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

Gregor Steiner

July 14, 2022

Abstract

The frequency and severity of natural disasters has increased recently, but we do not fully understand the impacts. We use an event-study design to estimate dynamic treatment effects of natural disasters on students' performance in standardized tests. There is a significant effect on the performance in mathematics in the period of the disaster. For achievement in Reading Language Arts we find no significant effect. Furthermore, we find a significantly negative effect of extreme heat exposure on standardized test scores for some groups.

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, 2021). Therefore, it is essential to have a good understanding of the consequences. Few studies have focused on the impact on the education system, yet this is a crucial part of understanding the economic consequences. Negative effects in education are likely to have long-term impact on earnings potential. Inequality in disaster risk exposure could therefore exacerbate economic inequality.

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). In the very short term, emotional stress caused by extended housing instability, food insecurity, parental job loss, or social disconnection may be the most important factor, especially so in socially vulnerabe communities (United States Government Accountability Office, 2022). Furthermore, some forms of disasters, e.g. extreme heat, may directly impair cognitive performance (Ramsey, 1995). Lastly, many students have to relocate and switch schools following a disaster. Such migration responses may also have a significant effect on the scholastic achievement of affected students (Pane et al., 2008; Sacerdote, 2012).

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 2017/2018. 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.

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

Furthermore, we investigate the effect of heat exposure as a very special case of natural disasters. We find negative effects on some groups, but there seems to be substantial heterogeneity.

Previous Work: This article contributes to the literature on the impact of natural disasters on the education system in the United States. Previous work has produced mixed results. Some authors find significant effects of natural disasters on the education system, while others find no or only small effects.

Holmes (2002) was among the first to study the effect of extreme weather events on academic achievement. Using a difference-in-differences approach, he finds a significantly negative effect of storms on the performance of North Carolina students. Baggerly and Ferretti (2008) find a statistically significant, but negligibly small effect of the 2004 hurricanes on the performance of students' test scores in Florida. Lamb et al. (2013) study the effects of hurricane Katrina and find a significant impact on mathematics achievement in Mississippi with the greatest effects in nonpoor and nonrural schools.

Park et al. (2020) find that cumulative heat exposure negatively impacts PSAT scores. At the same time, they find air conditioning to be very successful at mitigating the negative effect of heat exposure. Similarly, Park (2022) uses student level data from New York City to assess the impact of heat in high stakes exams. He finds a substantial negative effect on performance.

Many authors have focused on the role of student mobility as a consequence of natural disasters. Pane et al. (2008) focus on students who switch schools following hurricanes Katrina and Rita and find small negative effects of displacement on test scores. Similarly, Sacerdote (2012) finds sharp declines in test scores one year after the hurricanes, but a substantial improvement three to four years after. This is largely driven by the students' tendency to switch to higher quality schools.

Most authors have focused on a single type or even a single instance of disasters. In particular, storms and heat have been investigated predominantly. The main contribution of this article is the very comprehensive dataset which covers multiple years and many different types of disasters. To the best of our knowledge, this is the first analysis using such a broad set of natural disasters to analyze the effect on academic performance.

The rest of this paper is organized as follows: Section 2 gives some institutional background, 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 Background & Data

2.1 Natural Disasters in the United States

Natural disasters cause numerous fatalities and billions of dollars of infrastructure damage in the United States each year (Boustan et al., 2020). With more greenhouse gas in the atmosphere the number and severity of such disasters will increase (Intergovernmental Panel on Climate Change, 2021). Figure 1 shows the number of county-level disasters per year. However, note that a single disaster that affects multiple counties counts as multiple county-level disasters. The majority of natural disasters in the US is made up by storms, such as hurricanes or tornadoes, floodings, and fires. Some well-known examples include the 2020 wildfire season in the western United States, Hurricane Harvey in 2017, or Hurricane Katrina in 2005.

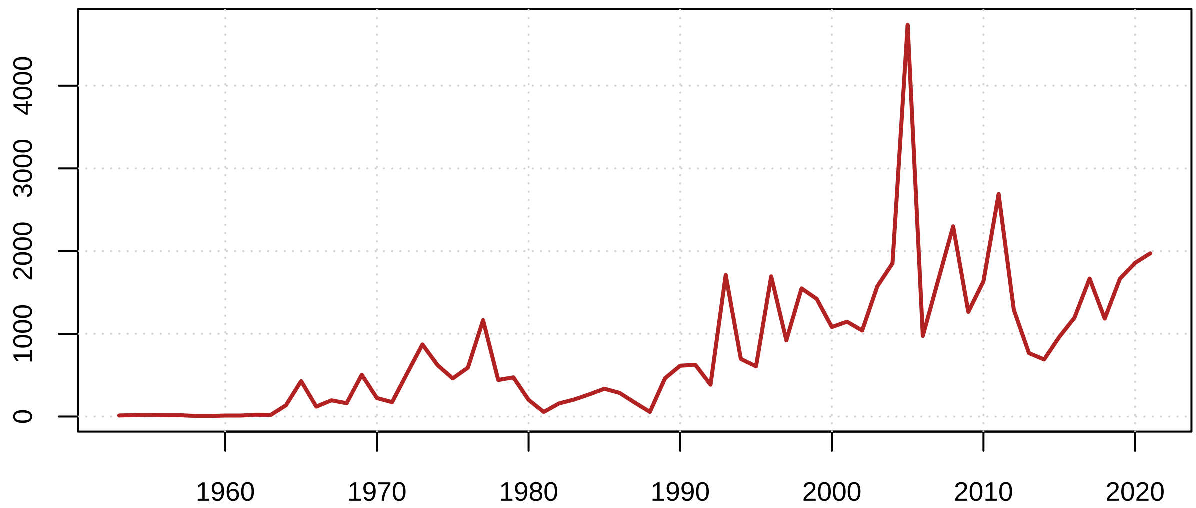

Figure 1: Number of county-level natural disasters by year

The handling of natural disasters is governed by the Robert T. Stafford Disaster Relief and Emergency Assistance Act (1988). It defines major disasters as follows:

> Major disaster means any natural catastrophe (including any hurricane, tornado, storm, high water, winddriven water, tidal wave, tsunami, earthquake, volcanic eruption, landslide, mudslide, snowstorm, or drought), or, regardless of cause, any fire, flood, or explosion, in any part of the United States, which in the determination of the President causes damage of sufficient severity and magnitude to warrant major disaster assistance under this Act to supplement the efforts and available resources of States, local governments, and disaster relief organizations in alleviating the damage, loss, hardship, or suffering caused thereby.

Based on the Stafford Act major disaster declarations are made solely by the president upon request by the affected state's governor. State officials must prove that the situation is beyond the capabilities of the state governemnt or involved local governments. The Federal Emergency Management Agency (FEMA) evaluates requests for major disasters and makes recommendations to the president.

A major disaster declaration provides a wide range of federal assistance programs for individuals and public infrastructure, including funds for both emergency and permanent work. In particular, affected state and local governments can receive federal disaster assistance as part of the Public Assistance program, while individuals and households can receive aid from the Individual Assistance program. Moreover, State, Tribal, and local governments and certain private nonprofit organizations can receive assistance for preventive actions taken based on the Hazard Mitigation Assistance program.

FEMA provides data on all federally declared natural disasters, beginning in 1953. The data is easily accessible via their API (Turner, 2022). This is the main source for natural disaster data used here. Figure 2 shows the number of declared disasters between 2009 and 2018 across the US. Table 1 shows the types of disasters and their proportion in the FEMA data. Storms make up the largest share of disaster events. Fires and floods also form a substantial part.

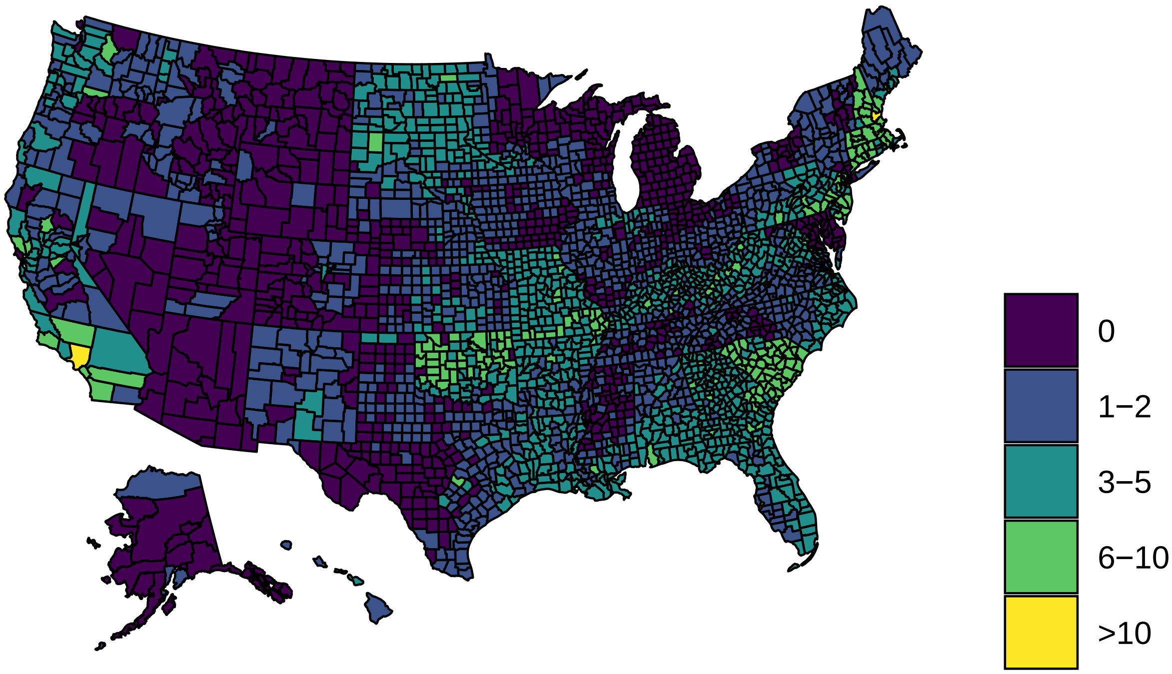

Figure 2: Number of declared natural disasters in school years 2008-09 through 2017-18

Table 1: Disasters from 2009 to 2018 by type

Variable	N	Percent
Disaster Type	7547	
... Chemical	9	0.1%
... Coastal Storm	11	0.1%
... Dam/Levee Break	3	0%
... Earthquake	5	0.1%
... Fire	220	2.9%
... Flood	1116	14.8%
... Freezing	1	0%
... Hurricane	2081	27.6%
... Mud/Landslide	28	0.4%
... Other	1	0%
... Severe Ice Storm	804	10.7%
... Severe Storm(s)	2691	35.7%
... Snow	519	6.9%
... Tornado	48	0.6%
... Tsunami	9	0.1%
... Volcano	1	0%

We repeat the analysis on two different datasets: Data on storms and data on heat exposure. First, this acts as a robustness check. These two datasets are based on obective measurements, while the FEMA data is based on subjective declarations. Second, this allows us to better understand heterogeneity by type of disaster.

The National Weather Service (NWS) provides data on storm events. In particular, this covers hurricanes, 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.

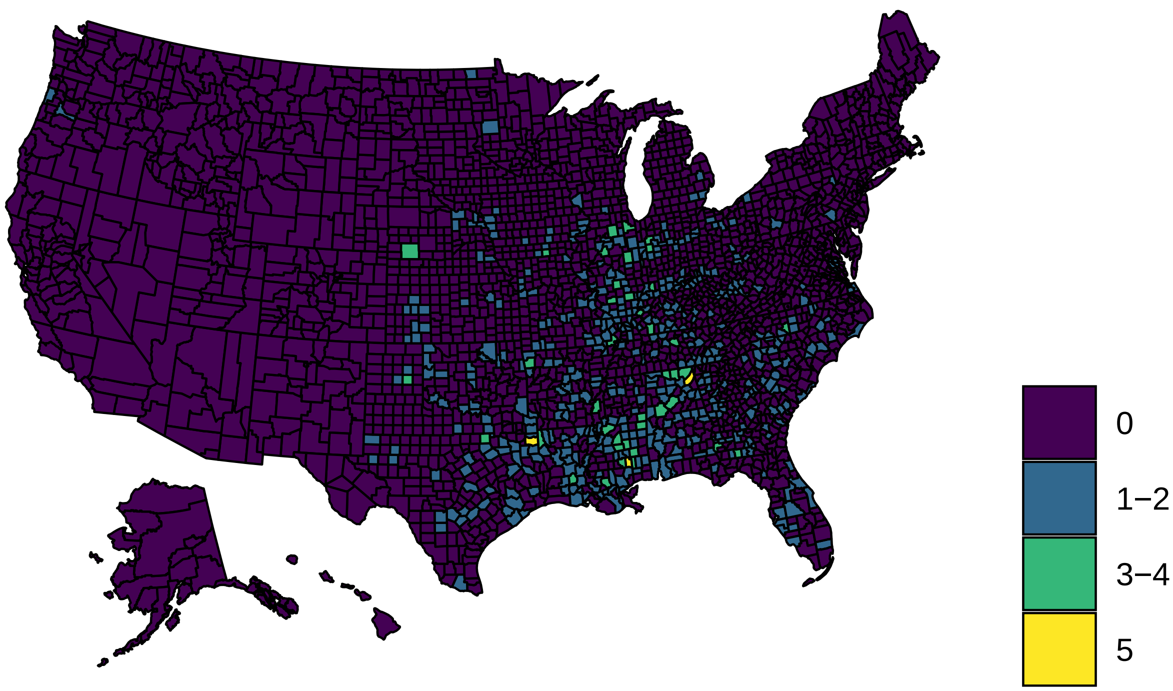

Figure 3: Number of storms in school years 2008-09 through 2017-18

We only consider severe storms which are likely to cause substantial damage. Tornadoes can be classified based on estimated peak wind speeds on the Enhanced Fujita (EF) scale (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 exclude all hurricanes with an estimated property damage of zero. Storm exposure by county is shown in figure 3.

To measure heat exposure, we exploit daily temperature data from the Global Historical Climatology Network (Menne et al., 2012). Each county is assigned the measurement station with the lowest distance to the county's center. Following Park et al. (2020), we use two measures of yearly cumulative heat exposure: The average daily maximum temperature and the number of days with a maximum temperature above 30°C in a school year. The threshold of 30°C is somewhat arbitrary, however, changing it slightly does not change the results in a meaningful way. Figures 4 and 5 show the distribution across counties. As expected, the variation in heat exposure is largely driven by location.

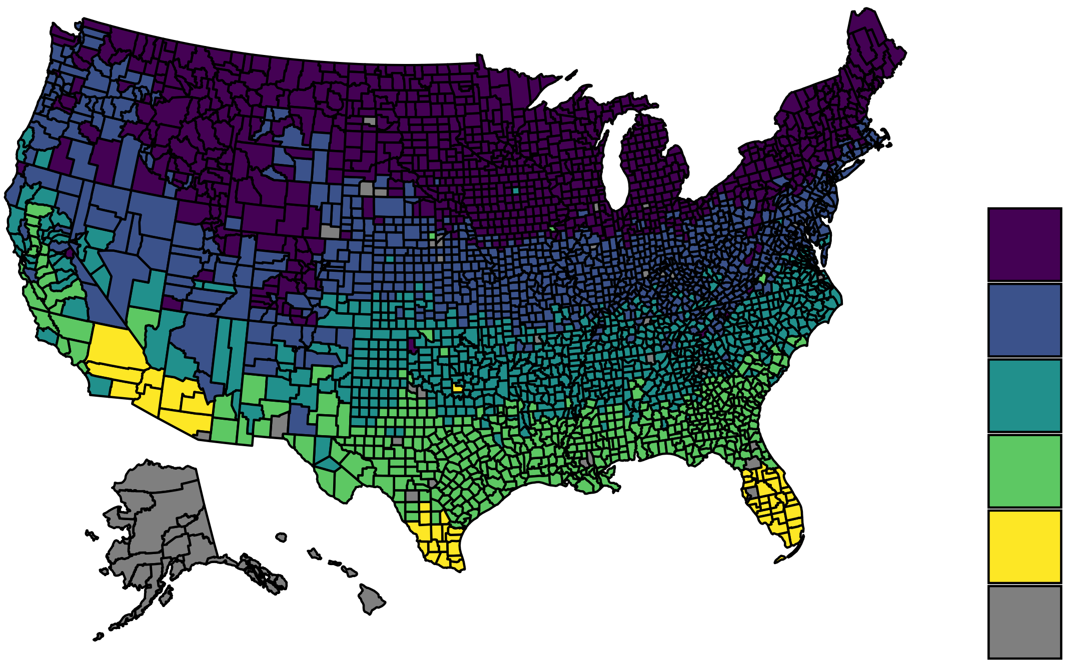

Figure 4: Average daily maximum temperature (in °C) in school years 2008-09 through 2017-18

2.2 Assistance Applications

After severe disasters counties frequently receive public assistance. This potentially creates a selection problem: The counties receiving state aid are likely coping better with the consequences of the natural disaster. If the assignment of federal aid is not independent of the disaster's consequences, this creates a problem for our identification. Since assistance provision is at least somewhat dependent on disaster severity, this is very probable.

FEMA provides a dataset on their Public Assistance Applicants Program Deliveries. It contains information on applicants and their recovery priorities, including the amount of damage caused and amount of federal disaster assistance granted. This is the main program for federal public disaster assistance, averaging USD 4.7 billion in assistance each year. However, note that counties that do not receive public assistance may still benefit from individual assistance or from the Hazard Mitigation Assistance program.

Unfortunately, this data is only available starting in October 2016. Based on the temporal overlap between this dataset and our main dataset, that is schoolyears 2016/2017 and 2017/2018, it is possible to analyze whether counties that are affected by disasters receive federal assistance. This can be done by checking whether a county that experienced a disaster in 2016/2017 or 2017/2018 appears in the Public Assistance Applicants Program Deliveries database. In fact, only about 17.78% of counties that experienced a disaster in that period did apply for federal assistance. Table 2 shows the number of disasters and the share of counties that applied for assistance following such a disaster by disaster type. It seems that the number varies dramatically by disaster type.

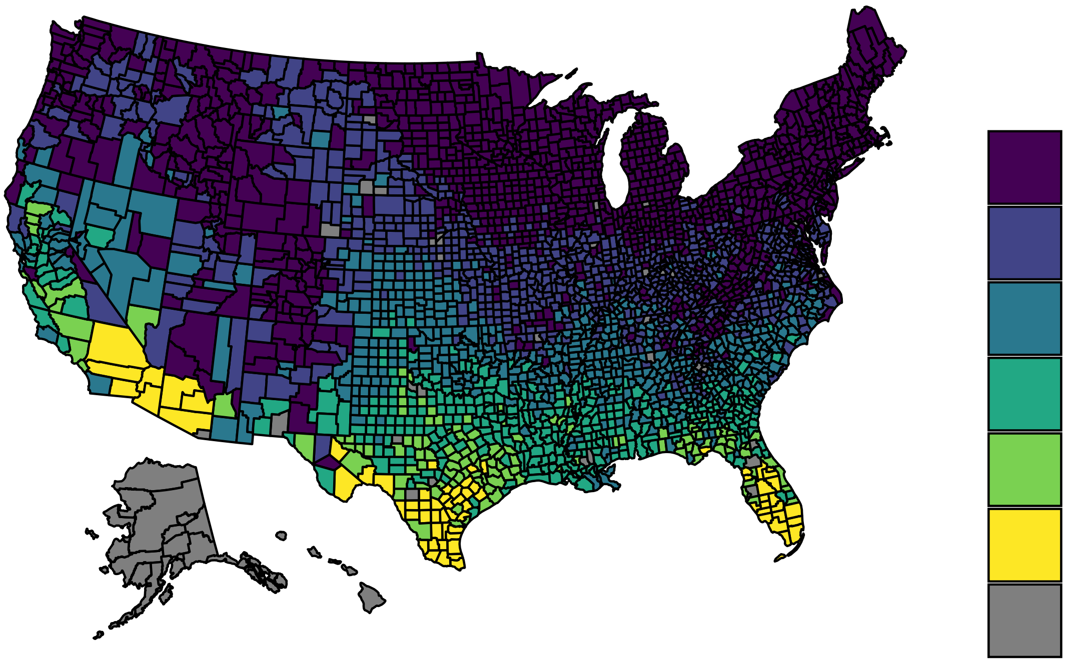

Figure 5: Average number of days above 30°C in school years 2008-09 through 2017-18

While about 80% of counties affected by a tornado applied for federal assistance, only about 10% of counties affected by fires or floods did so.

It may be interesting to see how these counties differ from the ones that did apply. Figure 6 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. This is consistent with the findings in United States Government Accountability Office (2022).

However, the direction of causality is not clear. Possibly these counties are more frequently affected by 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. While there is some correlation between disaster risk and economic conditions (for example Park et al., 2020), the latter explanation seems more likely overall.

It is also interesting whether variation in the federal assistance procedure may be driven by political factors. Visually, democratic votes in the 2016 election (almost coincides with the start of the Public Assistance Applicants Program Deliveries dataset) tend to be lower in counties that applied. That is, counties that applied for federal assistance tend to vote more Republican. Logistic regression results confirm the visual impression (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.

There are clear differences between counties that receive federal assistance after a disaster and counties that do not. This is a potential problem to the identification of a causal effect. However, counties that are severely affected and receive assistance would most likely experience even worse

Table 2: Share of counties that applied for federal assistance following a disaster by disaster type (schoolyears 2016-17 and 2017-18)

	Number of Cases	Applied for Assistance (in %)
Coastal Storm	3	33.33
Dam/Levee Break	3	0.00
Fire	100	11.00
Flood	270	41.85
Hurricane	1217	23.25
Mud/Landslide	22	50.00
Severe Ice Storm	20	0.00
Severe Storm(s)	164	28.66
Snow	36	8.33
Tornado	29	79.31
Total	1864	26.39

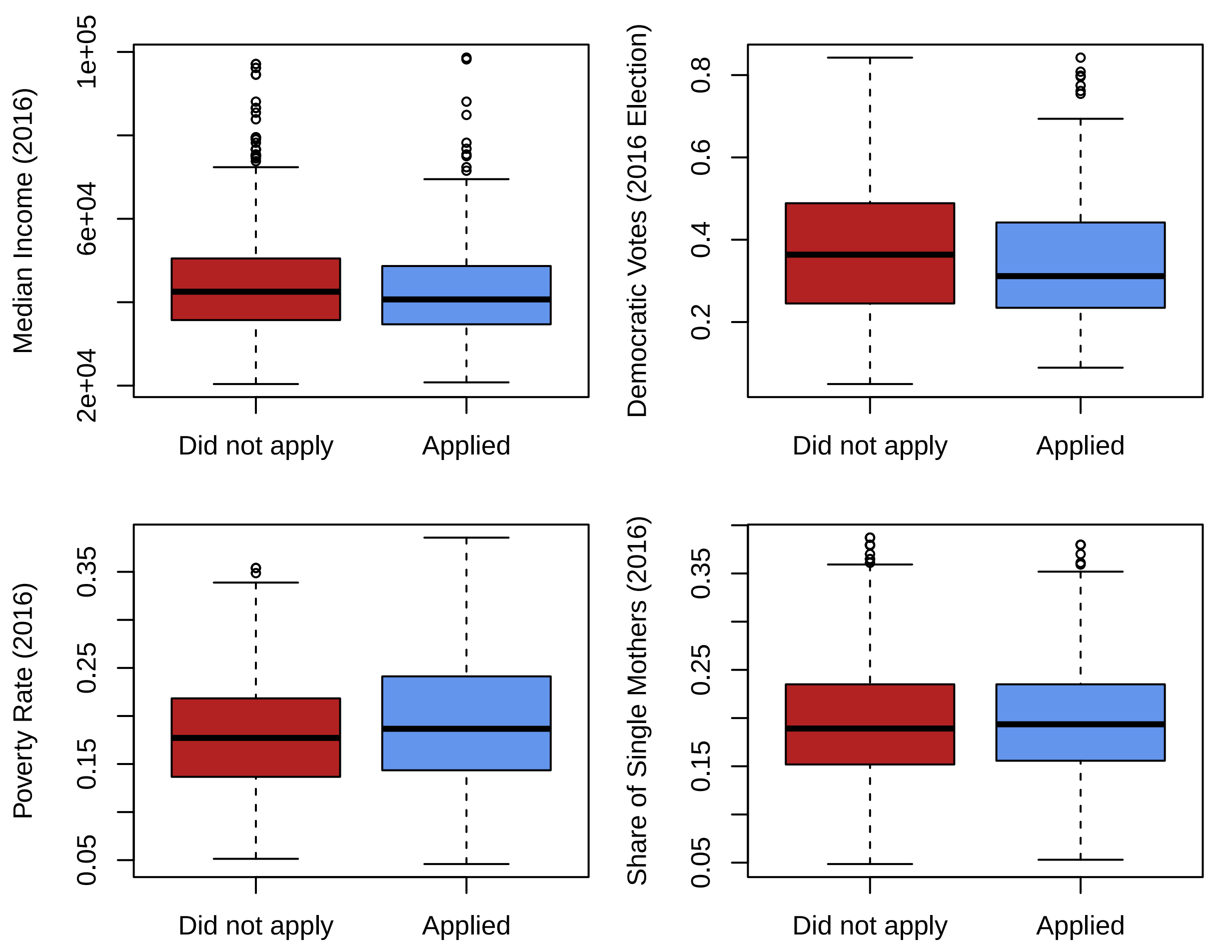

Figure 6: Boxplots by application status

Table 3: Summary statistics for dependent variables

	Overall	White	Black	Hispanic	Female	Econ. disadv.
Mean	-0.042	0.107	-0.483	-0.281	0.025	-0.284
Std. Dev.	0.294	0.262	0.273	0.266	0.295	0.256
Min.	-3.196	-2.936	-2.745	-1.699	-2.862	-3.007
Max.	1.669	1.700	1.394	1.374	1.496	1.312

consequences, did they not receive assistance. Thus, our estimates may only be a lower bound for the negative effect.

2.3 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)[1] 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. This makes the dataset very attractive, as test scores are nationally comparable. 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 white, black, hispanic, female, and economically disadvantaged students to investigate whether the effects differ by ethnicity, gender, or socioeconomic position. These are only reported if the subgroups' sample sizes are large enough to guarantee anonymization. Thus, the number of observations for some of the groups is substantially smaller.

The outcomes of interest are overall mean test scores by county, and mean test scores for white, black, hispanic, female, and, economically disadvantaged students. Figure 7 shows boxplots for the five outcomes of interest and Table 3 provides summary statistics. 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 tend to be lower than the overall means, while white students tend to perform slightly better than the overall average. Female mean scores are slightly above overall mean scores, meaning that female students perform slightly better than male students on average.

Natural disasters should only have an effect on test scores if they occur before the test. Standardized tests are generally administered in March, April, or May. 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.

[1] RLA assesses students' ability to understand what they read and to write clearly.

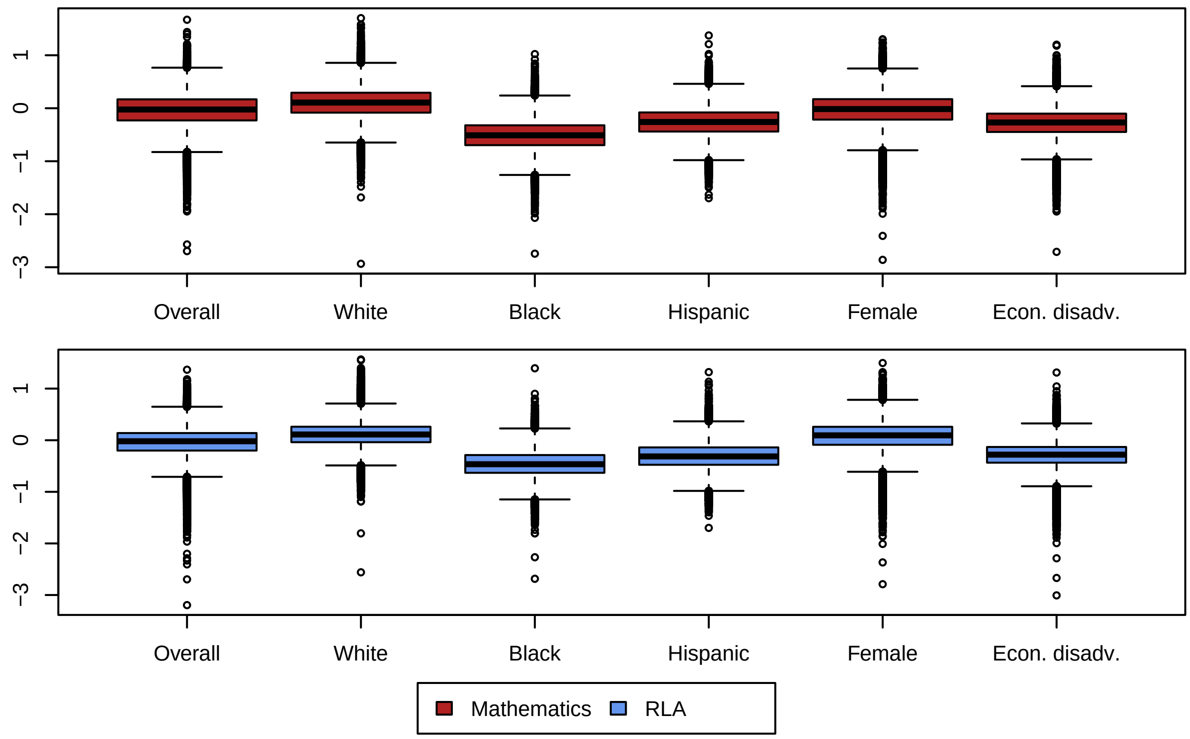

Figure 7: Boxplots of the outcomes of interest

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

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. While a disaster itself may be transient, the effects caused by the disaster may not be. 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

Table 4: Number of counties per relative period

	-9	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	9
FEMA Disasters	69	222	332	381	594	736	800	991	1490	2414	2345	2192	2082	2033	1820	1678	1614	1423	924
NWS Storms	32	110	175	191	248	275	387	446	474	602	518	440	375	359	302	275	163	104	76
FEMA Storms	92	223	381	391	437	721	787	946	1279	2035	1943	1812	1654	1644	1598	1314	1248	1089	756

fixed effects.

Consequently, the baseline specification is

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

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. E_i denotes the first year of treatment and

$$D_{i,t,l} := \mathbb{1}\{t - E_i = l\} = \begin{cases} 1 & t - E_i = l \\ 0 & \text{else.} \end{cases}$$

is a treatment indicator for county i l years after initial treatment. That is, it is 1 if the county had already experienced the first disaster l years ago. The set of counties which experience their first disaster in the same period, i.e. have the same E_i, are referred to as a cohort.

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 $l \leq -3$ and $l \geq 5$. This solves the collinearity problem.

Apart from avoiding the collinearity issue, there is also another reason for binning: Distant treatment leads and lags have fewer observations (see Table 4). In other words, few counties are first treated very late or and therefore only few counties experience many years before treatment. Similarly, fewer counties are treated very early and experience many years of treatment. In that sense, binning increases the sample size on the distant leads. This makes the estimates more precise and more robust to outliers. Thus, equation (1) turns into

$$y_{i,t,g} = \beta_{-3}\mathbb{1}\{t - E_i \leq -3\} + \sum_{l=-2, l\neq -1}^{4} \beta_l D_{i,t,l} + \beta_5 \mathbb{1}\{t - E_i \geq 5\} + \alpha_i + \lambda_t + \zeta_g + \varepsilon_{i,t,g} \ . \tag{2}$$

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), which will be explained below. A similar estimator was introduced by Callaway and Sant'Anna (2021). However, the latter is unable to handle multiple observations for the same unit-period combination. Since we have multiple grades for each county-year combination this would be a severe restriction in our setting. That is why, Sun and Abraham (2021) is better suited.

Heterogenous treatment effects also make it necessary to cluster standard errors (see Abadie et al., 2017). Following Sun and Abraham (2021), we cluster at the county 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:

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

   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: $\mathbb{E}[Y^\infty_{i,t} - Y^\infty_{i,s} | E_i = e]$ 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 for most cohorts.

No Anticipatory Behavior: There is no treatment effect prior to treatment, that is $\mathbb{E}[Y_{i,e+l} - Y^\infty_{i,e+l}] = 0$ 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.

3.4 Heat

For the effect of heat exposure we utilize a different design. A binary treatment indicator is not-well suited to capture the effect of cumulative heat exposure, as heat exposure is continous. Instead, we follow Park et al. (2020) in using the two aforementioned measures: average maximum daily temperature and days above 30°C[2]. These give rise to an interesting marginal interpretation: What is the effect of a 1°C hotter school year or of one additional day above 30°C on average test scores?

Again, conditional on location and year heat exposure should be independent of other factors determining academic performance. Also, there is no threat of endogenous selection into test-taking (as in Park et al., 2020), since all students are required to take these standardized tests. Thus, we estimate a simple linear model with county, year, and grade fixed effects,

$$y_{i,t,g} = \beta H_{i,t} + \alpha_i + \lambda_t + \zeta_g + \varepsilon_{i,t,g} \ , \tag{5}$$

[2]Of course, the threshold of 30°C is somewhat arbitrary. However, changing it slightly does not lead to meaningful differences in the results.

where $H_{i,t}$ is the measure of heat exposure at the county-schoolyear level. Since there are likely no confounding factors, a least squares estimate of β can be interpreted causally. A visual analysis of the residuals (available upon request) does not indicate any substantial deviations from a homoskedastic and diagonal covariance matrix. Therefore, we use classical standard errors without any adjustments.

4 Results

4.1 Main Results

Figure 8 shows estimated dynamic treatment effects and 95% confidence intervals for all students and the four subgroups of interest. Note that the precision of the estimates tends to decrease with the distance in time from treatment. Recall that the number of observed units decreases with the distance in time from treatment (see Table 4). 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 period of treatment there is a significant[3] effect of natural disasters on the performance in mathematics. Also, one year after the disaster there is an even larger significant decrease in test scores. For all subsequent periods the effect on mathematics scores is not significant. For RLA scores, on the other hand, we see significantly positive effects after three years. In other words, RLA test scores actually tend to improve three to five years after a natural disaster.

The estimated effects among the subgroups are relatively similar: Negative effects on mathematics test scores in the short term, but positive long-term effects in RLA. For black and hispanic students, however, we do not even find the negative effects on mathematics, but among black students there is a significant decrease in mathematics scores after five years.

The effect sizes for the short-term effect range from barely above zero to about -0.02 standard deviations of the national reference cohort. The positive long term effects go up to 0.05 standard deviations of the national reference cohort. Here, the effect on black and hispanic students seems to be larger than on white students. Also, female students seem to be more positively affected. Their test scores improve by between 0.02 and 0.04 standard deviations of the national reference cohort four or five years after the disaster.

The positive medium- and long-term effects may be surprising, but 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), who attributes them to student mobility after the disaster. Due to destroyed infrastructure, many students have to switch schools and some may even benefit from attending a higher quality school after the disaster. Thus, if a disaster leads to students switching from lower quality to higher quality schools, positive effects are very plausible.

Figures 9 shows the same graphs based on the storm treatment. Overall, we also find a significantly negative effect on mathematics test scores in the same year.

Similarly, white students experience a significant decline in mathematics scores in the same year, but they see an improvement in RLA scores three to five years after the disaster. For other subgroups, however, we do not find this effect in the storm data. For black and hispanic students, we find no significant effects, while for female and disadvantaged students we find a significant decrease in both mathematics and RLA.

[3] 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.

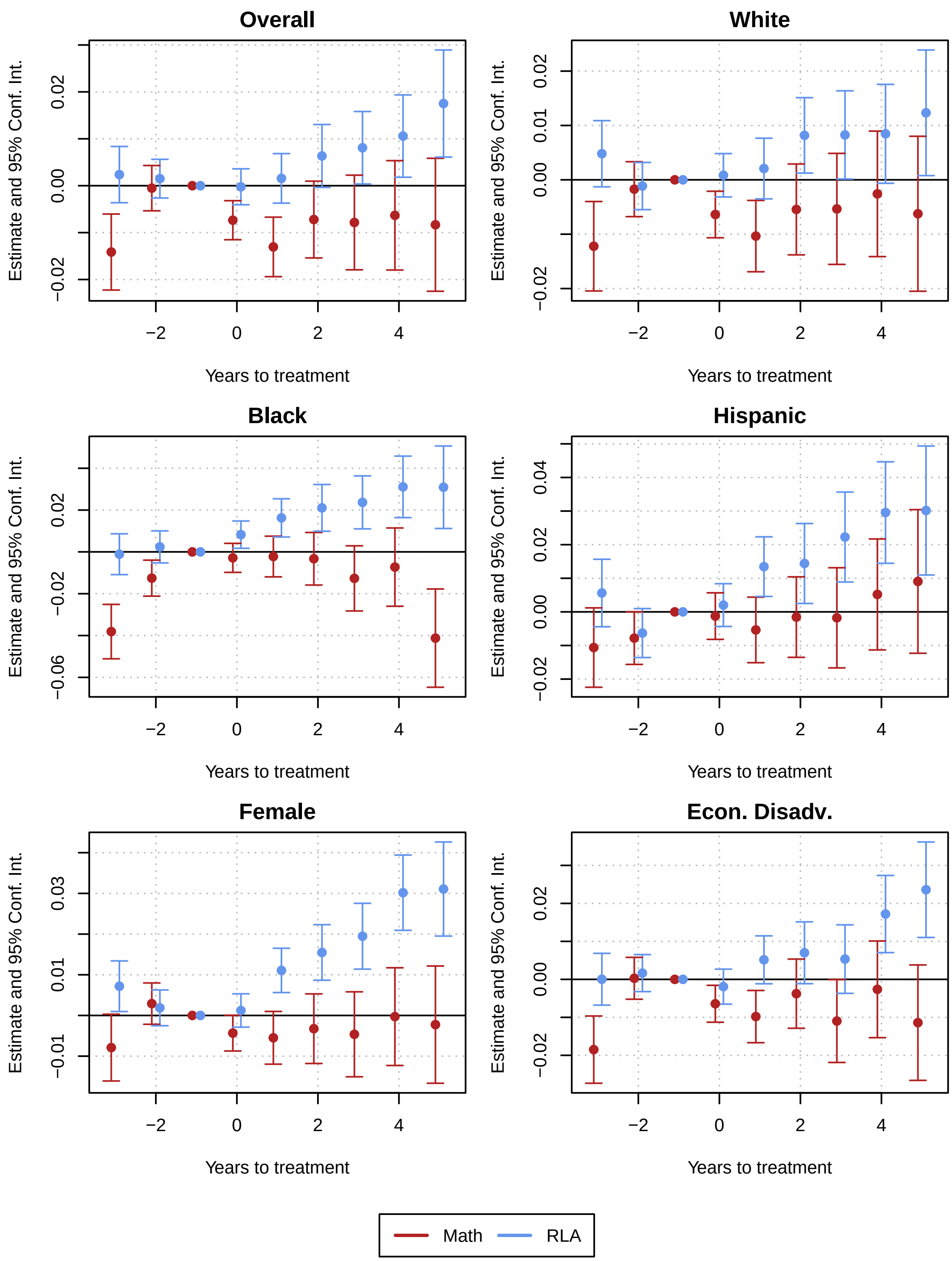

Figure 8: Dynamic Treatment effects in relative time: FEMA disaster data

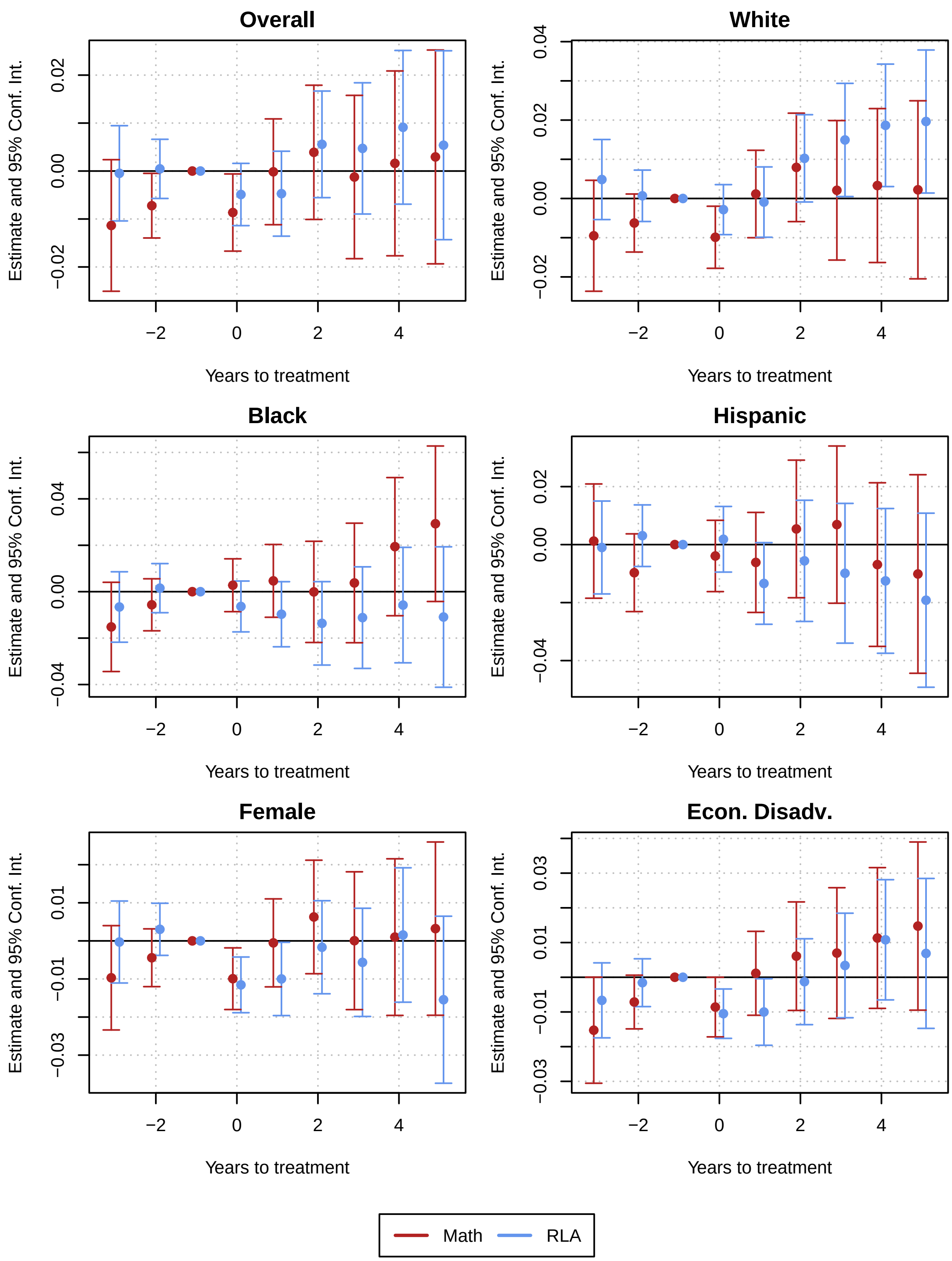

Figure 9: Dynamic Treatment effects in relative time: NWS storm data

4.2 Heat Results

Table 5 provides the estimation results for the heat models. Rows 1 and 2 show the estimated coefficients and standard errors for the maximum temperature variable for mathematics and RLA

Table 5: Estimated coefficients for heat models

	Overall	White	Black	Hispanic	Female	Econ. Disadv.
Max. Temp. (Math)	-0.00074^{***}	-0.00032	-0.0007	-0.0023^{***}	-0.00112^{***}	-0.00123^{***}
	(0.00034)	(0.00036)	(0.00072)	(0.0006)	(0.00037)	(0.00037)
Max. Temp. (RLA)	-0.00002	0.00056	-0.00124^{***}	-0.00102	-0.00015	-0.00036
	(0.00028)	(0.0003)	(0.00063)	(0.00053)	(0.00032)	(0.00032)
Days ab. 30 (Math)	-0.000157	-0.000096	0.000097	-0.00017	0.000002	0.000005
	(0.000087)	(0.000094)	(0.000138)	(0.000132)	(0.000093)	(0.000093)
Days ab. 30 (RLA)	-0.000091	-0.000165^{***}	-0.000116	-0.000447^{***}	-0.000195^{***}	-0.000027
	(0.00007)	(0.000077)	(0.000117)	(0.000113)	(0.000077)	(0.000077)
Mean	-0.042	0.107	-0.483	-0.281	0.025	-0.284

Note: Standard errors in parentheses. Stars indicate significance at a 5% level.

respectively. We find a significant effect of daily maximum temperature on mathematics scores. This is also significant and more pronounced among hispanic, female and, economically disadvantaged students. The effect on RLA is only significant among black students. Based on these point estimates, a year that is 5-10°C hotter has a similar effect as a natural disaster.

Rows 3 and 4 of Table 5 show the same results for the number of days above 30°C. Here, we only find a significant effect on RLA among white, hispanic and female students in RLA. For hispanic students the point estimate is -0.0004, meaning about 25 additional days above 30°C would be somewhat comparable to the average effect of a natural disaster in the same year, a rather small effect. For white and femal students, the effect size is even smaller.

In total, we find only limited evidence for a negative effect of heat on the achievement in standardized tests. However, it seems to be more pronounced among some subgroups. For hispanic students the point estimates are substantially larger when significant. This is somewhat consistent with other authors' findings (for example Park et al., 2020)

One explaining factor could be the presence of air-conditioning. The results in Park et al. (2020) suggest that air conditioning offsets about three quarters of the negative effect of heat exposure. Possibly, counties with higher shares of minority students tend to have a lower presence of air-conditioning in schools. Also, black and hispanic households tend to be poorer on average and may therefore be less likely to have air-conditioning at home. This could explain some of the heterogeneity in the effect of heat.

4.3 Discussion

It could be the case that medium- and long-term effects are largely driven by migration. This is a prominent theme in previous studies on disasters and learning (Pane et al., 2008; Sacerdote, 2012). That's why it may be interesting to have a look at the ethnic composition of the counties relative to initial treatment. Figure 10 shows ethnic shares of enrolled students for the treated counties in relative time.

In the left panel we see that the share of black students decreases in counties that experienced a disaster. This supports the hypothesis that black students disproportionately switch schools after disasters and may explain the positive long term effect as described above. However, the share of black students already decreases before treatment, so this may not be a causal effect of the disaster.

The same plot based on the NWS storm data shows a different picture. Here, all three shares remain somewhat constant until five years after treatment. Then the share of black students in-

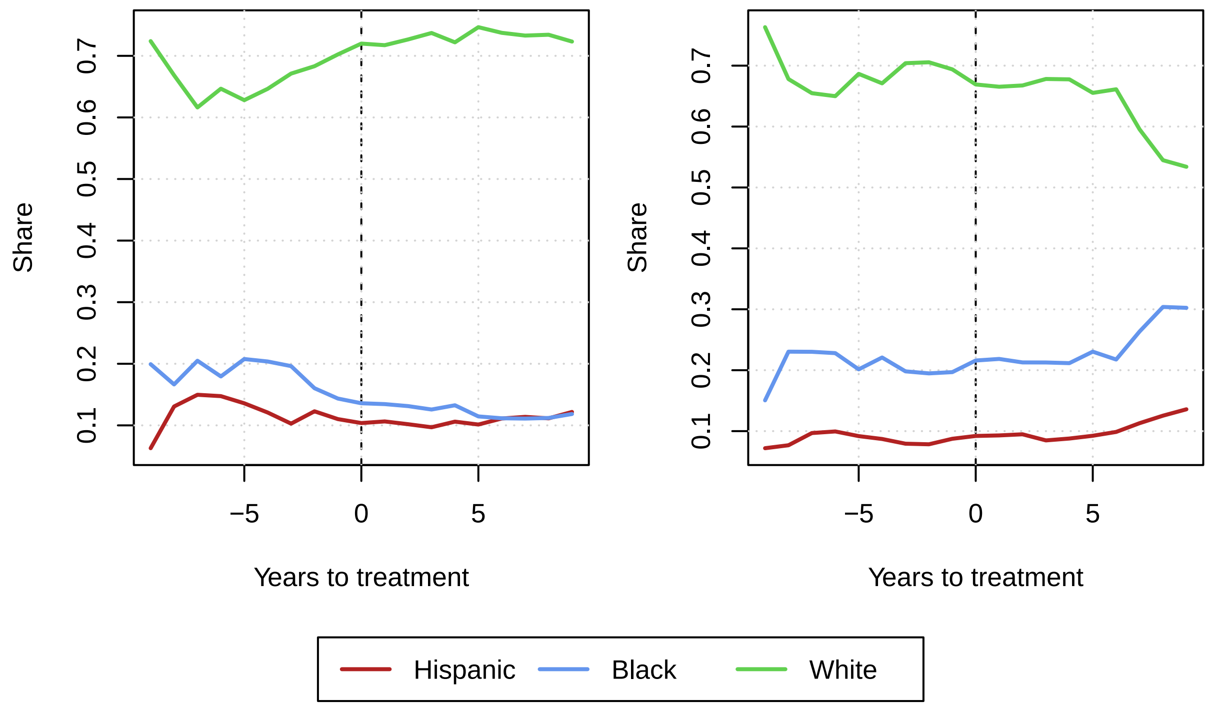

Figure 10: Aggregated ethnic shares of enrolled students by treatment timing based on FEMA disasters (left) and on NWS storms (right)

creases, while the share of white students decreases. This suggests that the migration response to storms may be qualitatively different than the response to other forms of disasters. At least for the storms data this does not seem to be a major driver of the results.

A more in-depth analysis of migration responses to natural disasters and their role in academic achievement is unfortunately not possible with this data. There is some prior research based on individual level data indicating that it does play an important role (for example Sacerdote, 2012). Analyzing migration responses and their link to academic achievement for different types of natural disasters may be a promising area for future research.

For the heat results, we find substantial heterogeneity: Minorities seem to be more prone to adverse effects of heat exposure. The results in Park et al. (2020) indicate that air-conditioning mitigates a large part of the negative effect. Based on that, we speculate that inequality in air-conditioning is the main driver of said heterogeneity, since minority students tend to have less access to air-conditioning. Improving the air-conditioning coverage in more affected schools could be an easy way for policymakers to adress the adverse effects of heat on academic achievement. However, to the best of our knowledge Park et al. (2020) is the only paper to study this relationship. More research on the interconnection between heat, learning, and air-conditioning may be useful to make policy decisions.

Obviously, our results come with limitations. First, we find significant pre-treatment effects. Therefore, the treatment led to a significant difference in average test scores between treatment and control group before the treatment even began. This is likely caused by a violation of one of the two identifiying assumptions: Either the trends in outcomes of treatment and control group are not parallel or treated units anticipated treatment and therefore already saw an effect earlier. Since the latter is rather implausible, we conclude that there might be a violation of the parallel trends assumption, despite the well behaved pre-treatment trends (see Appendix B). Thus, the affected results have to be interpreted with caution.

Based on the results from Section 2, in particular Figure 6, it seems that there are significant differences between counties that apply for (and receive) federal assistance after a disaster and those that do not. If such federal assistance offsets the negative effect on the education system, our estimates could be biased. However, this would suggest that the effects could be even larger in the absence of federal aid. Therefore, our estimates may only provide a lower bound on the true effect, which could be much larger.

For the heat results, the measures of heat exposure used are based on daily maximum temperature only. However, this likely does not fully capture the physiological impact of heat. Other variables, like humidity or minimum temperatures overnight, can also contribute to the negative effect of heat on learning (for an extensive discussion of heat exposure measurement see Rennie et al., 2021). It would be desirable to use a more complete measure of heat exposure. The reason why we nevertheless used only the daily maximum temperature is that the data availability for more comprehensive measures is much worse. A reasonable option would be to use gridded satellite data, such as those provided by the Copernicus project[4]. Unfortunately, these datasets are huge and require much more computational power.

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, that is they only appear in one of the datasets used.

Based on temperature data, we also analyze the effect of heat exposure on academic performance. For the overall student population, we do not find a significant effect. However, for hispanic and to a weaker degree also black and economically disadvantaged students we find significantly negative effects. These discrepancies are likely driven by differing access to air-conditiong.

Mitigating such negative effects should be a concern for policymakers. Even if effect sizes are small, such negative effects can quickly compound in regions that are frequently affected by disasters.

[4] Available here

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. More specifically, the applicant variable is 1 if the county experienced a disaster and applied for dederal assistance, that is itappears in the Public Assistance Applicants Program Deliveries database, and 0 otherwise. This is regressed on a few covariates, including the share of democratic votes in the 2016 election. All independent variables are as of 2016.

Table 6: Determinants of Assistance Application

Dependent Variable:	Applied
Model:	(1)
Variables	
(Intercept)	-26.13***
	(2.053)
Share of democratic voters (2016)	-1.712***
	(0.1700)
Median Income (logs)	2.197***
	(0.1824)
Poverty Rate	10.93***
	(0.8545)
Share of single mothers	5.153***
	(0.6200)
Fit statistics	
Observations	9,894
Squared Correlation	0.05322
Pseudo R^2	0.03890
BIC	13,004.3

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

A.2 Results for storms based on FEMA data

Below we repeat the main analysis based on the FEMA data only considering storms. That is, all disasters that are categorized as Tornadoes, Hurricanes, or Severe Storms. The results are shown in Figure 11.

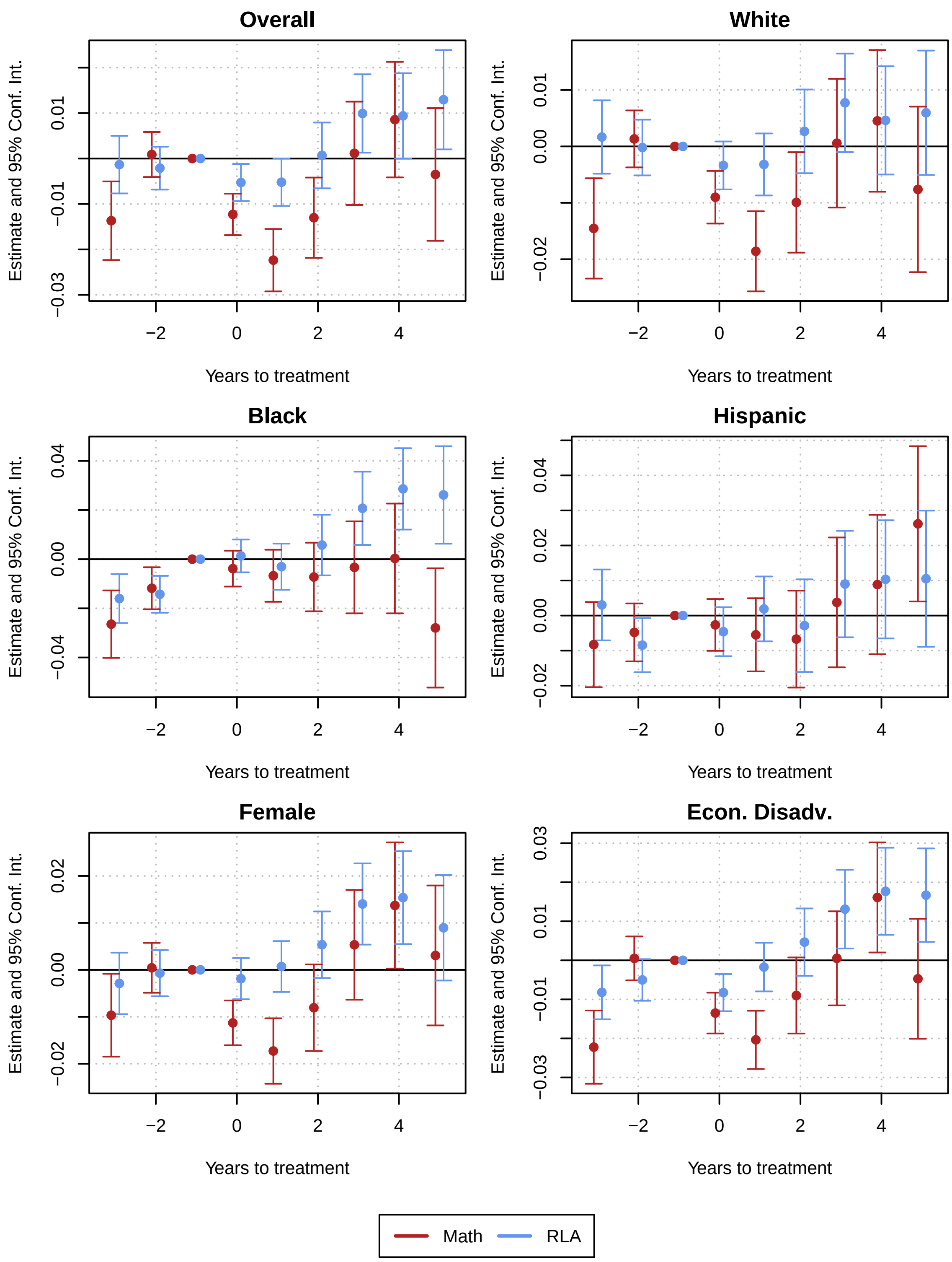

Figure 11: Dynamic Treatment effects in relative time: FEMA data (storms only)

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.

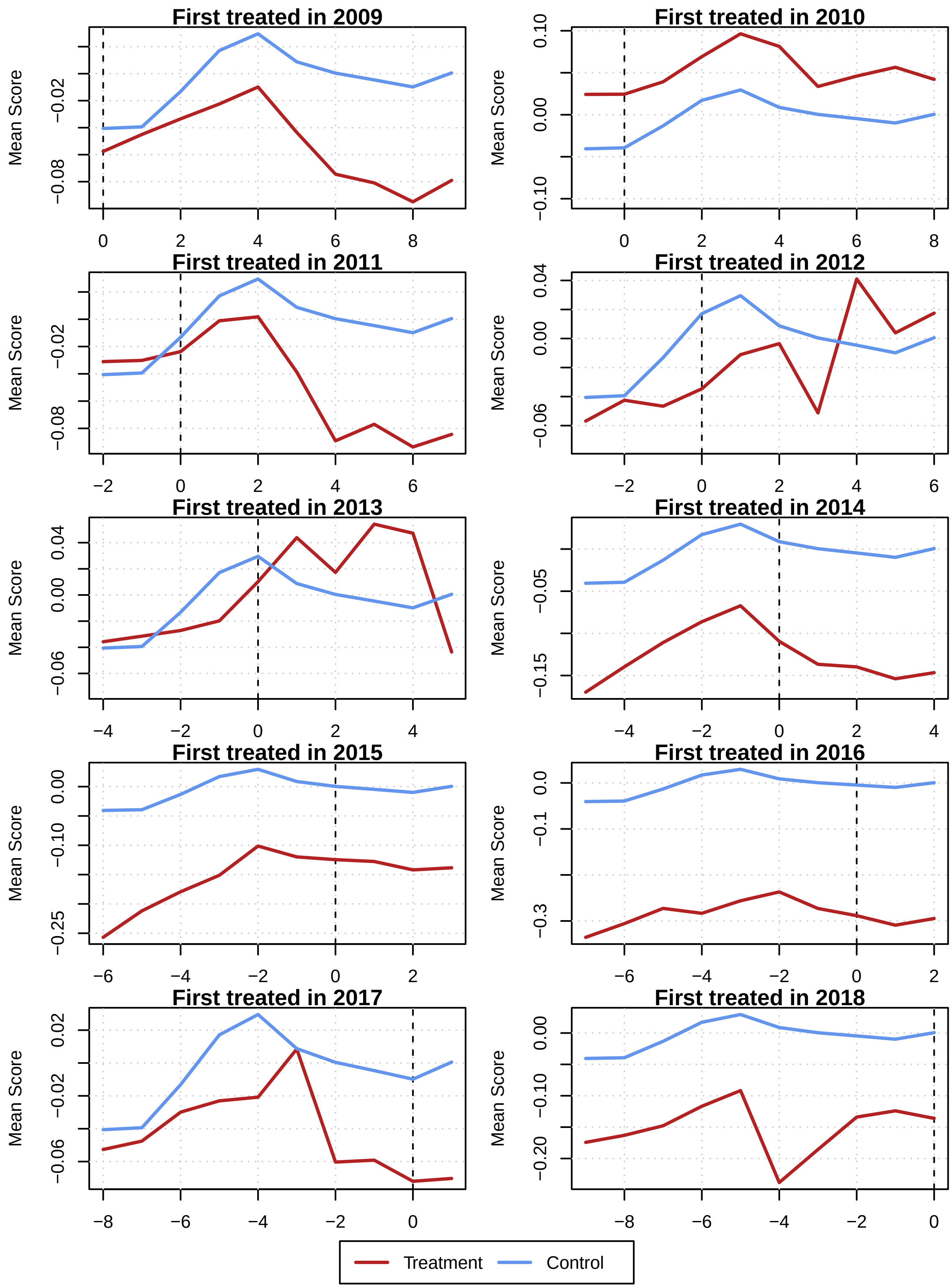

Figure 12: Aggregated mean scores in mathematics based on FEMA data in relative time to treatment

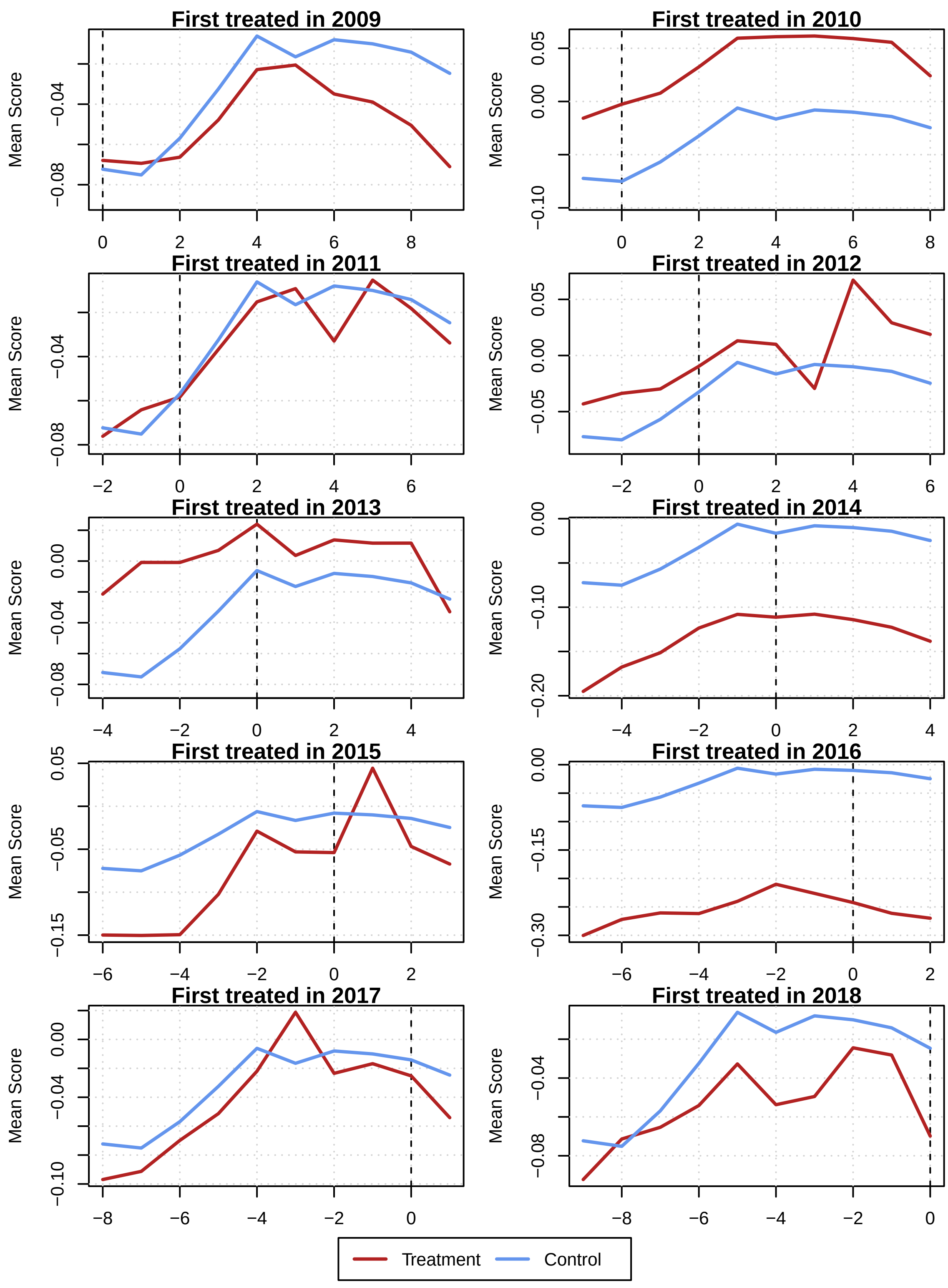

Figure 13: Aggregated mean scores in RLA based on FEMA data in relative time to treatment

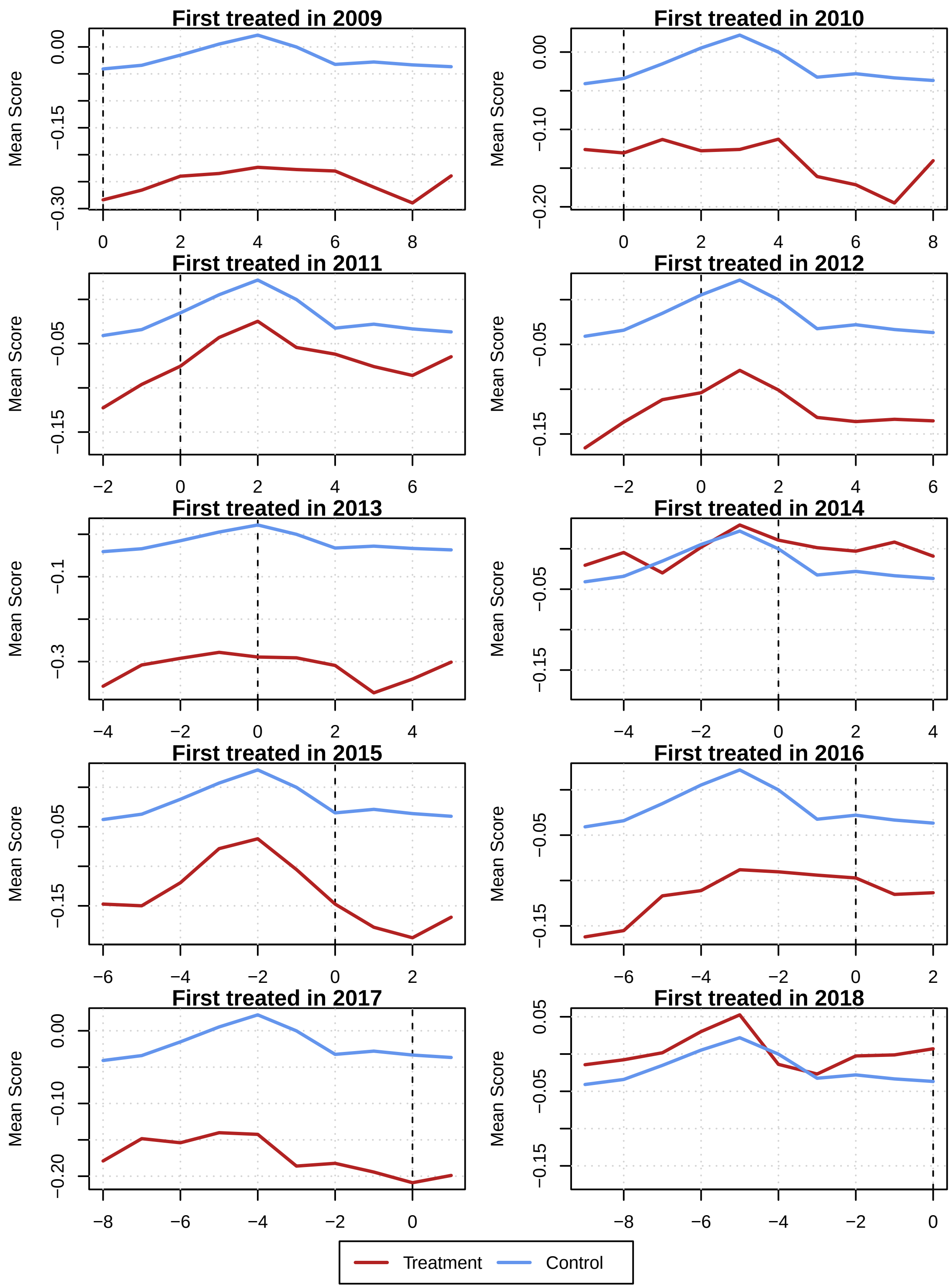

Figure 14: Aggregated mean scores in mathematics based on NWS storm data in relative time to treatment

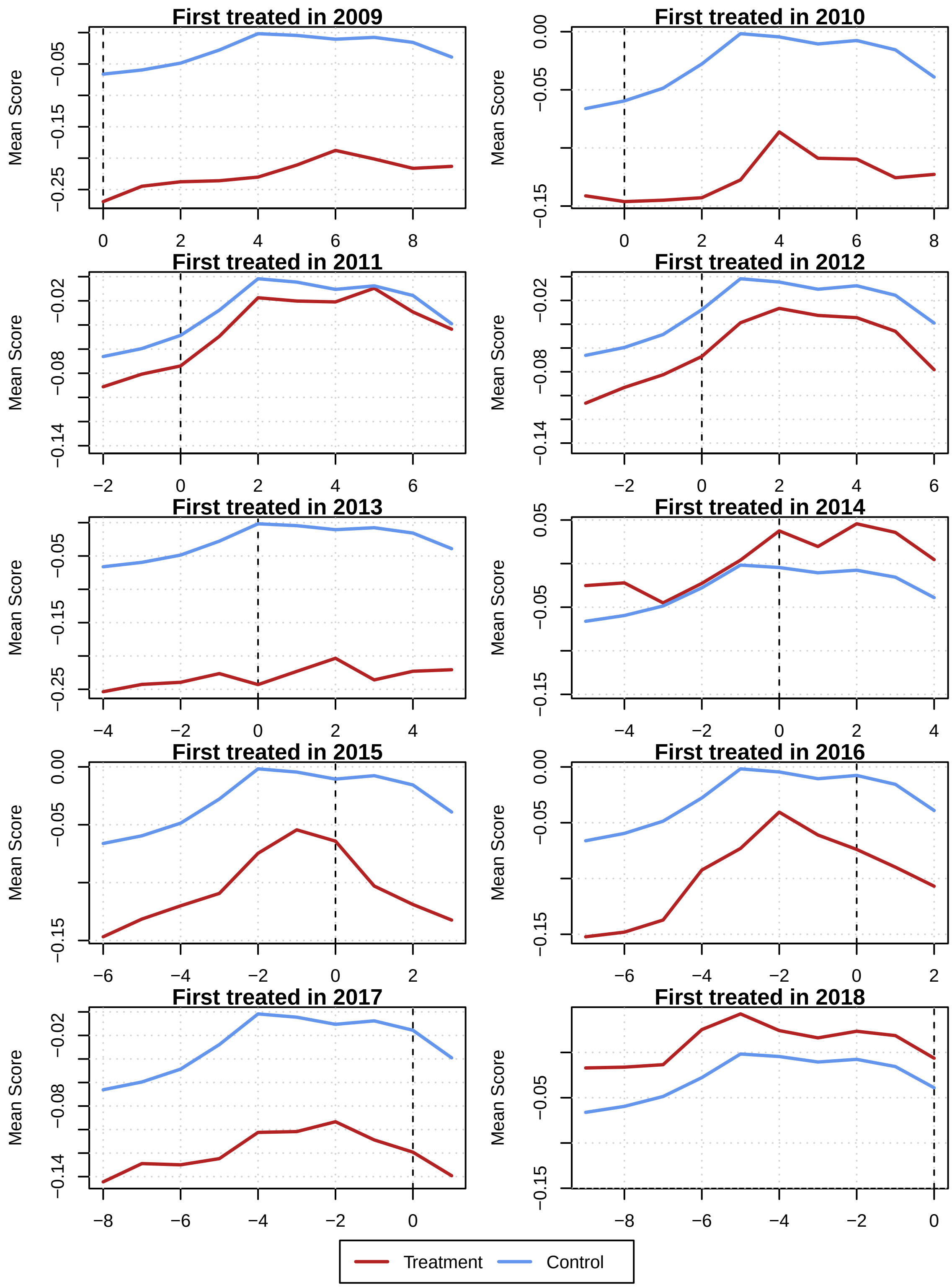

Figure 15: Aggregated mean scores in RLA based on NWS storm data in relative time to treatment

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*.

Boustan, L. P., Kahn, M. E., Rhode, P. W., and Yanguas, M. L. (2020). The effect of natural disasters on economic activity in us counties: A century of data. *Journal of Urban Economics*, 118:103257.

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

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.

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 (2021). *Climate Change 2021: The Physical Science Basis.* Cambridge University Press.

Lamb, J., Gross, S., and Lewis, M. (2013). The hurricane katrina effect on mathematics achievement in mississippi. *School Science and Mathematics*, 113(2):80–93.

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

Menne, M. J., Durre, I., Vose, R. S., Gleason, B. E., and Houston, T. G. (2012). An overview of the global historical climatology network-daily database. *Journal of atmospheric and oceanic technology*, 29(7):897–910.

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. (2022). Hot temperature and high-stakes performance. *Journal of Human Resources*, 57(2):400–434.

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).

Rennie, J. J., Palecki, M. A., Heuser, S. P., and Diamond, H. J. (2021). Developing and validating heat exposure products using the u.s. climate reference network. *Journal of Applied Meteorology and Climatology*, 60(4):543 – 558.

Robert T. Stafford Disaster Relief and Emergency Assistance Act (1988).

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*.

United States Government Accountability Office (2022). Disaster recovery: School districts in socially vulnerable communities faced heightened challenges after recent natural disasters.

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