The Impact of Natural Disasters on Education: Evidence from Standardized Testing

Gregor Steiner

May 20, 2022

Abstract

1 Introduction

Natural disasters are responsible for massive economic damage and due to climate change the frequency of such disasters will increase in most regions (Intergovernmental Panel on Climate Change (IPCC), 2021). Therefore, it is essential to have a good understanding of the consequences. While much research has been done on the macroeconomic consequences of natural disasters, few studies have focused on the impact on education. This study investigates the effect of natural disasters on academic performance.

A causal effect of natural disasters may be driven by school closures (Grewenig et al., 2021) or lowered attendance (Spencer et al., 2016), destroyed infrastructure, and emotional stress (Vogel and Schwabe, 2016). Furthermore, some forms of disasters, e.g. extreme heat, may directly impair cognitive performance (Ramsey, 1995).

To identify dynamic causal effects of experiencing natural disasters, this paper uses an event-study design. In particular, we estimate dynamic treatment effects for up to eight years after initial treatment. As a result, not only short-term but also medium to long-term effects can be found. Since treatment effects are likely very heterogenous, we use the estimator by Sun and Abraham (2021).

This article uses standardized test data on a US county level for grades 3 through 8 in mathematics and reading language arts (RLA) to measure academic performance, covering the school years 2008/2009 to 2018/2019. This measure is very attractive as the test scores are standardized relative to a national reference cohort. Therefore, the outcomes are nationally comparable. For the same period, we obtain data on natural disasters from declarations by the Federal Emergency Management Agency and data on storms from the National Weather Service.

This article contributes to a rich literature on economic effects of natural disasters. More specifically, this paper contributes to the literature on the impact of natural disasters on education. Previous work has produced mixed results. Some authors find significant effects of natural disasters on the education system (Holmes, 2002; Crespo Cuaresma, 2010; Sacerdote, 2012; Park et al., 2020), while others find no or only small effects (Baggerly and Ferretti, 2008; Pane et al., 2008). Morover, some authors find large differences by subject (Spencer et al., 2016).

Previous studies have overwhelmingly concentrated on a single type of disasters, or even a single instance. For example, hurricane Katrina has been the subject of many studies (e.g. Sacerdote, 2012; Deryugina et al., 2018). Contrarily, we exploit a very comprehensive dataset of natural disasters containing various different types of disasters.

We find strong evidence for a negative effect of disasters on performance in mathematics in the same school year. Evidence for medium and long term effects is weak.

The rest of this paper is organized as follows: Section 2 introduces the data used and presents some descriptive statistics. Section 3 explains the empirical strategy. Section 4 discusses the results and section 5 concludes.

2 Data

2.1 Natural Disasters

Natural disasters are declared as such by the president, usually upon request by the affected state's governor. Once a disaster is federally declared, states or local governments can receive federal assistance. The Federal Emergency Management Agency (FEMA) provides data on all federally declared natural disasters, beginning in 1953. The data is easily accessible via their API (Turner, 2022). Figure 1 shows the number of declared disasters between 2009 and 2018 across the US.

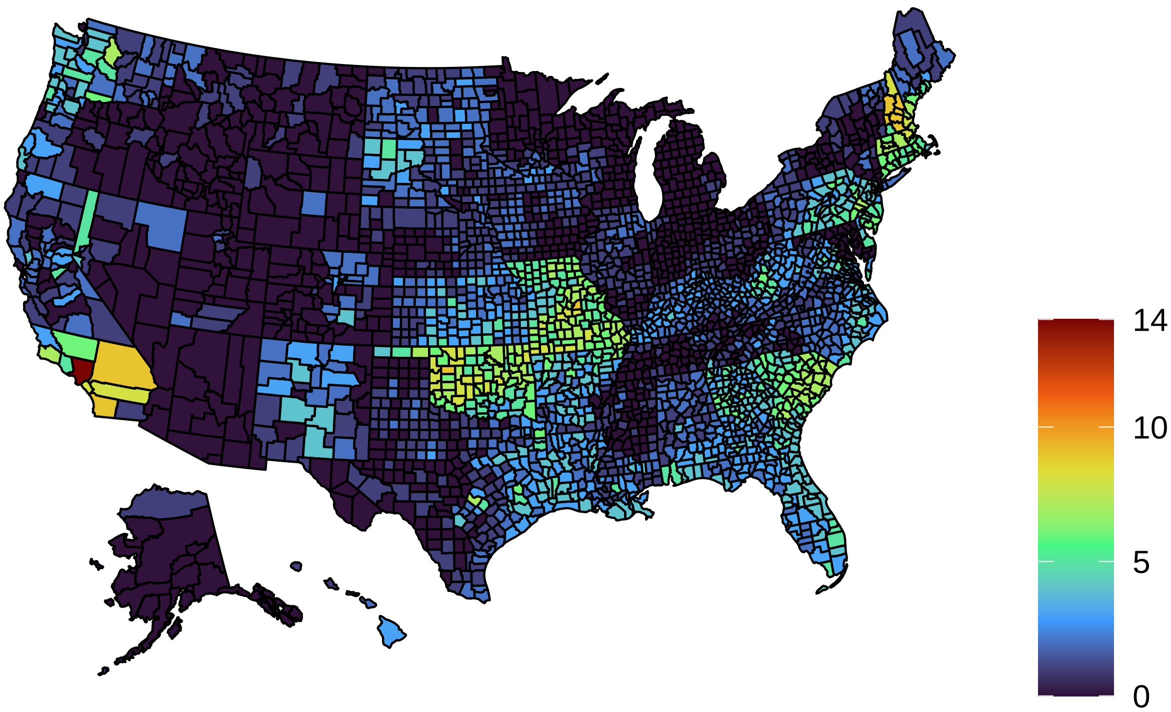

Figure 1: Number of declared natural disasters from 2009 to 2018

Table 1 shows the types of disasters and their proportion in the FEMA data. Storms make up the largest share of disaster events. Fire and floods are also a substantial part.

FEMA also provides a dataset on their Public Assistance Applicants Program Deliveries. This contains information on applicants and their recovery priorities, including the amount of damage caused and amount of federal disaster assistance granted. Unfortunately, this data is only available since October 2016. Figure 2 shows the total federal assistance awarded to counties.

Figure 3 shows boxplots by county application status. Counties that did apply for federal disaster assistance tend to have lower median income, higher poverty rates, and higher shares of single motherhood. Thus, it seems that counties that had to apply for federal disaster assistance were more socially vulnerable in the first place. However, the direction of causality is not clear. Possibly these counties are more vulnerable to natural disasters and are also poorer or more socially vulnerable because of it. Alternatively, counties that are poorer could be more likely to apply for public disaster aid as they have fewer private resources.

It is also interesting whether variation in the federal assistance procedure may be driven by political factors. Visually, the distribution of democratic votes in the 2016 election (almost coincides

Table 1: Disasters from 2009 to 2018 by type

Variable	N	Percent
Disaster Type	13230	
... Chemical	9	0.1%
... Coastal Storm	12	0.1%
... Dam/Levee Break	3	0%
... Earthquake	19	0.1%
... Fire	886	6.7%
... Flood	2006	15.2%
... Freezing	1	0%
... Hurricane	3094	23.4%
... Mud/Landslide	28	0.2%
... Other	7	0.1%
... Severe Ice Storm	803	6.1%
... Severe Storm(s)	5644	42.7%
... Snow	577	4.4%
... Terrorist	4	0%
... Tornado	114	0.9%
... Toxic Substances	1	0%
... Tsunami	9	0.1%
... Typhoon	11	0.1%
... Volcano	2	0%

with the start of the Public Assistance Applicants Program Deliveries dataset) does not seem to be different in the two groups. However, logistic regression results indicate a significant negative relationship between a county's application status and its share of democratic votes in the 2016 election (see Appendix A). While this is not necessarily a causal effect, it could be an indication that a Republican president may be more hesitant awarding disaster assistance to Democratic counties. Similarly, there is a negative relationship between the democratic voter share in the 2008 election and whether a county received any disaster declarations in th 2009-2018 period. However, since the president was a Democrat from 2009-2016, this speaks against partisan declaration of disasters.

Since FEMA declarations may be driven by politics, we repeat the analysis on a second disaster dataset from an alternative source. The National Weather Service (NWS) provides data on storm events. In particular, this covers hurricanes, mainly affecting southern coastal regions, tornadoes, and other severe storms. These make up a very large part of all natural disasters experienced in the United States (see table 1). Combined they account for more than 80% of all disaster damage in the FEMA Public Assistance Applicants Program Deliveries database.

We only consider severe storms which are likely to cause substantial damage. Tornadoes can be classified with the Enhanced Fujita (EF) scale based on estimated peak wind speeds (for more details see Mcdonald et al., 2004). Tornadoes with an EF scale of 0 or 1 (wind speeds of up to 110 mph) are characterized as weak. Therefore, we exclude those and only keep tornadoes with an EF scale of at least 2 (wind speeds of at least 111 mph). Unfortunately, the hurricane data does not include a similar measure, but it does include an estimated amount of property damage. We

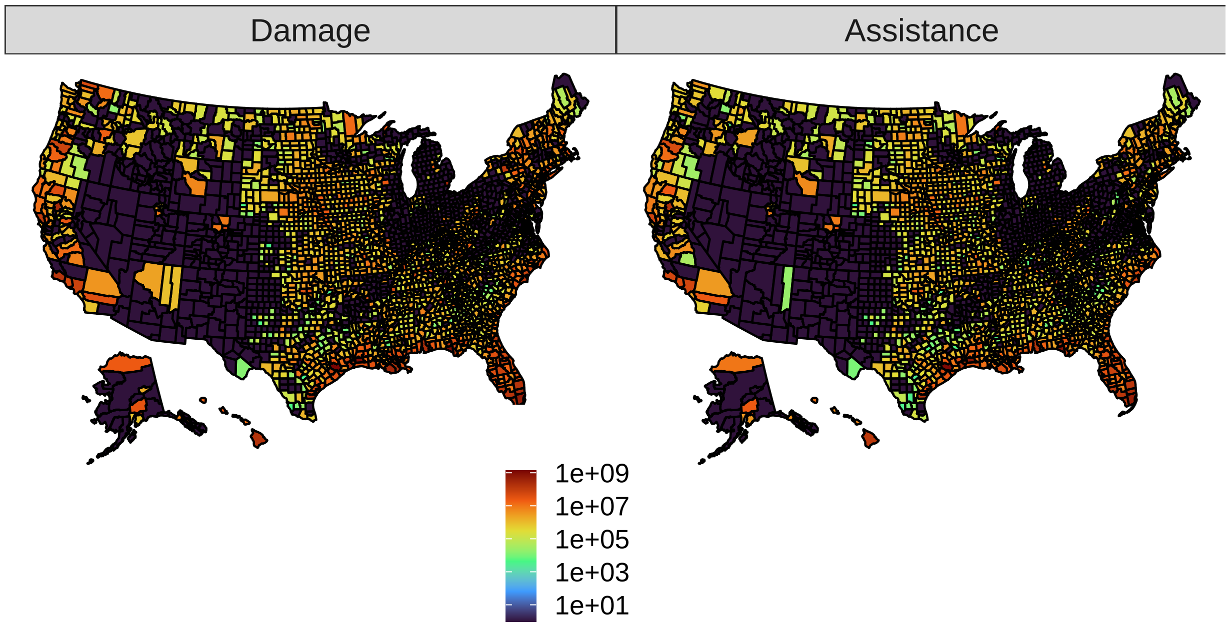

Figure 2: Amount of disaster damage reported by and federal disaster assistance awarded to counties since October 2016 (both in USD)

exclude all hurricanes with an estimated property damage of zero. Storm exposure by county is shown in figure 4.

2.2 Standardized Testing Data

Data on academic achievement is available from the Stanford Education Data Archive (Reardon et al., 2021). They provide mean test results from standardized tests by county, year, grade and subject among all students and various subgroups (including race, gender, and economically disadvantaged). The most recent version 4.1 covers grades 3 through 8 in mathematics and Reading Language Arts (RLA) over the 2008-09 through 2017-18 school years.

Test scores are cohort-standardized, meaning they can be interpreted relatively to an average national reference cohort in the same grade. For instance, a county mean of 0.5 indicates that the average student in the county scored approximately one half of a standard deviation higher than the average national student in the same grade.

In addition to overall mean test scores, the data includes mean test scores for various subgroups, e.g. by ethnicity. In particular, we consider mean test scores for black, hispanic, female, and economically disadvantaged students. These are only reported if the subgroups' sample sizes are large enough. Thus, the number of observations for some of them is substantially smaller.

Furthermore, the Stanford Education Data Archive maintains a large set of covariates for each county and year. They include variables like the county's median income, unemployment and poverty rate.

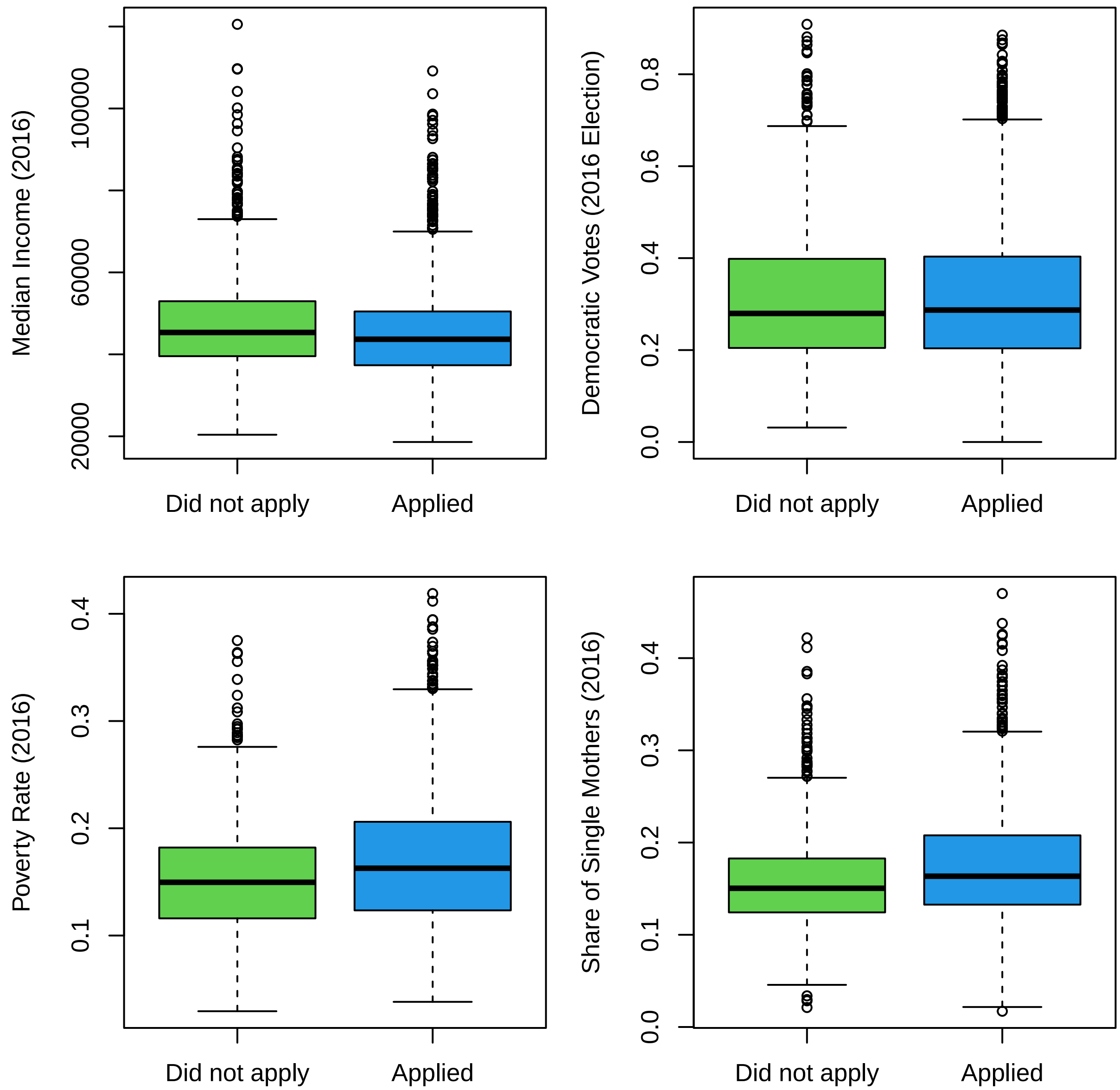

Figure 3: Boxplots by application status

2.3 Combining disaster and testing data

Natural disasters should only have an effect on test scores if they occur before the test. Standardized tests are generally administered during spring. We will use March 1st as a cut-off point. Thus, any disaster happening within the same school year before the 1st of March will be considered. School years tend to start in late August or early September with some variation across states. We will use September 1st, meaning any disaster happening between September 1st and March 1st will be counted for a given school year. Disasters occuring in the summer or in the spring after the exams should have much less influence on performance. Thus, we do not consider disasters that occur between March 1st and September 1st.

Each disaster is assigned to a school year as described above. Then, disaster and test score data

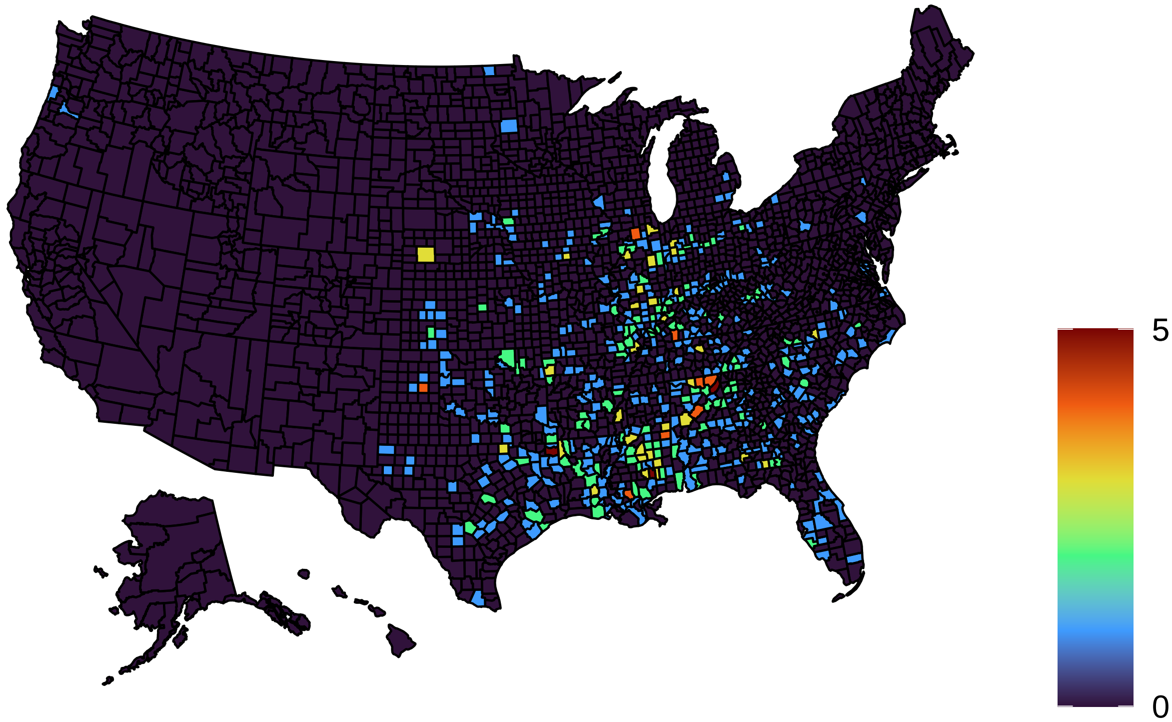

Figure 4: Number of storms from 2009 to 2018

can be merged by school year and county. This yields a panel data set with six grades and two subjects for each county-year combination.

The outcomes of interest are overall mean test scores by county, and mean test scores for black, hispanic, female, and economically disadvantaged students. Figure 5 shows boxplots for the five outcomes of interest. All five mean test scores are measured on the cohort standardized scale, that is a given observation measures the distance in standard deviations from the national reference cohort.

Due to the way the scale is constructed, overall test scores are distributed symmetrically around zero, except for a few outliers. The mean scores for black, hispanic, and economically disadvantaged students are shifted downwards by -0.48, -0.281, and -0.283 standard deviations respectively. Female mean scores are slightly above overall mean scores, meaning that female students perform slightly better than male students on average.

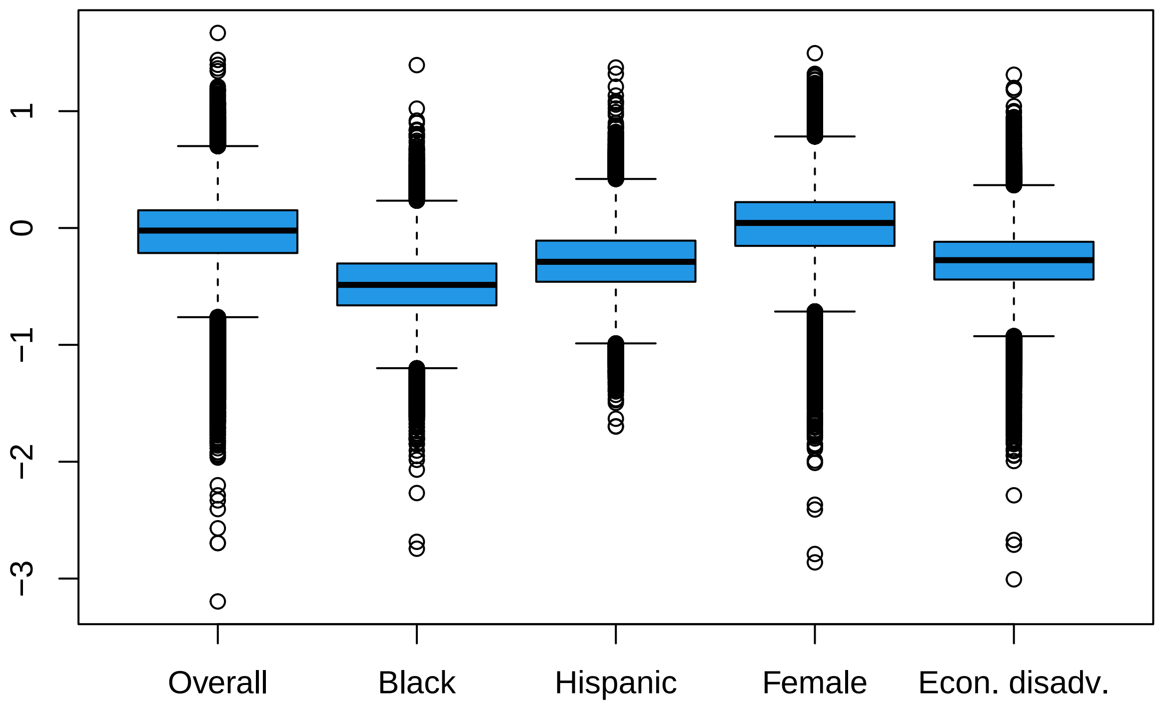

Figure 5: Boxplots of the outcomes of interest

3 Empirical Strategy

3.1 Setting

We employ an event study design to measure the effect of natural disasters on standardized test outcomes. An event study design is a staggered adoption design where units are first-treated at different points in time, and there may or may not be never-treated units (Sun and Abraham, 2021).

Note that treatment must be absorbing, meaning the sequence of treatment indicators $(D_{i,t})_{t=1}^{T}$ must be a non-decreasing sequence of 0s and 1s. In other words, after being treated for the first time a county stays treated. In the present application this means treatment refers to having experienced a disaster rather than experiencing a disaster in that year. This is common practice and does not cause bias due to the conditionally random nature of natural disasters (Deryugina, 2017). Thus, the emphasis lies on cumulative long-term effects rather than instantaneous short-term effects.

In order to identify a causal effect, unobservable determinants of a county's test scores must be unrelated to natural disasters conditional on observable characteristics of that county. The occurrence of natural disasters is plausibly random conditional on location. Furthermore conditioning on the year should account for an increasing trend in natural disasters due to climate change. Thus, independence of mean test scores and natural disasters is plausible conditional on county and year fixed effects.

Consequently, the baseline specification is

$$y_{i,t,g} = \sum_{l=-9, l\neq -1}^{8} \beta_l D_{i,t-l} + \alpha_i + \lambda_t + \zeta_g + \varepsilon_{i,t,g} \ , \qquad (1)$$

where $y_{i,t,g}$ is the outcome of interest for county i, year t, and grade g. County, year, and grade fixed-effects are given by α_i, λ_t, and ζ_g respectively. $D_{i,t-l}$ is a treatment indicator for county i in year $t-l$. That is, $D_{i,t-l} = 1$ if the county had already experienced a disaster l years ago at time t.

Since we consider the time period 2009-2018, $-9 \leq l \leq 9$, but note that $l = 9$ would correspond to a unit that experienced a disaster in the first period and is therefore always treated. As recommended by Sun and Abraham (2021) and Callaway and Sant'Anna (2021), these units are dropped from estimation. Neither can treatment effects be identified for that group nor are they useful as a comparison group under standard parallel trends assumptions.

Also, we need to drop at least two leads or lags to avoid a multicollinearity problem. A complete set of treatment leads and lags is perfectly collinear with unit and time fixed-effects (for an extensive discussion of this issue see Borusyak et al., 2021, section 3.2). It is common to drop the first relative indicator prior to treatment (i.e. $\beta_{-1} = 0$). This acts as a normalization of treatment relative to the period before treatment. Furthermore, we bin the distant leads, that is we combine the treatment indicators for $l \leq -5$. Thus, equation (1) turns into

$$y_{i,t,g} = \beta_{-5} D_{i,t-5} + \sum_{l=-4, l\neq -1}^{8} \beta_l D_{i,t-l} + \alpha_i + \lambda_t + \zeta_g + \varepsilon_{i,t,g} \ , \qquad (2)$$

where $D_{i,t-5}$ indicates treatment for any $l \leq 5$.

It is implausible that the treatment effects are constant in our setting. The extent of disasters varies substantially, and also the level of preparation for such disasters likely displays high variance across counties. Also, some counties may experience additional natural disasters after the first one, while others only experience one. Naturally, we would expect larger treatment effects for the former group.

With heterogenous treatment effects, standard two-way fixed-effects estimators are inadequate (de Chaisemartin and D'Haultfœuille, 2020; de Chaisemartin and D'Haultfoeuille, 2021; Sun and Abraham, 2021). Therefore, we use an alternative estimation procedure by Sun and Abraham (2021).

Treatment adoption varies in time and is therefore assigned in clusters: Counties that are first treated in a given year form a cluster. Following the recommendation by Abadie et al. (2017), standard errors are clustered at the cohort level.

3.2 Interaction-weighted estimator

We utilize the interaction-weighted (IW) estimator proposed by Sun and Abraham (2021) that is robust to treatment effects heterogeneity. The main interest lies on the cohort average treatment effect on the treated (CATT),

$$CATT_{e,l} := \mathbb{E}\left[Y_{i,t+l} - Y^{\infty}_{i,t+l} | E_i = e\right],$$

where $Y^{\infty}_{i,t+l}$ is the counterfactual of being never treated and E_i denotes the first treatment period. Thus, $CATT_{e,l}$ is the average treatment effect on the treated l years after being treated for the first time for the cohort that was first treated in year e.

The estimation procedure consists of three main steps:

1. Estimate $CATT_{e,l}$ using a linear fixed effects specification with interactions between relative period indicators and cohort indicators:

$$y_{i,t,g} = \sum_{e \notin C} \sum_{l \neq -1} \delta_{e,l}(\mathbb{1}\{E_i = e\} D_{i,t-l}) + \alpha_i + \lambda_t + \zeta_g + \varepsilon_{i,t,g}\ , \tag{3}$$

   where C is the set of comparison cohorts. In our case C is the never treated cohort, i.e. $C = \infty$. If there is a cohort that is always treated, i.e. that already receives treatment in the first period, then we need to exclude this cohort. The coeffiecient estimator $\widehat{\delta}_{e,l}$ that we obtain from (3) estimates $CATT_{e,l}$.

2. Weight the estimators by the share of the respective cohort in the sample in that period. Let $\hat{W}^l$ be a weight matrix with element (t, e)

$$[\widehat{W}^l]_{t,e} := \frac{\mathbb{1}\{t - e = l\} \sum_{i=1}^{N} \mathbb{1}\{E_i = e\}}{\sum_{e \in h^l} \sum_{i=1}^{N} \mathbb{1}\{E_i = e\}},$$

   where $h^l := \{e : 1 - l \leq e \leq \max(E_i) - 1 - l\}$ is the set of cohorts that experience at least l periods of treatment.

3. Take the average over all $CATT_{e,l}$ estimates weighted by the cohort shares in the weight matrices. Let $vec(A)$ be the vectorize operator that vectorizes matrix A by stacking its columns and let $\widehat{\delta}$ be the vector that collects $\widehat{\delta}_{e,l}$ for all e and l. Then, the IW estimator $\widehat{v}_g$ for bin g can be written as

$$\widehat{v}_g := \frac{1}{|g|} \sum_{l \in g} [vec(\widehat{W}^l)]^{\intercal} \widehat{\delta}. \tag{4}$$

   For a singleton bin $g = \{l\}$, this simplifies to

$$\widehat{v}_g := [vec(\widehat{W}^l)]^{\intercal} \widehat{\delta}.$$

Under some standard assumptions, $\widehat{v}_g$ is asymptotically normal (for a proof and a detailed description of said assumptions see Sun and Abraham, 2021, Appendix C). Under the additional assumptions of parallel trends and no anticipatory behavior, $\widehat{v}_g$ is consistent, that is it converges in probability to

$$\widehat{v}_g \xrightarrow{p} [vec(W^l)]^\intercal \delta = \sum_{e \in h^l} \mathbb{P}(E_i = e | E_i \in h^l) CATT_{e,l} \ ,$$

where W^l is the probability limit of the weight matrix $\widehat{W}^l$.

We use $\widehat{v}_g$ as an estimator for β_g in equation (2) and we exploit the existing implementation in the **fixest** R package (Bergé, 2018).

3.3 Identifying assumptions

Below we discuss the identifying assumptions.

Parallel Trends: Parallel trends in the sense of Sun and Abraham (2021) refers to the following: $\mathbb{E}[Y_{i,t}^{\infty} - Y_{i,s}^{\infty} | E_i = e]$ does not depend on e for any $s \neq t$. That is, the expected temporal difference, i.e. the trend, in the potential outcomes of being never-treated is the same for all treatment timings. A conditional version of the assumption, as in Callaway and Sant'Anna (2021), should definitely hold, as test scores and natural disasters are plausibly independent given location. However, we cannot be sure about the unconditional version required by Sun and Abraham (2021).

Testing for parallel trends is problematic for two reasons: These tests tend to have very low power and they introduce selective inference type issues if inference is conditional on passing a parallel trends test (Rambachan and Roth, 2019). A purely visual inspection of pre-treatment trends does not indicate a violation of the parallel trends assumption (see appendix B). In fact, the trends look almost identical for treated and control (never-treated) units.

No Anticipatory Behavior: There is no treatment effect prior to treatment, that is $\mathbb{E}[Y_{i,e+l} - Y_{i,e+l}^{\infty}] = 0$ for all e and all $l < 0$. This assumption is plausible as the treatment path is not known. Natural disasters are quasi-random and cannot be reliably forecast more than a few days in advance. Thus, anticipatory behavior is implausible.

Both identifying assumptions should be fulfilled and the IW-Estimator consistently estimates a weighted average of the cohort average treatment effects on the treated.

4 Results

Figure 6 and 7 show estimated dynamic treatment effects and 95% confidence intervals for all students and the four subgroups of interest respectively.

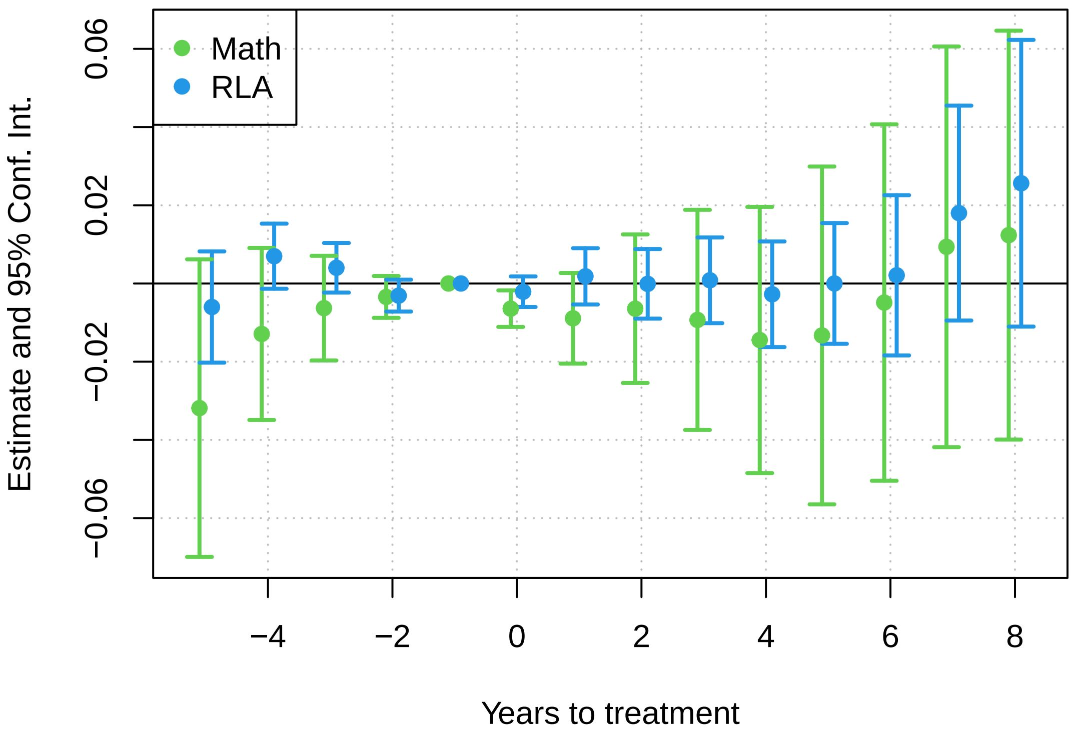

Figure 6: Dynamic Treatment effects for all students: FEMA disaster data

For the period of treatment there is a significant[1] effect of natural disasters on the performance in mathematics. The effect size is between just above zero and -0.01 standard deviations. For all subsequent periods the effect is not significant. There are some point estimates well below zero, but the uncertainty around those is relatively large. For performance in RLA, there are no significant effects.

Note that the number of observed units decreases with the distance in time from treatment. The reason for this is that in order to experience eight treated years, the county has to experience its first disaster very early. Similarly, it has to receive treatment very late to experience more than five years before treatment. As a result, the uncertainty increases with the distance in time from treatment.

For the subgroups we find some surprising results. Black students seem to perform better in RLA in the medium term after a disaster. That is, there are significantly positive results one to seven years after treatment. The effect sizes are substantial: Seven years after treatment the increase in

[1] Significant is used here in the sense that a confidence interval with nominal coverage of 95% does not include zero, that is a corresponding t-test would reject the null hypothesis of a zero effect at the 5% level.

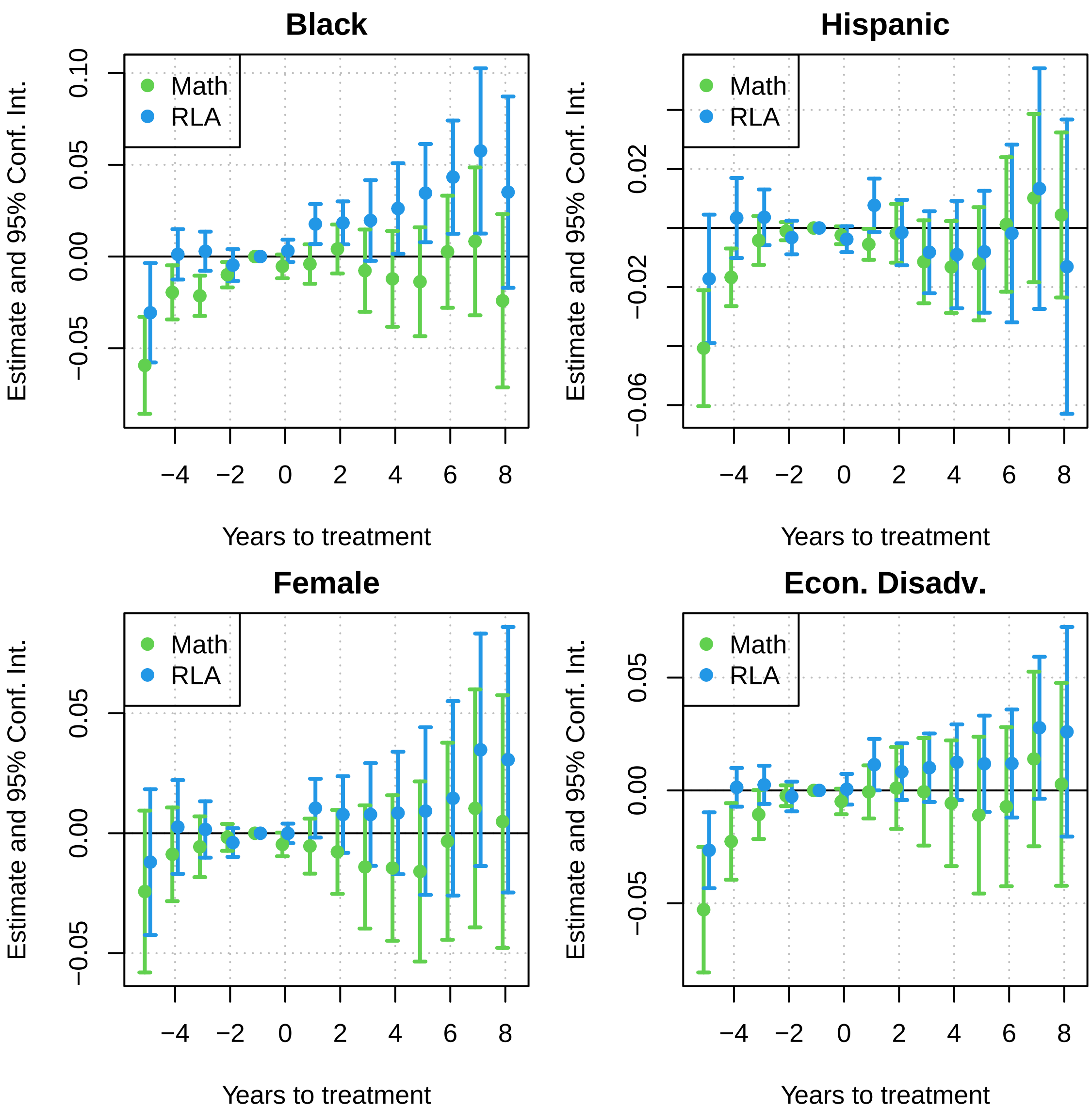

Figure 7: Dynamic Treatment effects for various subgroups: FEMA disaster data

RLA performance goes up to 0.1 standard deviations. In other words, the average black student sees an increase in performance of up to 0.1 standard deviations of the national reference cohort. Also, hispanic students score significantly lower in mathematics in the year following a disaster.

Positive effects of disasters on performance are not unheard of in the literature. In fact, this is somewhat consistent with the findings by Sacerdote (2012). Many students have to switch schools and some may even benefit from attending a higher quality school after the disaster. Black students may disproportionally attend lower quality schools and are therefore more likely to benefit from having to switch schools.

Figures 8 and 9 show the same graphs based on the storm treatment. The results look very similar. In the period of the storm there is a significant decrease in mathematics scores of up to -0.015 standard deviations. For the years following treatment there are no significant effects.

For female students there is a significant decrease in both subjects in the period of the storm.

For RLA we even find a significantly negative effect one year after the storm. Similarly, economically disadvantaged students perform worse in the period of treatment and in RLA one year after treatment. The effect sizes range from barely above zero up to -0.015 or even -0.02 standard deviations. For black and hispanic students we do not find any significant effects of the storm treatment.

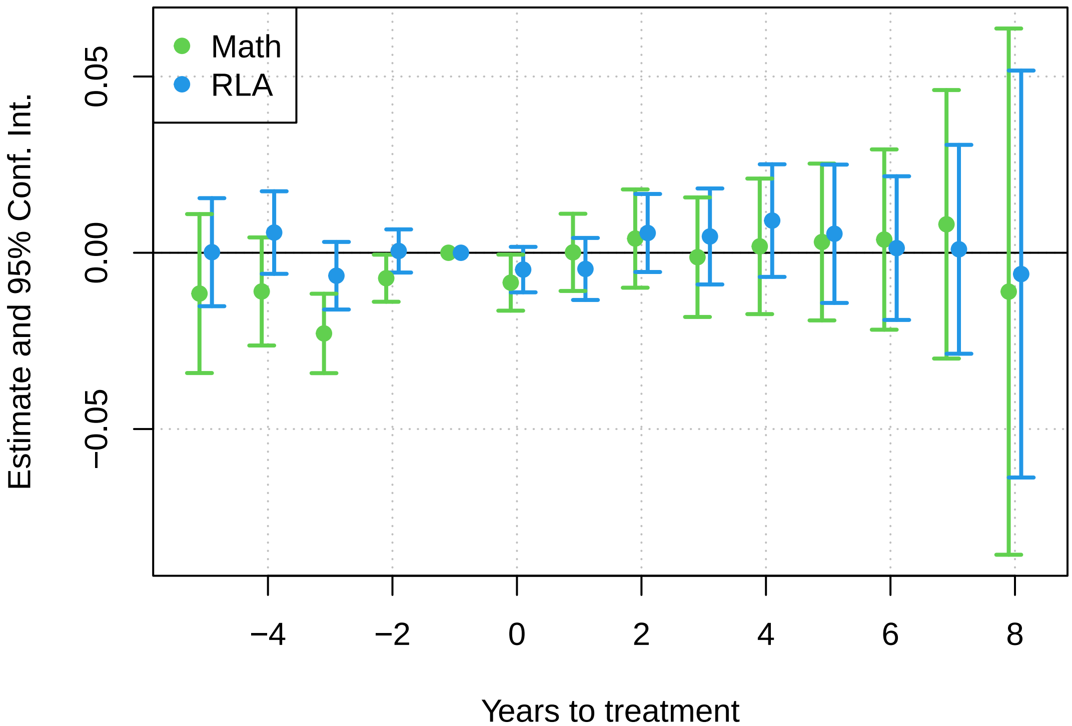

Figure 8: Dynamic Treatment effects for all students: NWS storm data

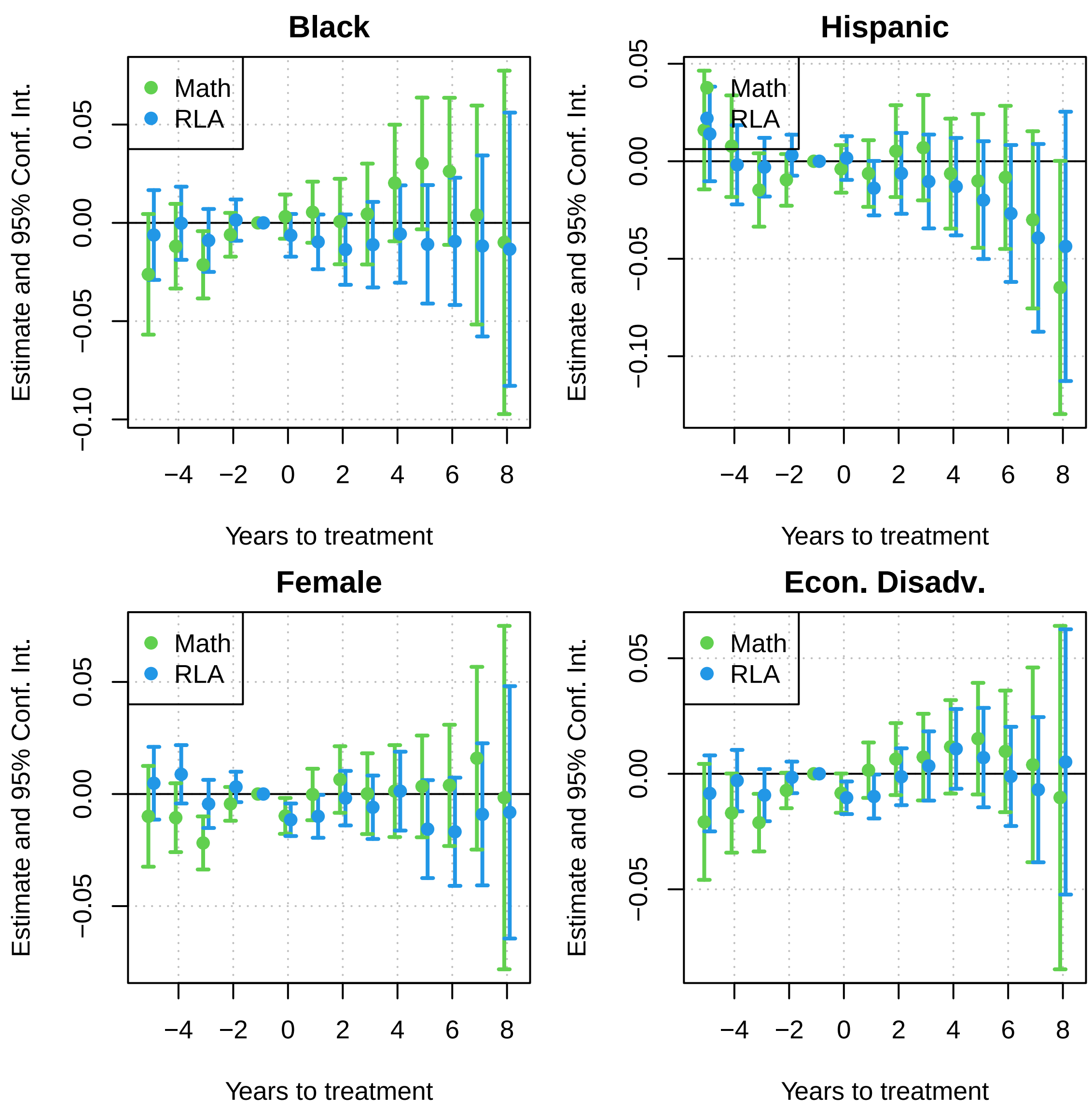

Figure 9: Dynamic Treatment effects for various subgroups: NWS storm data

5 Conclusion

This study estimates dynamic effects of natural disasters on academic performance measured by standardized test results in mathematics and Reading Language Arts (RLA). For both datasets we find a negative effect on the performance in mathematics in the year the disaster occured. The effect reaches up to 0.01 or even 0.015 standard deviations of the national reference cohort. For RLA we find no significant effects on the overall mean score.

Based on FEMA natural disaster data we find that the performance in RLA among black students increases substantially in the years following a natural disaster. The reason could be that black students may disproportionately benefit from having to switch schools after a disaster (Sacerdote, 2012). However, the same model estimated on the NWS storm data does not confirm these findings.

In total, there is strong evidence for a negative effect of disasters on performance in mathematics in the same school year. For RLA, on the other hand, there is no significant effect. Evidence for medium and long term effects is weak. There are some significant effects among minority students, but they do not seem to be very robust.

Mitigating such negative effects should be a concern for policymakers. Otherwise negative performance effects could compound in strongly affected regions.

A Additional Results

A.1 Logistic regression for assistance applications

Below we report logistic regression results for the applicant status, that is whether a county applied for federal disaster assistance based on the Public Assistance Applicants Program Deliveries data. This is regressed on a few variables, including the share of democratic votes in the 2016 election. The other independent variables are also from 2016.

Similarly, the declaration status, that is whether a county had any natural disasters declared during the 2009 to 2018 period. This is regressed on the share of democratic votes in the 2008 election and the set of same control variables.

Table 2: Determinants of Assistance Application

Dependent Variables: Model:	Applicant (1)	Declared (2)
Variables		
(Intercept)	-3.622	-10.83***
	(3.723)	(4.000)
Share of democratic voters (2016)	-0.8362**	
	(0.3469)	
Median Income (logs)	0.2939	0.9862***
	(0.3331)	(0.3586)
Poverty Rate	4.033**	0.2160
	(1.575)	(1.730)
Share of single mothers	4.068***	12.18***
	(1.144)	(1.241)
Share of democratic voters (2008)		-1.496***
		(0.3624)
Fit statistics		
Observations	2,882	2,882
Squared Correlation	0.02306	0.05222
Pseudo R^2	0.01966	0.04725
BIC	3,724.7	3,122.7

IID standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

B Pre-Treatment Trends

Here we show plots of aggregated pre-treatment trends to justify the parallel trends assumption. Mean test scores are aggregated by cohort (year of first treatment) and relative time to treatment, and never treated units act as the control group. We only display these plots for overall test scores for both datasets, but not for subgroups. However, the plots for the subgroups look very similar.

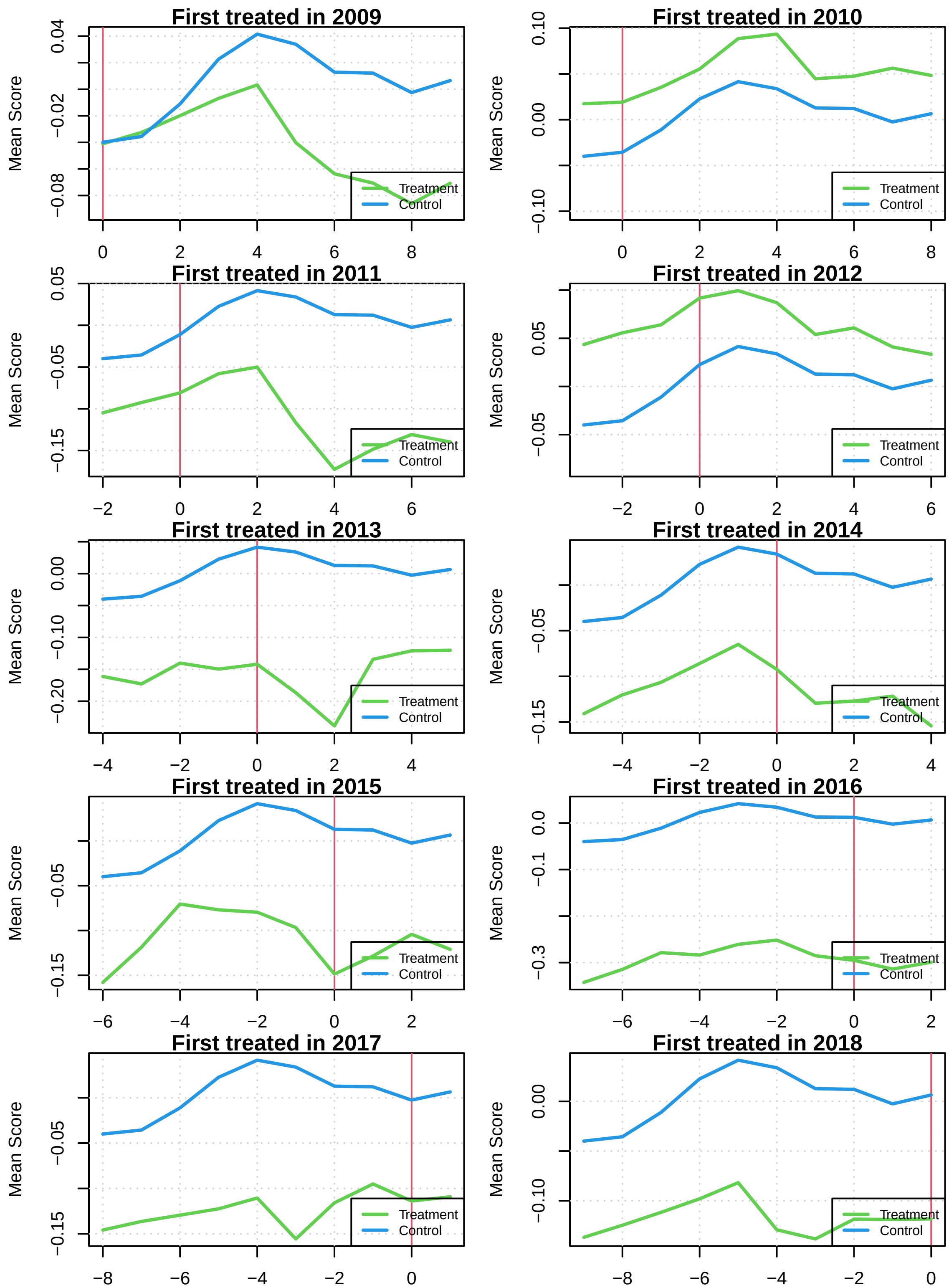

Figure 10: Aggregated mean scores in mathematics based on FEMA data in relative time to treatment

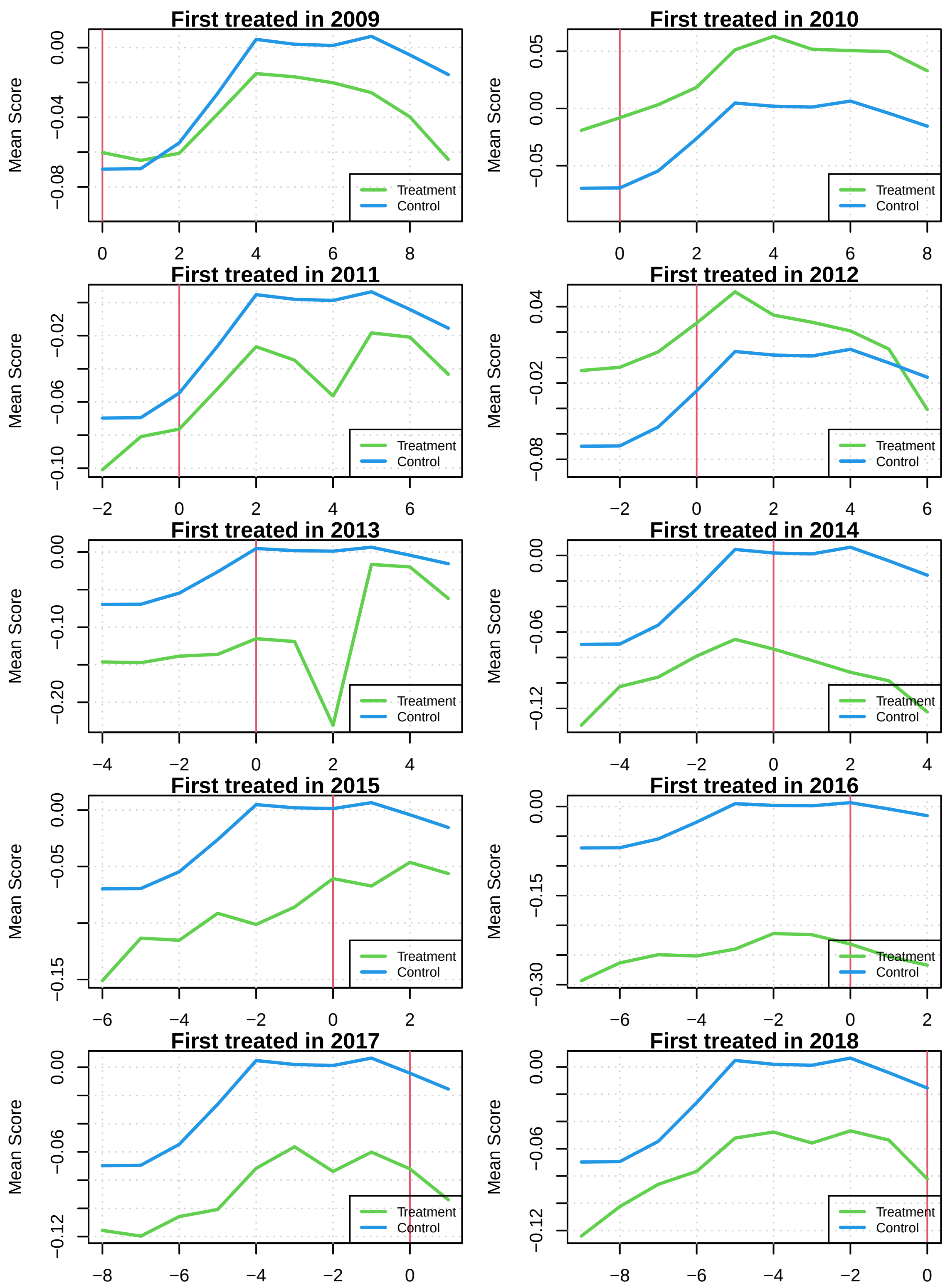

Figure 11: Aggregated mean scores in RLA based on FEMA data in relative time to treatment

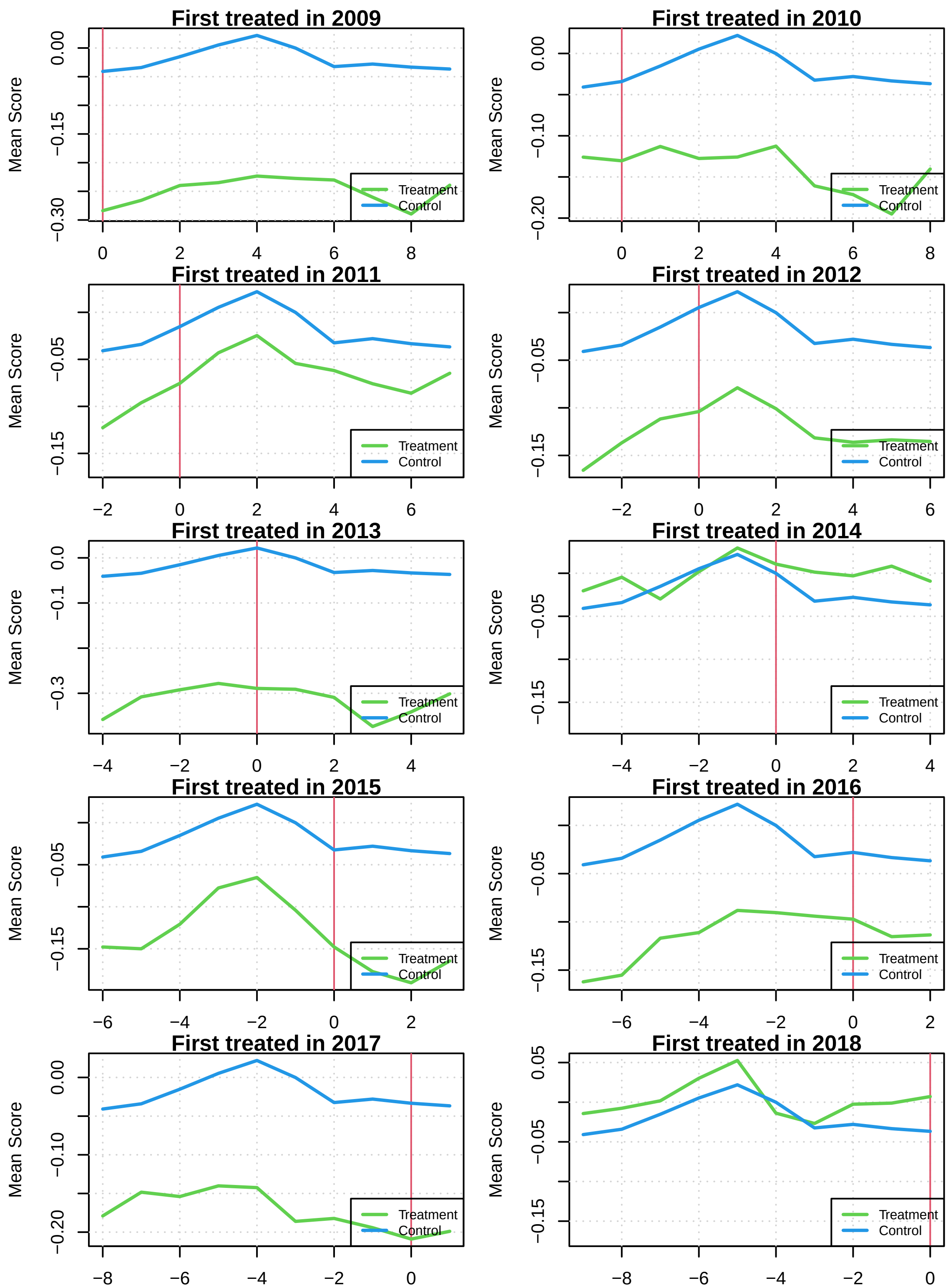

Figure 12: Aggregated mean scores in mathematics based on NWS storm data in relative time to treatment

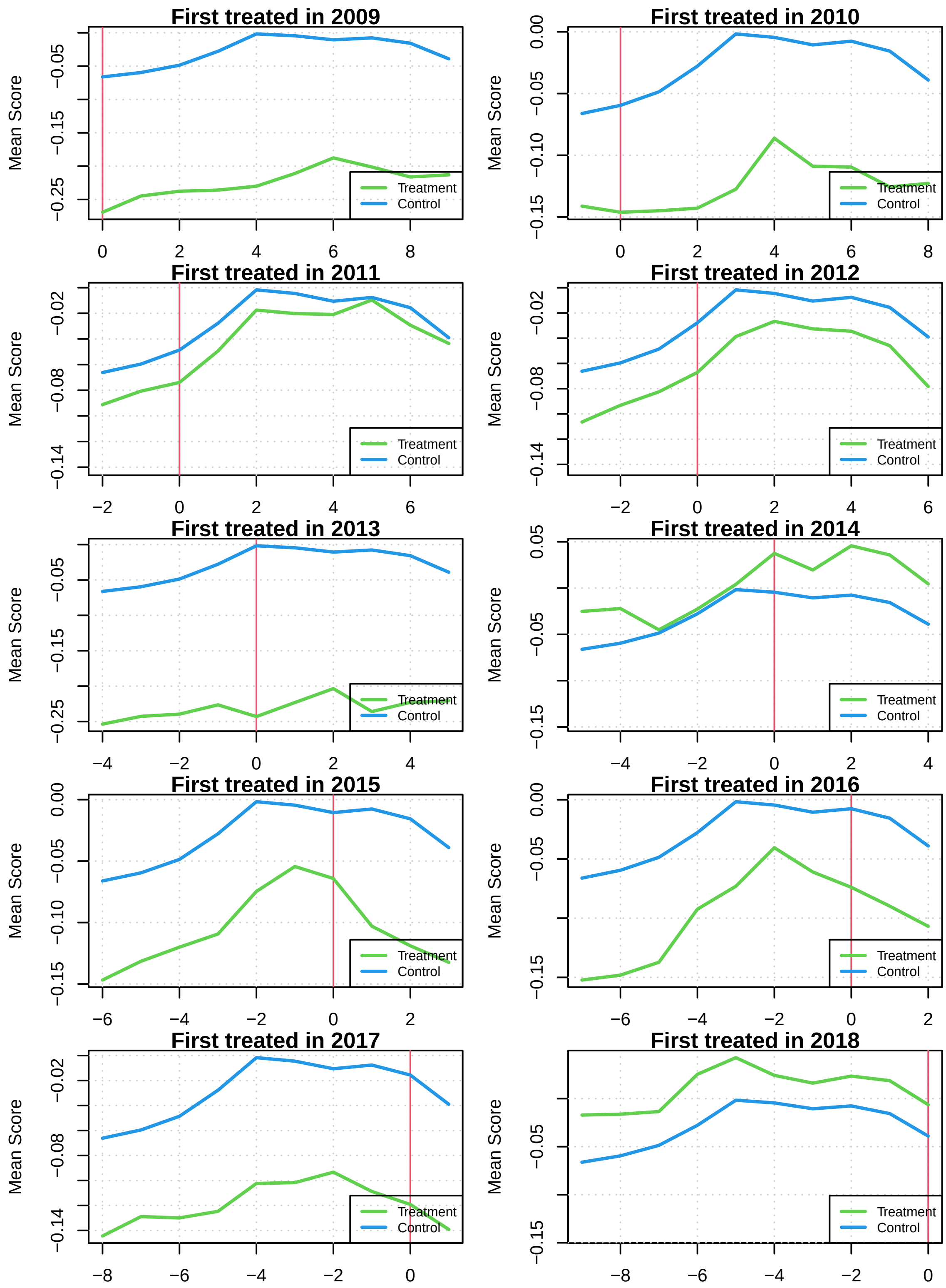

Figure 13: Aggregated mean scores in RLA based on NWS storm data in relative time to treatment

References

Abadie, A., Athey, S., Imbens, G. W., and Wooldridge, J. (2017). When should you adjust standard errors for clustering? Technical report, National Bureau of Economic Research.

Baggerly, J. and Ferretti, L. K. (2008). The impact of the 2004 hurricanes on florida comprehensive assessment test scores: Implications for school counselors. *Professional School Counseling*, 12(1):1–9.

Bergé, L. (2018). Efficient estimation of maximum likelihood models with multiple fixed-effects: the R package FENmlm. *CREA Discussion Papers*, (13).

Borusyak, K., Jaravel, X., and Spiess, J. (2021). Revisiting event study designs: Robust and efficient estimation. *arXiv:2108.12419*.

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

Crespo Cuaresma, J. (2010). Natural disasters and human capital accumulation. *The World Bank Economic Review*, 24(2):280–302.

de Chaisemartin, C. and D'Haultfoeuille, X. (2021). Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: A survey. *arXiv:2112.04565*.

de Chaisemartin, C. and D'Haultfœuille, X. (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review*, 110(9):2964–96.

Deryugina, T. (2017). The fiscal cost of hurricanes: Disaster aid versus social insurance. *American Economic Journal: Economic Policy*, 9(3):168–98.

Deryugina, T., Kawano, L., and Levitt, S. (2018). The economic impact of hurricane katrina on its victims: Evidence from individual tax returns. *American Economic Journal: Applied Economics*, 10(2):202–33.

Grewenig, E., Lergetporer, P., Werner, K., Woessmann, L., and Zierow, L. (2021). Covid-19 and educational inequality: How school closures affect low- and high-achieving students. *European Economic Review*, 140:103920.

Holmes, G. M. (2002). Effect of extreme weather events on student test performance. *Natural Hazards Review*, 3(3):82–91.

Intergovernmental Panel on Climate Change (IPCC) (2021). *Climate Change 2021: The Physical Science Basis.* Cambridge University Press.

Mcdonald, J., Forbes, G., and Marshall, T. (2004). The enhanced fujita scale (ef).

Pane, J., Mccaffrey, D., Kalra, N., and Zhou, A. (2008). Effects of student displacement in louisiana during the first academic year after the hurricanes of 2005. *Journal of Education for Students Placed at Risk (jespar)*, 13:168–211.

Park, R. J., Goodman, J., Hurwitz, M., and Smith, J. (2020). Heat and learning. *American Economic Journal: Economic Policy*, 12(2):306–39.

Rambachan, A. and Roth, J. (2019). An honest approach to parallel trends. *Unpublished manuscript, Harvard University.*

Ramsey, J. D. (1995). Task performance in heat: a review. *Ergonomics*, 38(1):154–165. PMID: 7875117.

Reardon, S., Kalogrides, D., Ho, A., Shear, B., Fahle, E., Jang, H., and Chavez, B. (2021). Stanford education data archive (version 4.1).

Sacerdote, B. (2012). When the saints go marching out: Long-term outcomes for student evacuees from hurricanes katrina and rita. *American Economic Journal: Applied Economics*, 4(1):109–35.

Spencer, N., Polachek, S., and Strobl, E. (2016). How do hurricanes impact scholastic achievement? a caribbean perspective. *Natural Hazards*, 84(2):1437–1462.

Sun, L. and Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, 225(2):175–199.

Turner, D. (2022). rfema: Access the openfema api. *rOpenSci.*

Vogel, S. and Schwabe, L. (2016). Learning and memory under stress: implications for the classroom. *npj Science of Learning*, 1(1).