcome variable to the initiative itself. The purpose of this indicator is to capture whether a household has qualities which make it more at risk of being food insecure. This strategy is based on "Endogenous Stratification in Randomized Experiments" (Abadie et al, 2013)[2]. To evaluate the reliability of the indicator, I run a regression of the risk indicator on the main outcome variable — the binary variable indicating if a household is either *food insecure* or *very food insecure*. Results for this regression are shown in Table 5, which indicate that being food insecure is associated with a 16.3

[2] See Appendix Section B for more details regarding the construction of the indicator itself.

percentage point increase in being predicted to be at risk of food insecurity. This coefficient estimate is both positive and statistically significant at the 1% level, corroborating the quality of the indicator.

4 Empirical Methods

4.1 Difference-in-Differences Specification

To analyze the impact of the 2004 PA FFFI on the change in the probability of being food insecure, I estimate the following difference-in-differences model:

$$Y_{ist} = \gamma PA_s + \lambda POST_t + \beta(PA_s x\, POST_t) + \delta_t + X_i\alpha + \epsilon_{ist} \quad (1)$$

βcaptures the effect of the 2004 PA FFFI on *overall food insecurity* Y_{ist} (probability of being either *food insecure* or *very food insecure*) among households in Pennsylvania after 2004. δ_t are year fixed effects and X_i is a vector of controls for race, food stamps/SNAP beneficiary status, annual family income, and poverty status. I include these controls to combat potential endogeneity concerns, in which ϵ_{st} may be correlated with $PA_S x\, POST_t$as well as the outcome variable Y_{ist}. As shown in the balance tests, these variables have statistically significant differences between NJ and PA in the two years preceding the initiative. Additionally, these variables likely have an effect on food insecurity status; evidence suggests that high risk of food insecurity among people of color persists even after socioeconomic factors are controlled for (Odoms-Young, 2019). Moreover, eligibility for SNAP is determined through particular low-income thresholds, and research finds a statistically significant relationship between poverty and food insecurity among children (Wight et al., 2014). Thus, I include controls for these plausibly confounding variables as they are likely to be associated with higher risk of food insecurity.

When running the analysis, I use robust standard errors to account for potential heteroskedasticity, but do not cluster standard errors at the state level since there are only two states, NJ and PA, in my model. Since the number of states corresponds to the number of clusters, clustered standard errors would

not be statistically meaningful in my analysis. Standard errors are thus calculated at the individual level and reflect sampling uncertainty, but not uncertainty about the causal effect of the initiative (since I am unable to see the true counterfactual for PA). Results from this model, which is the preferred specification in this paper, are presented in Table 2 and further discussed in Section 5.

4.2 Assumptions

The identification assumption is that absent the treatment — the 2004 PA FFFI — PA and NJ would have seen the same trend in food insecurity levels. To test this assumption, I evaluate the trend of percent food insecure and percent very food insecure in NJ and PA from 2000-2004 (prior to PA FFFI implementation) to assess if my outcome variable of interest follows the same pattern prior to any statewide initiative being implemented. Figure 1 displays the results of this analysis, showing that food insecurity levels follow largely similar and fairly stable trends in NJ and PA prior to 2004.

To analyze if the initiative even had an effect on grocery stores and that the impact was captured within a specified time frame within the treatment, I evaluate an array of case studies. The first store in PA to receive financing through the FFFI, Brown's Shoprite, did so in 2005. The impacts of the program also appear to be immediate as upon receiving a grant, grocery stores were expected to complete expansion efforts within a couple months, upon which fresh food options should have been available. As an example, Ha Ha's Market located in Philadelphia had improvements in their refrigeration units made in 2005 to prolong storage of and thereby increase their offerings in fresh food; sales at Ha Ha's Market increased ever since, revealing the demand for fresh food options. Thus, funds seem to have been used immediately for supermarket development and increasing fresh food availability such that the effects of financing on food security should have been captured within a reasonable time frame of the treatment.

4.3 Triple Difference Estimator

To create more power in detecting the causal impact of the 2004 PA FFFI, I use a triple difference-in-differences specification to zoom into the portion of the population that is particularly at risk

of food insecurity. Using the food insecurity risk indicator I constructed, I classify households with a risk indicator value above the median to be more at risk of food insecure, meaning that these households have qualities (income, race, food stamp/SNAP status) that are associated with an increased risk of food insecurity and thus are more likely to be affected by the initiative. In addition to my original specification, I add a full set of two-way interactions with the food insecurity indicator $AtRisk$ (which $= 1$ for households more at risk, $= 0$ otherwise) and a triple interaction term. Primarily used as a robustness check by conditioning on the subset of households most likely to be at risk of food insecurity, the following model strengthens the estimates from equation (1) (the preferred specification in this paper):

$$Y_{ist} = \gamma_1 PA_S + \gamma_2 POST_t + \gamma_3 AtRisk_i + \lambda_1(PA_s \, x \, POST_t) + \lambda_2(PA_s \, x \, AtRisk_i) + \lambda_3(POST_t \, x \, AtRisk_i) + \beta(PA_s \, x \, POST_t \, x \, AtRisk_i) + \delta_t + \delta_{st} + X_i\alpha + \epsilon_{ist} \quad (2)$$

β, the parameter of interest, measures the effect of the initiative on PA households predicted to be more at risk of food insecurity. δ_{st} are state-year fixed effects, which control for state and time-specific trends across all households.

5 Results and Discussion

5.1 Results

Table 2 shows the estimates from the preferred difference-in-differences model. Column (6) reveals that the addition of controls for employment status has no significant effect on the estimates; thus, I focus on column (5), which presents the results of equation (1). In this specification, I find that the implementation of the 2004 PA FFFI is associated with a 2.6 percentage point decrease in the probability of being food insecure, with this result being statistically significant at the 1% level.

Table 3 analyzes the effect of the 2004 PA FFFI on only the binary *food insecure* variable and only the binary *very food insecure* variable. Estimates for the former show that the implementation of the policy is associated with a 2.2 percentage point decrease in the probability of being *food insecure*, with this result being significant at the 1% level. For the latter, estimates show that the implementation of the policy is associated with a 0.3 percentage point decrease in the probability of being *very food insecure*; this result is not significant at the 1% level. These results suggest that the initiative was not effective for the most food insecure households; rather, the effectiveness of the policy was primarily driven by households which have some level of food insecurity (*food insecure*), but not the most severe.

5.2 Alternative Specifications and Robustness Checks

In the following section, I present a series of robustness checks by evaluating the effect of the initiative on subpopulations which may be correlated with higher food insecurity and on households predicted to be more at risk of food insecurity. If my model is indeed valid, the estimates from these sub-analyses are expected to produce larger coefficients, indicating an even larger decrease in food insecurity levels among the most vulnerable groups as a result of the initiative.

1. Zooming into Specific Subpopulations. Table 4, which presents estimates of the effect of the initiative conditional on being in a certain subpopulation, suggests that the initiative had a significantly larger effect on households below the 185% poverty level, which is expected as poverty status is plausibly a key predictor of food insecurity. Interestingly, however, the initiative does not seem to have a significant effect on neither the black nor the unemployed. Additional analysis regressing only on households that are *food insecure* however (as opposed to the *overall food insecurity* variable in the preferred specification) finds large and significant effects of the initiative on both of these subpopulations; these results support the findings from Table 3, which show that the policy decreased food insecurity for households which experienced some levels of food insecurity, but not for households with the most severe conditions of food insecurity.

2. Triple Difference Specification. To assess whether the policy had a disproportionate effect on households most likely to be food insecure, I run the triple difference specification in equation (2); Table 6 presents regression estimates from this specification. Column (5) shows that the probability of the initiative reducing food insecurity is 5.1 percentage points more among households which are predicted to be more at risk of food insecurity; this result is statistically significant at the 1% level. Since a potential concern is state and time specific shocks, I include state-by-year fixed effects in column (6) to acknowledge that these potential shocks may be correlated with both the regressors and *overall food insecurity*. The inclusion of these controls does not change my point estimates, suggesting that state-year shocks do not bias my estimates for the probability of a more at-risk household being food insecure.

5.3 Discussion

In the linear probability model estimated by equation (1), the coefficient estimate βis interpreted as the change in probability that a household is food insecure with a unit change in $PA_s \, x \, POST_s$, which represents being in the treatment state, Pennsylvania (PA), during the treatment period, 2005-2010. To assess the magnitude of the expected reduction in the number of households which were food insecure, I use data on the number of households in PA from the Decennial Census of Population and Housing and predict that as a result of the 2004 PA FFFI, the expected decrease in the number of households being food insecure was 127,347 households. Using the confidence interval for the coefficient estimate, I am 95% confident that the expected decrease in the number of food insecure households in PA was between 88,163 and 166,530 households. To further back out the expected decrease in the number of individuals who were food insecure, I use data on the average household size in PA from the same Decennial Census and predict that, as a result of the initiative, the expected decrease in the number of food insecure individuals was 313,274 individuals; I am 95% confident that the expected decrease in the number of food insecure individuals in PA was between 216,880 and 404,663 individuals[3].

[3] See Appendix Section C for additional details regarding how the magnitudes of the results are backed out and quantified.

Since Figure 1 graphically shows that the trends for food insecurity in both NJ and PA prior to the implementation of this initiative were fairly stable, I can reasonably interpret my estimates as an absolute reduction in the number of individuals who were food insecure in Pennsylvania as a result of the policy. These estimates are also consistent with data from The Food Trust, which reports that the 2004 PA FFFI improved food access to roughly 400,000 Pennsylvanians. Potential channels through which the initiative could have decreased the total number of food insecure individuals include supermarkets' purchase of equipment to increase the availability of fresh foods in underdeveloped communities and economic revitalization to provide low-income households increased means to afford balanced meals. My findings and these potential explanations are consistent with Stack's 2015 paper "Examining Relationship Between Food Deserts and Health," which shows that socioeconomic status and the travel distance required to purchase fresh produce are correlated with eating a healthy, balanced diet.

Although my paper shows that the 2004 PA FFFI had a significant effect in reducing food insecurity, with an even larger effect after conditioning on households being below the 185% poverty line, my estimates in Table 3 also indicate that the initiative was not effective in reducing food insecurity among the most food insecure households. One potential explanation is that the effect of the initiative was limited such that despite increasing the availability of fresh foods, the most food insecure households were still unable to afford them. This view is consistent with a Southeastern Pennsylvania Survey, which finds that despite access to fresh groceries, many food insecure households still reduced food intake or skipped meals as a result of financial constraints (Mayer et al., 2014). An alternative explanation expands on findings from the paper "Food Deserts and the Causes of Nutritional Inequality," which found that supply-side policies that help out grocery stores in low-income areas have limited effects on healthy eating in those communities due to their lack of demand for healthy groceries (Allcott et al., 2017). The paper finds that education is strongly correlated with demand for healthy groceries, and since low-income is plausibly correlated with both food insecurity and education, it may be the case that the households categorized to be the most food insecure also have the least demand for fresh food.

6 Conclusion

This paper examines the causal impact of the 2004 Pennsylvania Healthy Food Financing Initiative (PA FFFI) on food insecurity levels. Motivated by the persistence of nutritional inequality and the phenomenon of supermarket redlining, I aim to understand the effectiveness of statewide policies in knocking down barriers to fresh food access in disadvantaged communities. Using data from the CPS Food Security Supplement and employing a differences-in-differences model for my identification strategy, I find that the implementation of the 2004 PA FFFI is associated with a 2.6 percentage point decrease in the probability of being food insecure, translating to an expected 313,274 decrease in the number of individuals being food insecure. To strengthen these estimates, I employ a triple-difference specification and find that the probability of the initiative reducing food insecurity is 5.1 percentage points more among households which are predicted to be at risk of food insecurity, suggesting a disproportionate effect of the policy on households most likely to be food insecure at baseline.

The results of this paper have several implications for future policy. Primarily, the results suggest the importance of supply-side policies in decreasing food insecurity for households which experience some degree of food insecurity. Analysis of specific grocery stores which have received funding implies that the impact of the statewide food financing policy may have been driven by both investment in supermarket development/equipment and the expansion of jobs in their respective communities. In contrast, these same supply-side policies do not have an impact on the most food insecure households, suggesting both limitations of supermarket food financing in improving conditions for the worse-off households and the potential merits of providing need-based subsidies for only healthy groceries.

While the results are robust to conditioning on households below the 185% poverty level and corroborated by the triple difference-in-differences model, the insignificance of the policy on households which were *very food insecure* biases my estimates to reveal no impact of the policy on *overall food insecurity* for predominantly black or unemployed households. Moreover, one of the assumptions of the

empirical strategy is that the impact of the initiative was captured within a specified time frame within implementation. Although I was able to evaluate specific case studies in Pennsylvania, comprehensive data on the timing of effects and distribution of finances remain scarce. In future research, it would be valuable to investigate why statewide financing initiatives do not appear to have an effect on households with the most severe cases of food insecurity. Further analysis regarding the timing of the impact of food financing initiatives and channels through which these initiatives are impactful would be beneficial in informing future policy as well.

References

"2010 Census: Pennsylvania Profile." *US Census Bureau,* https://www2.census.gov/geo/pdfs/reference/guidestloc/42_Pennsylvania.pdf.

Abadie, A., Athey, S., Imbens, G.W. and Wooldridge, J.M. (2020), Sampling‐Based versus Design‐Based Uncertainty in Regression Analysis. Econometrica, 88: 265-296. https://doi.org/10.3982/ECTA12675

Alberto Abadie, Matthew M. Chingos, Martin R. West; Endogenous Stratification in Randomized Experiments. The Review of Economics and Statistics 2018; 100 (4): 567–580. doi: https://doi.org/10.1162/rest_a_00732

Arguinzoni, Jennifer and Marie Lawrence. "State Legislation on Hunger" *National Conference of State Legislatures,* January 2012, https://www.ncsl.org/research/human-services/state-legislation-on-hunger.aspx.

Chrisinger, Benjamin W. "Taking Stock of New Supermarkets in Food Deserts: Patterns in Development, *Financing, and Health Promotion." Working paper (Center for Community Development Investments) vol. 2016 (2016): 4.*

Coleman-Jensen, Alisha et al."Measurement." *United States Department of Agriculture Economic Research Service - Measurement*, December 2020, https://www.ers.usda.gov/topics/food-nutrition-assistance/food-security-in-the-us/measurement.aspx.

Eisenhauer, E. In poor health: Supermarket redlining and urban nutrition. GeoJournal 53, 125–133 (2001). https://doi.org/10.1023/A:1015772503007

Goldstein, Ira. "CDFI Financing of Supermarkets in Underserved Communities: A Case Study." *The Reinvestment Fund*, August 2008, https://www.reinvestment.com/wp-content/uploads/2015/12/CDFI_Financing_of_Supermarkets_in_Underserved_Communities_A_Case_Study-Report_2008.pdf.

"Healthy Food Access in Pennsylvania." *The Food Trust,* 2015, http://thefoodtrust.org/uploads/media_items/pabifoldfinal.original.pdf.

"History & Background." *USDA ERS - History & Background*, www.ers.usda.gov/topics/food-nutrition-assistance/food-security-in-the-us/history-background/.

Hunt Allcott, Rebecca Diamond, Jean-Pierre Dubé, Jessie Handbury, Ilya Rahkovsky, Molly Schnell, Food Deserts and the Causes of Nutritional Inequality, The Quarterly Journal of Economics, Volume 134, Issue 4, November 2019, Pages 1793–1844, https://doi.org/10.1093/qje/qjz015

"Intervention: Pennsylvania Fresh Food Financing Initiative." *Center TRT*, http://centertrt.org/content/docs/Intervention_Documents/Intervention_Templates/PA_FFFI_template.pdf.

Mayer, Victoria L et al. "Food insecurity, neighborhood food access, and food assistance in Philadelphia." Journal of urban health : bulletin of the New York Academy of Medicine vol. 91,6 (2014): 1087-97. doi:10.1007/s11524-014-9887-2

Odoms-Young, Angela, and Marino A Bruce. "Examining the Impact of Structural Racism on Food Insecurity: Implications for Addressing Racial/Ethnic Disparities." Family & community health vol.

41 Suppl 2 Suppl, Food Insecurity and Obesity, Suppl 2 FOOD INSECURITY AND OBESITY (2018): S3-S6.

"Pennsylvania: Census 2000 Profile." *US Census Bureau,* August 2002, https://www.census.gov/prod/2002pubs/c2kprof00-pa.pdf.

"Pennsylvania Fresh Food Financing Initiative." *The Reinvestment Fund*, consumerfed.org/wp-content/uploads/2011/10/Evans_Food_Deserts_panel_FPC_2011.pdf.

Peter Berck & Sofia B. Villas-Boas (2015): A note on the triple difference in economic models, Applied Economics Letters, DOI: 10.1080/13504851.2015.1068912

Robbins, Jessica M et al. "Prevalence, disparities, and trends in obesity and severe obesity among students in the Philadelphia, Pennsylvania, school district, 2006-2010." Preventing chronic disease vol. 9 (2012): E145. doi:10.5888/pcd9.120118

Stack, Rebecca L. (2015) "Examining Relationship Between Food Deserts and Health," *SPACE: Student Perspectives About Civic Engagement*: Vol. 1 : Iss. 1, Article 4. https://digitalcommons.nl.edu/space/vol1/iss1/4.

"The Healthy Food Financing Initiative." *PolicyLink,* 14 April 2015, https://www.frbsf.org/community-development/files/healthy_food_financing_initiative.pdf

"The Success of HFFI." The Food Trust, thefoodtrust.org/administrative/hffi-impacts/the-success-of-hffi.

United States. Bureau of the Census, and United States Department of Labor. Bureau of Labor Statistics. Current Population Survey, September 2000: Food Security Supplement. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2011-09-09. https://doi.org/10.3886/ICPSR03908.v2

Walker, Renee E et al. "Disparities and access to healthy food in the United States: A review of food deserts literature." Health & place vol. 16,5 (2010): 876-84. doi:10.1016/j.healthplace.2010.04.013

Wight, Vanessa et al. "Understanding the Link between Poverty and Food Insecurity among Children: Does the Definition of Poverty Matter?." Journal of children & poverty vol. 20,1 (2014): 1-20. doi:10.1080/10796126.2014.891973

Zhang, M. and Ghosh, D. (2016), Spatial Supermarket Redlining and Neighborhood Vulnerability: A Case Study of Hartford, Connecticut. Transactions in GIS, 20: 79-100. https://doi.org/10.1111/tgis.1214

Tables and Figures

Table 1: Differences in Means Test (NJ vs PA)

	2002			2003		
	NJ	PA	t	NJ	PA	t
Food Security Status						
Food Secure	0.92	0.90	-3.07	0.93	0.92	-2.67
Food Insecure	0.05	0.07	3.81	0.04	0.06	3.81
Very Food Insecure	0.03	0.03	-0.19	0.03	0.02	-0.81
Annual Family Income	61044.96	50583.93	-14.71	73158.87	60251.02	-11.22
Race						
White	0.83	0.90	7.63	0.84	0.91	8.79
Black	0.11	0.08	-4.68	0.11	0.08	-4.94
American Indian	0.008	0.001	-4.18	0.003	0.0002	-2.89
Asian	0.05	0.02	-5.01	0.04	0.01	-7.5
Employment Status						
Employed	0.48	0.48	-0.5	0.49	0.48	-0.98
Unemployed	0.03	0.03	0.9	0.02	0.02	-0.11
Not in Labor Force	0.26	0.29	2.17	0.26	0.28	2.14
Food Stamps/SNAP Beneficiary	0.04	0.05	3.16	0.03	0.05	5.09
Below 185% Poverty Level	0.18	0.25	7.19	0.18	0.23	5.09

Notes: Test of equality of means in New Jersey and Pennsylvania among key variables in our dataset. 2002 and 2003 are the years before the 2004 PA FFFI was implemented, with this balance test table presenting the t-statistics for each balance test.

Table 2: Effect of 2004 Pennsylvania Fresh Food Financing Initiative (FFFI) on Food Insecurity

	(1)	(2)	(3)	(4)	(5)	(6)
PA x 2005 − 2010	-0.009**	-0.009**	-0.011***	-0.016***	-0.026***	-0.026***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
White	—	—	0.064	0.060	0.045	0.044
			(0.002)	(0.009)	(0.014)	(0.014)
Black	—	—	0.163	0.115	0.082	0.079
			(0.005)	(0.010)	(0.015)	(0.015)
American Indian	—	—	0.162	0.0148	0.120	0.119
			(0.031)	(0.031)	(0.035)	(0.035)
Asian	—	—	0.038	0.041	0.033	0.031
			(0.004)	(0.009)	(0.015)	(0.015)
Food Stamps/SNAP Beneficiary	—	—	—	0.333	0.236	0.232
				(0.008)	(0.009)	(0.009)
Annual Family Income	—	—	—	—	-5.05e-07	-5.67e-07
					(2.48e-08)	(2.56e-08)
Below 185% Poverty Level	—	—	—	—	0.109	0.109
					(0.004)	(0.004)
Employed	—	—	—	—	—	-0.015
						(0.003)
Unemployed	—	—	—	—	—	0.042
						(0.009)
Not in Labor Force	—	—	—	—	—	-0.038
						(0.003)
Year FE	No	Yes	Yes	Yes	Yes	Yes
R^2	0.001	0.004	0.016	0.092	0.141	0.145
Observations	68,654	68,654	68,654	68,654	58,722	58,722

Notes: This table reports regression DD estimates of the effect of the 2004 PA FFFI on the probability of being food insecure (among households without missing values for food insecurity status). The table shows the percentage point effect in the probability of being food insecure. Estimates in column (2) include controls for year fixed effects, (3) include additional controls for race, (4) for food stamp/SNAP status, (5) for annual family income/poverty status, and (6) for employment status. Standard errors are reported in parentheses.

Table 3: Effect of 2004 Pennsylvania FFFI on Food Insecure vs Very Food Insecure

	(1)	(2)	(3)
	Food Insecure OR Very Food Insecure	Food Insecure	Very Food Insecure
PA x 2005 − 2010	-0.026***	-0.022***	-0.003
	(0.004)	(0.003)	(0.002)
Controls	Yes	Yes	Yes
Year FE	Yes	Yes	Yes
R^2	0.141	0.084	0.061
Observations	58,722	58,722	58,722

Notes: This table reports regression DD estimates of the effect of the implementation of the 2004 PA FFFI on the probability of being food insecure OR very food insecure, only food insecure, and only very food insecure (among households without missing values for food insecurity status). The table shows the percentage point effect in the probability of being either food insecure OR very food insecure in column (1), food insecure in column (2), and very food insecure in column (3). Standard errors are reported in parentheses.

Table 4: Effect of 2004 Pennsylvania FFFI on Food Insecurity (Sub-population Analysis)

	(1)	(2)	(3)	(4)	(5)
	All	Below 185% Poverty Level	Black	Food Stamp/SNAP Beneficiary	Unemployed
PA x 2005 − 2010	-0.026***	-0.062***	-0.013	-0.026	-0.029
	(0.004)	(0.014)	(0.020)	(0.041)	(0.037)
Controls	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
R^2	0.141	0.069	0.143	0.038	0.105
Observations	58,722	15,666	4,590	3.098	1,613

Notes: This table reports regression DD estimates of the implementation of the 2004 PA FFFI on the probability of being food insecure (among households without missing values for food insecurity status). The table shows the percentage point effect in the probability of being food insecure after controlling for year fixed effects. Columns (2), (3), (4), and (5) show the zoomed in effect of the PA FFFI on different subpopulations within my sample. Standard errors are reported in parentheses.

Table 5: Regression of Risk Indicator on Food Insecurity Variable

	Food Insecurity Dummy (PA before 2005)
Food Insecurity Indicator	0.163***
	(0.003)
Constant	0.085
R^2	0.169
Observations	18,020

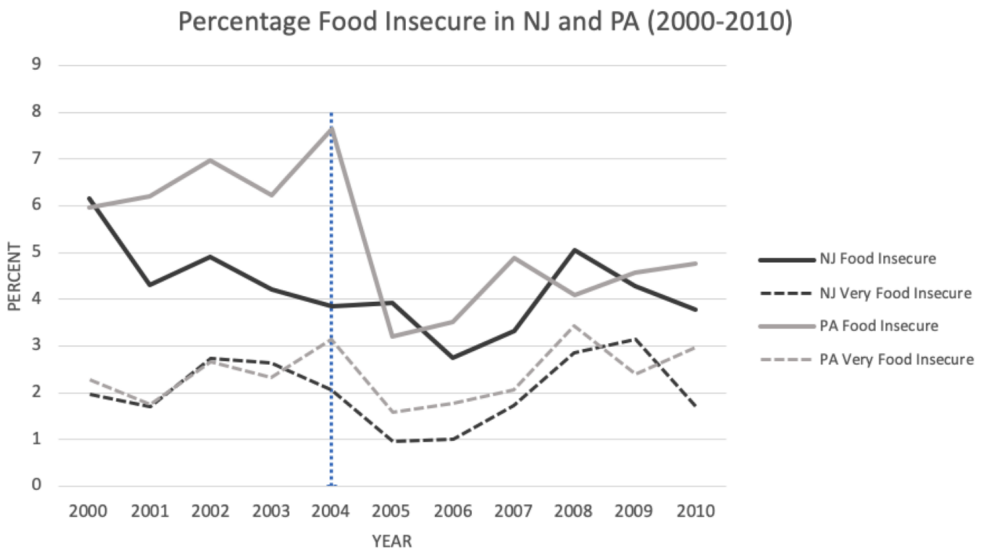
Notes: This table shows the regression of the food insecurity risk indicator on overall food insecurity (looking at households in Pennsylvania before 2005). The coefficient is positive and statistically significant at the 1% level, indicating that being either food insecure or very food insecure is associated with a 16.3 percentage point increase in being predicted to be at risk of food insecurity. In addition to the presented summary statistics, the regressions presented here corroborates the quality of the food insecurity indicator.

Table 6: Effect of 2004 Pennsylvania FFFI on Food Insecurity (Triple DiD Specification)

	(1)	(2)	(3)	(4)	(5)	(6)
PA x 2005 – 2010 x At Risk	-0.045***	-0.043***	-0.045***	-0.055***	-0.051***	-0.051***
	(0.010)	(0.011)	(0.009)	(0.010)	(0.010)	(0.011)
White	—	—	0.029	0.036	0.047	0.050
			(0.009)	(0.012)	(0.015)	(0.015)
Black	—	—	0.092	0.067	0.084	0.087
			(0.011)	(0.014)	(0.016)	(0.016)
American Indian	—	—	0.069	0.080	0.114	0.120
			(0.037)	(0.037)	(0.036)	(0.036)
Asian	—	—	0.031	0.038	0.037	0.039
			(0.010)	(0.013)	(0.015)	(0.015)
Food Stamps/SNAP Beneficiary	—	—	—	0.290	0.238	0.238
				(0.009)	(0.010)	(0.010)
Annual Family Income	—	—	—	—	-4.78e-07	-4.81e-07
					(3.23e-08)	(3.24e-08)
Below 185% Poverty Level	—	—	—	—	0.108	0.108
					(0.004)	(0.004)
Year FE	No	Yes	Yes	Yes	Yes	Yes
Year-by-State FE	No	No	No	No	No	Yes
R^2	0.056	0.060	0.063	0.119	0.142	0.143
Observations	53,958	53,958	53,958	53,958	53,958	53,958

Notes: This table reports regression estimates for the triple diff-in-diff specification, looking at the effect of the 2004 PA FFFI on the probability of being food insecure among households more at risk of being food insecure. The data includes all households without missing values for food insecurity status in the CPS Food Security Supplement from 2000-2010. The table shows the percentage point effect in probability of being food insecure for households more at risk of being food insecure. Estimates in column (2) include controls for year fixed effects, (3) include additional controls for race, (4) for food stamp/SNAP status, (5) for annual family income/poverty status, and (6) for year-by-state fixed effects. Standard errors are reported in parentheses.

Figure 1: Food Insecurity Levels in NJ and PA from 2000-2010



Notes: From the CPS Food Security Supplement 2000-2010. The graph shows the change in food insecurity levels (distinguished between households that are categorized as food insecure and very food insecure, which is the most severe level of food insecurity) between respondents in NJ vs in PA based on responses to specific questions on the survey. The blue dotted line indicates the year that the 2004 Pennsylvania Fresh Food Financing Initiative was implemented.

Appendix

A Data Construction and Background

To construct the master dataset used for my analysis, I use the Current Population Survey: Food Security Supplement (CPS-FSS) for every year from 2000 to 2010. As a first step, I identify fourteen variables to keep for the purpose of my analysis. To combine the datasets into one large dataset for my analysis, I standardize these variables across all years. Across these 11 different datasets, some of the variables are inconsistent over the years; for example, starting from the 2003 CPS-FSS, the race variable was renamed and also changed to include mixed races. In this regard, I rename variables across all datasets to match, and I also re-code variables, such as race, such that the values of each of the variables corresponded to the same label across all years. Next, since the income variable in particular is coded as different numerical values corresponding to specific income ranges, I re-code each income range to be the average income value of the range such that the income variable can be interpreted in dollars.

After making these sample restrictions and re-coding variables for consistency across each CPS-FSS from 2000-2010, I combine every year's dataset into one master dataset. There are 68,654 observations in the master dataset, with ~6,000-7,000 observations for each year. In Appendix Tables A1 and A2, I display NJ and PA summary statistics for 2002 (prior to the 2004 PA FFFI) and 2008 (post implementation of the 2004 PA FFFI). Appendix Figures A1 and A2 display a histogram of the proportion of *food insecure* and *very food insecure* in NJ and PA in both 2002 and 2008, respectively. 2008 was chosen as a representative year since as of July 2008, over $25M have already been dispersed in grants and loans. Appendix Table A3 presents a detailed breakdown of the loans and grants requested, approved, and disbursed as of July 2008.

As additional information regarding supermarkets and grocery stores which received financing through FFFI, stores approved for FFFI financing in urban areas ranged from 17,000 to 65,000 square feet, with full-service supermarkets employing 150-200 employees; stores in rural areas were mainly

family-owned businesses ranging from 12,000 to 22,000 square feet and employing 10-84 employees. The FFFI has supported supermarkets in the following counties in PA: Adams, Allegheny, Armstrong, Beaver, Blair, Bradford, Berks, Bucks, Cambria, Carbon, Chester, Columbia, Dauphin, Delaware, Lackawanna, Lancaster, Lebanon, Lehigh, Luzerne, Northumberland, Philadelphia, Somerset, Schuylkill, Tioga, Washington, Westmoreland, and York.

B Food Insecurity Risk Indicator

To zoom into populations more at risk of food insecurity, and thus would have been more impacted by the program, I construct a food insecurity risk indicator to capture whether or not a household has qualities that make it more at risk of being food insecure. Based on the paper "Endogenous Stratification in Randomized Experiments," the process to construct the indicator is as follows (Abadie et al, 2013). First, I take a random sample of 20% of the NJ households, which is the control group in my data. Within my sample, I regress the *overall food insecurity* (main outcome) binary variable on a set of baseline characteristics, including race, income, unemployment, food stamp status. Using the estimation results from this regression, I predict food insecurity risk for the remaining households in my dataset, which include both NJ and PA households. Lastly, I split the households at the median, classifying households with a food insecurity risk indicator above the median to be more at risk of food insecurity.

Appendix Table B1 shows summary statistics for the food insecurity risk indicator, as well as the summary statistics for the indicator conditional on the household being categorized as *food insecure*, *very food insecure*, or *food secure* by the CPS-FSS. The summary statistics show that households that are *very food insecure* have the highest mean food insecurity risk indicator value of 0.264, indicating the highest predicted level of food insecurity. Households classified as *food insecure* have a mean food insecurity risk indicator value of 0.226 and households classified as *food secure* have a mean food insecurity risk indicator value of 0.075. The results from our summary statistics support the validity of the food insecurity risk indicator.

C Quantifying the Results

To get a sense of the magnitudes associated with the regression estimates and back out a value for the expected decrease in the number of households and individuals who were food insecure after the policy was passed, I use data from the Decennial Census of Population and Housing (US Census Bureau). Since the treatment period is 2005-2010, data on the number of households in PA in 2005 would have been preferable. However, since the Decennial Census is only available for 2000 and 2010, I estimate the number of households in PA in 2005 by taking the average of the number of households in PA in 2000 and 2010, which were 4,777,003 and 5,018,904 respectively, to estimate the number of households in PA in 2005 to be 4,897,953. The estimate from my preferred specification for the decrease in the probability of being food insecure after the policy was passed is -0.026 (2.6 percentage points), with a 95% confidence interval of [-0.034, -0.018]. Multiplying the coefficient estimate, as well as the 95% confidence interval, by the estimated number of PA households in 2005 yields the expected decrease in the number of food insecure households in PA to be 127,347, with 95% confidence that the expected decrease in the number of food insecure households was between 88,163 and 166,530 households.

To quantify the decrease in the number of individuals who were food insecure after the initiative was implemented, I use data from the same Decennial Census for the average household size in PA. Similarly, data on the average household size in PA in 2005 would have been preferable. However, since the Decennial Census is only available for 2000 and 2010, I estimate the average household size in PA in 2005 by taking the mean of the average household size in PA in 2000 and in 2010, which were 2.44 and 2.48 respectively, to estimate the average PA household size in 2005 to be 2.46 individuals. Multiplying 2.46 by the expected decrease in the number of food insecure households, as well its 95% confidence interval, yields the expected decrease in the number of food insecure individuals in PA to be 313,247, with 95% confidence that the expected decrease in the number of food insecure individuals was between 216,880 and 404,663 individuals.

Appendix: Tables and Figures

Table A1: 2002 Summary Statistics (NJ vs PA)

	NJ (2002)					PA (2002)				
	Mean	SD	Min	Max	Observations	Mean	SD	Min	Max	Observations
Food Security Status										
Food Secure	0.92	0.27	0	1	3035	0.90	0.30	0	1	4585
Food Insecure	0.05	0.22	0	1	3035	0.07	0.25	0	1	4585
Very Food Insecure	0.03	0.16	0	1	3035	0.03	0.16	0	1	4585
Annual Family Income	61044.96	27759.02	2500	87500	2597	50583.93	28141.91	2500	87500	3789
Race										
White	0.83	0.37	0	1	3035	0.90	0.31	0	1	4585
Black	0.11	0.32	0	1	3035	0.08	0.27	0	1	4585
American Indian	0.01	0.09	0	1	3035	0.00	0.03	0	1	4585
Asian	0.05	0.21	0	1	3035	0.02	0.15	0	1	4585
Employment Status										
Employed	0.48	0.50	0	1	3035	0.48	0.50	0	1	4585
Unemployed	0.03	0.16	0	1	3035	0.03	0.17	0	1	4585
Not in Labor Force	0.26	0.44	0	1	3035	0.29	0.45	0	1	4585
Food Stamps/SNAP Beneficiary	0.04	0.18	0	1	3035	0.05	0.22	0	1	4585
Below 185% Poverty Level	0.18	0.39	0	1	3035	0.25	0.43	0	1	4585

Notes: This table shows summary statistics for main outcome variable — food security status — as well as other variables that are known to be correlated with food security. All variables with a minimum value of 0 and maximum value of 1 are encoded as binary variables such that the mean and standard deviation can be interpreted as percentages. Since family income (HUFAMINC in the original CPS-FSS dataset) is coded as different numerical values corresponding to specific income ranges, "Annual Family Income" has been re-coded to take on the average income of each income range and can thus be interpreted as dollars.

Table A2: 2008 Summary Statistics (NJ vs PA)

	NJ (2008)					PA (2008)				
	Mean	SD	Min	Max	Observations	Mean	SD	Min	Max	Observations
Food Security Status										
Food Secure	0.92	0.27	0	1	2314	0.92	0.26	0	1	3292
Food Insecure	0.05	0.22	0	1	2314	0.04	0.20	0	1	3292
Very Food Insecure	0.03	0.17	0	1	2314	0.03	0.18	0	1	3292
Annual Family Income	85976.21	46638.22	2500	150000	2018	63474.94	41447.76	2500	150000	2813
Race										
White	0.81	0.39	0	1	2314	0.89	0.31	0	1	3292
Black	0.09	0.28	0	1	2314	0.09	0.29	0	1	3292
American Indian	0.00	0.05	0	1	2314	0.00	0.04	0	1	3292
Asian	0.09	0.29	0	1	2314	0.01	0.12	0	1	3292
Employment Status										
Employed	0.50	0.50	0	1	2314	0.48	0.50	0	1	3292
Unemployed	0.03	0.18	0	1	2314	0.03	0.18	0	1	3292
Not in Labor Force	0.26	0.44	0	1	2314	0.30	0.46	0	1	3292
Food Stamps/SNAP Beneficiary	0.04	0.20	0	1	2314	0.07	0.25	0	1	3292
Below 185% Poverty Level	0.18	0.38	0	1	2314	0.30	0.46	0	1	3292

Notes: This table shows summary statistics for main outcome variable — food security status — as well as other variables that are known to be correlated with food security. All variables with a minimum value of 0 and maximum value of 1 are encoded as binary variables such that the mean and standard deviation can be interpreted as percentages. Since family income (HUFAMINC in the original CPS-FSS dataset) is coded as different numerical values corresponding to specific income ranges, "Annual Family Income" has been re-coded to take on the average income of each income range and can thus be interpreted as dollars. The 2008 dataset also has 18 additional values to represent mixed races, which have been re-coded into one of the four original categories.

Table A3: FFFI Program Summary as of July 31, 2008

FFFI Metric	Philadelphia	Non-Phila	PA Totals	% Non-Phila
Loans Requested	$33,683,909	$44,467,758	$78,151,667	56.9%
Grants Requested	$11,728,060	$29,496,873	$41,224,933	71.6%
Loans Approved	$24,789,760	$9,870,498	$34,660,258	28.5%
Grants Approved	$2,944,660	$6,013,930	$8,958,590	67.1%
Loans Disbursed	$17,556,000	$3,082,500	$20,638,500	14.9%
Grants Disbursed	$3,218,550	$2,192,900	$5,411,450	40.5%
Total Project Costs: Approved Applicants	$103,561,815	$63,646,123	$167,207,938	38.1%
Total Jobs: Approved Applicants	1,980	1,730	3,710	46.6%
Total Square Footage: Approved Applicants	588,748	867,544	1,456,292	59.6%
Total # of Applicants	39	117	156	75.0%
Total # of Eligible Applicants	36	107	143	74.8%
Total # of Applicants Approved for Funding	21	44	65	67.7%

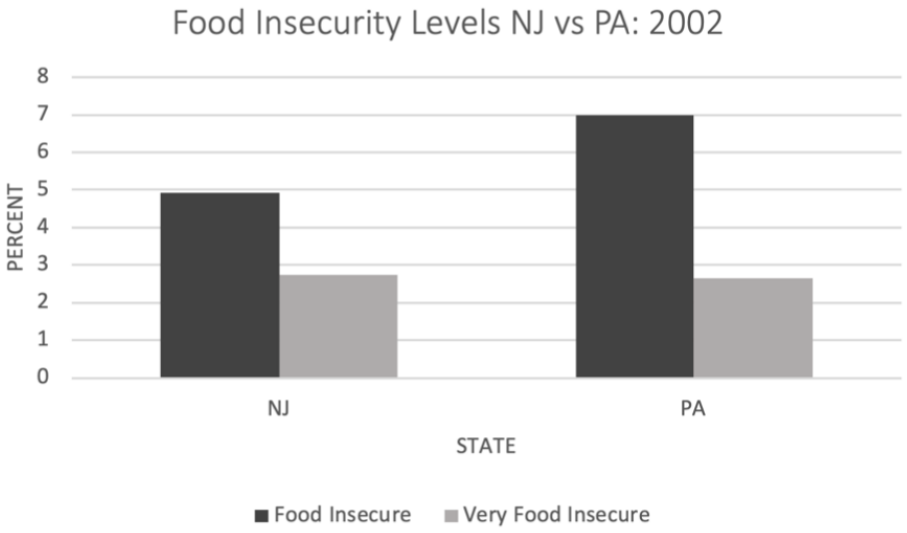
Notes: This table is from the paper "CDFI Financing of Supermarkets in Underserved Communities: A Case Study," published by the Reinvestment Fund in August 2008. This table details the financial figures and metrics for the PA Fresh Food Financing Initiative as of July 31, 2008. .

Table B1: Food Insecurity Risk Indicator Summary Statistics

	(1)	(2)	(3)	(4)	(5)
	Mean	SD	Min	Max	Observations
Food Insecurity Indicator	0.087	0.117	-0.053	0.599	53,958
If Food Insecure	0.226	0.154	-0.038	0.572	2,799
If Very Food Insecure	0.264	0.164	-0.015	0.599	1,318
If Food Secure	0.075	0.104	-0.053	0.582	49,841

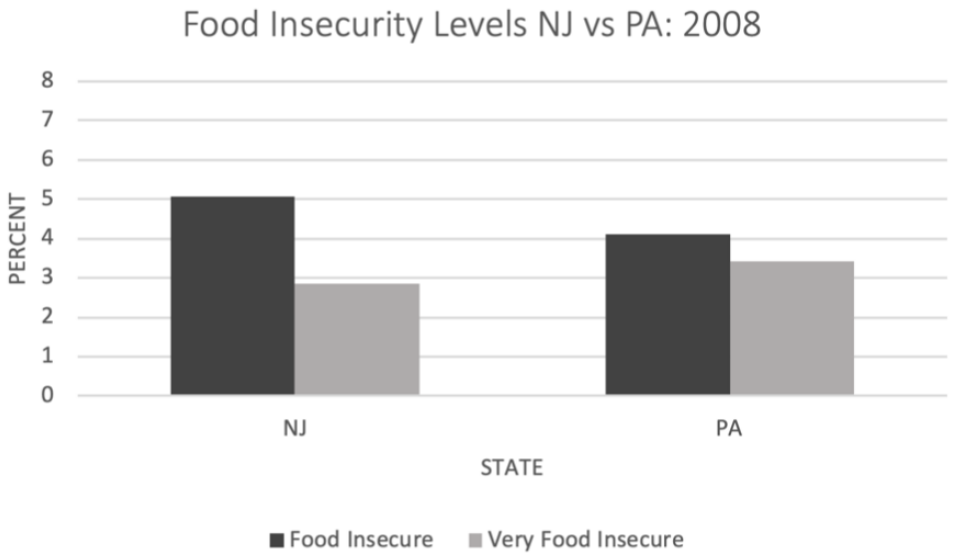
Notes: This table shows summary statistics for the food insecurity risk indicator overall, as well as the summary statistics for the indicator among households that are food insecure, very food insecure, and food secure. This indicator was constructed by regressing on baseline characteristics of a random sample of NJ households and predicting food insecurity risk for the remaining households. Households with indicator values which are above the median indicator value are classified to be more at risk of food insecurity.

Figure A1: Proportion of respondents that are food insecure in NJ and PA in 2002



Notes: December 2002 CPS Food Security Supplement. Compares percent food insecure and very food insecure between a sample of 3,035 respondents in NJ and 4,585 respondents in PA in 2002. The figure shows that the proportion of population that is food insecure is lower in NJ than in PA in 2002, but both states exhibit relatively similar proportions of population which are very food insecure (most severe conditions of food insufficiency).

Figure A2: Proportion of respondents that are food insecure in NJ and PA in 2008



Notes: December 2008 CPS Food Security Supplement. Compares percent food insecure and very food insecure between a sample of 2,314 respondents in NJ and 3,292 respondents in PA in 2008. The figure shows that the proportion of population that is food insecure is lower in PA than in NJ in 2008, while PA exhibits a bit higher proportion of its population which is very food insecure (most severe conditions of food insufficiency).