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

Gregor Steiner

May 6, 2022

**Abstract**

# 1 Introduction

A causal effect of natural disasters may be driven by school closures (Grewenig et al., 2021), destroyed infrastructure, and emotional stress (Vogel and Schwabe, 2016).

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.

This article contributes to a rich literature on economic effects of natural disasters. Many authors have focused on one specific type of natural disasters, e.g. hurricanes (Deryugina, 2017; Deryugina et al., 2018).

More specifically, this paper contributes to the literature on the impact of natural disasters on education. Some authors have investigated the link between natural disasters and human capital formation from a growth theory perspective, e.g. Crespo Cuaresma (2010) finds a strong negative effect of geologic natural disaster risk on secondary school enrolment rates.

Many authors working with standardized test data have focused on one particular type of disaster. There is extensive evidence for a negative effect of heat exposure on learning (e.g. Park et al., 2020b,a). Spencer et al. (2016) find a negative effect of hurricanes on performance in the sciences.

## 2 Data

### 2.1 Natural Disaster Data

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. It seems that the variation in the number of declared disasters may be driven by the governor's proactiveness in requesting a declaration. Thus, it could be interesting to compare counties on different sides of state borders, whose actual disaster exposure is likely very similar in order to analyze the effect of a declaration.

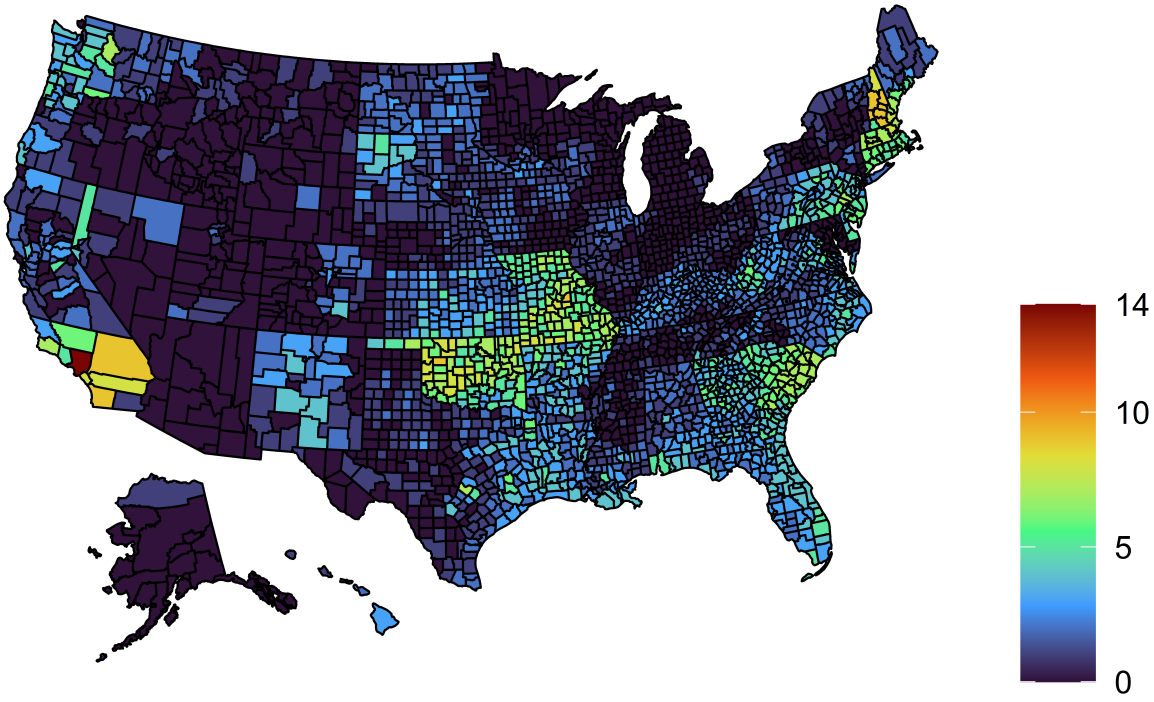


Figure 1: Number of declared natural disasters from 2009 to 2018

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 and figure 4 show boxplots and kernel density estimates 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

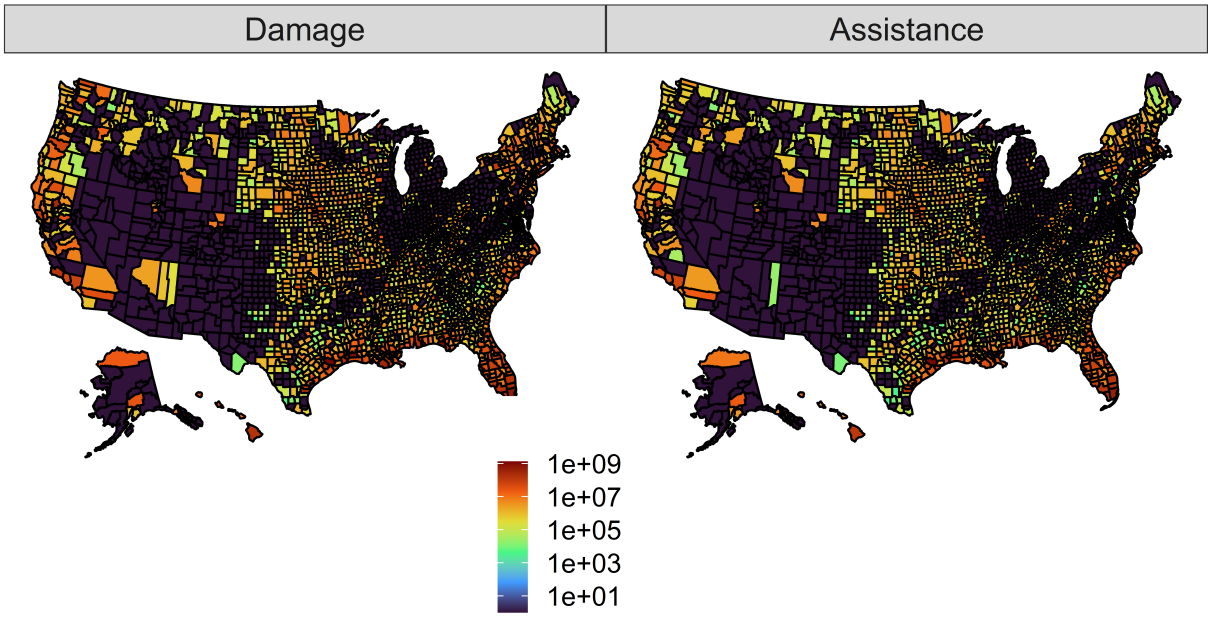


Figure 2: Amount of disaster damage and federal disaster assistance (in USD) awarded to counties since October 2016

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 with the start of the Public Assistance Applicants Program Deliveries dataset) does not seem to be different in the two groups.

### 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

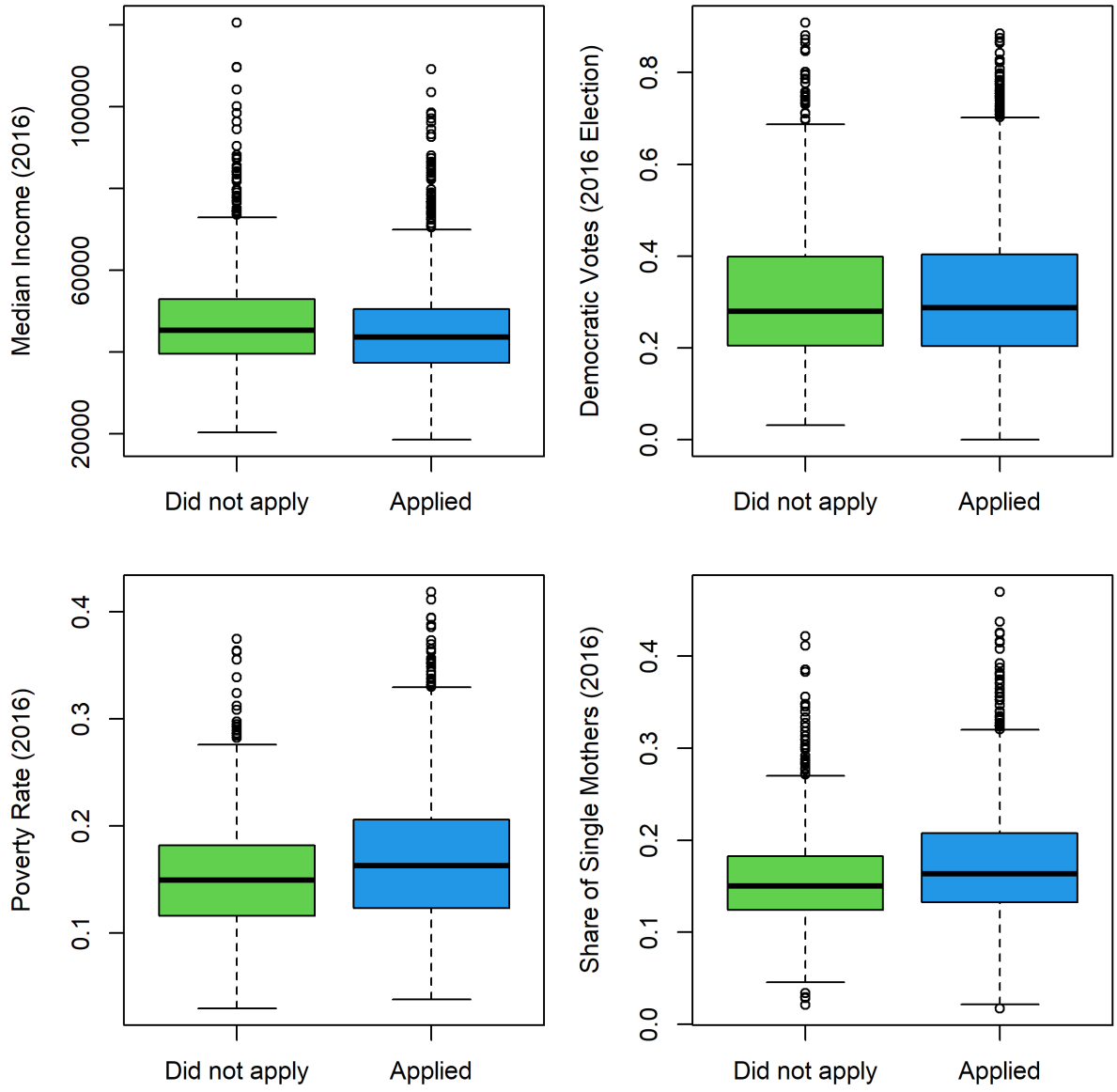


Figure 3: Boxplots by application status

poverty rate.

### 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.

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

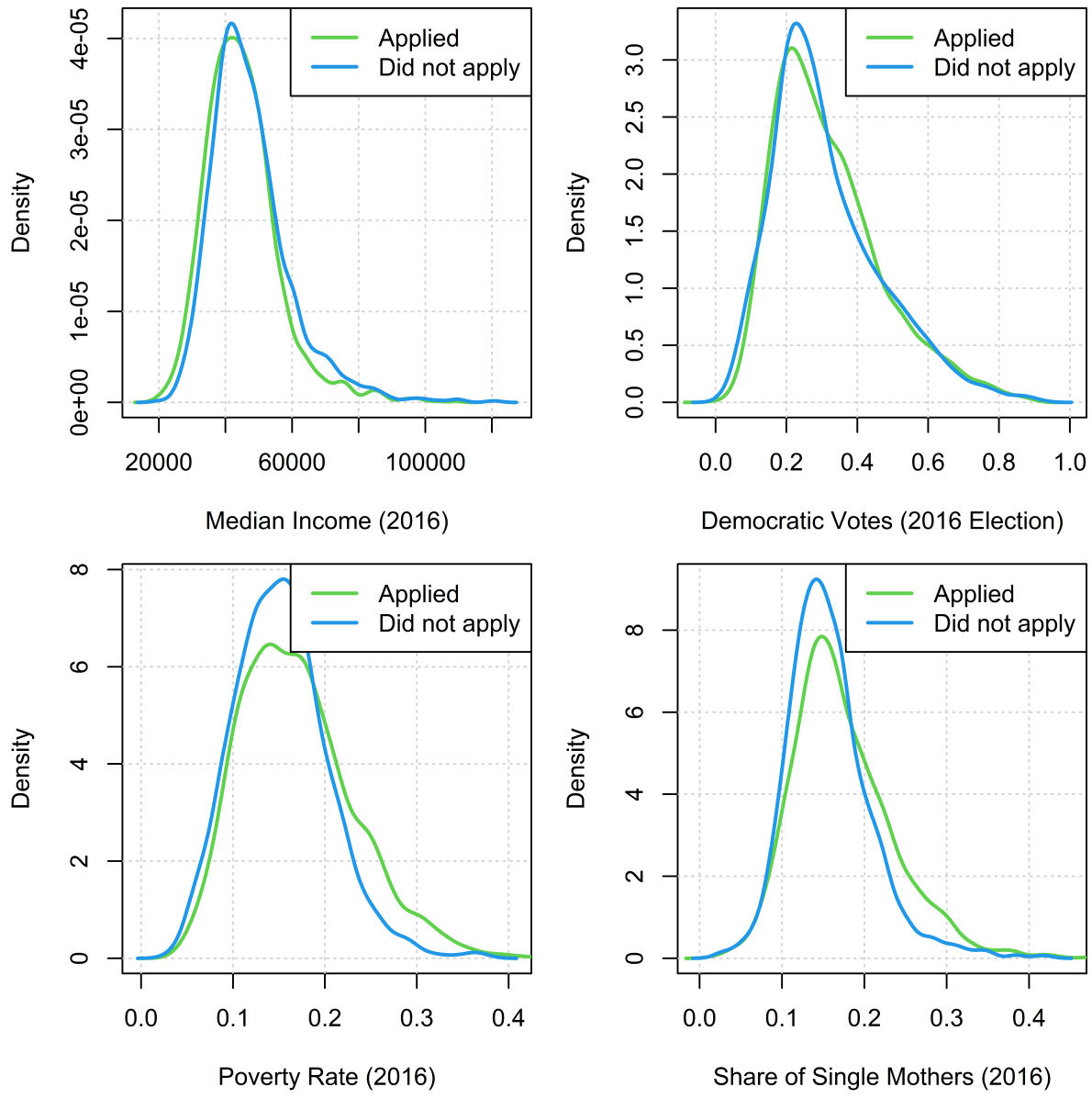


Figure 4: Kernel density estimates by application status

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. Table 1 shows summary statistics and 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.

Table 1: Summary Statistics

| Variable | N | Mean | Std. Dev. | Min | Pctl. 25 | Pctl. 75 | Max |
|---|---|---|---|---|---|---|---|
| Disasters | 336415 | 0.221 | 0.566 | 0 | 0 | 0 | 6 |
| Treatment | 336415 | | | | | | |
| ... 0 | 150685 | 44.8% | | | | | |
| ... 1 | 185730 | 55.2% | | | | | |
| Subject | 338349 | | | | | | |
| ... Mathematics | 165352 | 48.9% | | | | | |
| ... RLA | 172997 | 51.1% | | | | | |
| Mean test score | 327358 | -0.042 | 0.294 | -3.196 | -0.214 | 0.152 | 1.669 |
| Mean test score (black students) | 136886 | -0.483 | 0.273 | -2.745 | -0.662 | -0.304 | 1.394 |
| Mean test score (hispanic students) | 144303 | -0.281 | 0.266 | -1.699 | -0.46 | -0.108 | 1.374 |
| Mean test score (female students) | 310350 | 0.025 | 0.295 | -2.862 | -0.153 | 0.222 | 1.496 |
| Mean test score (economically disadvantaged students) | 305395 | -0.284 | 0.256 | -3.007 | -0.441 | -0.118 | 1.312 |

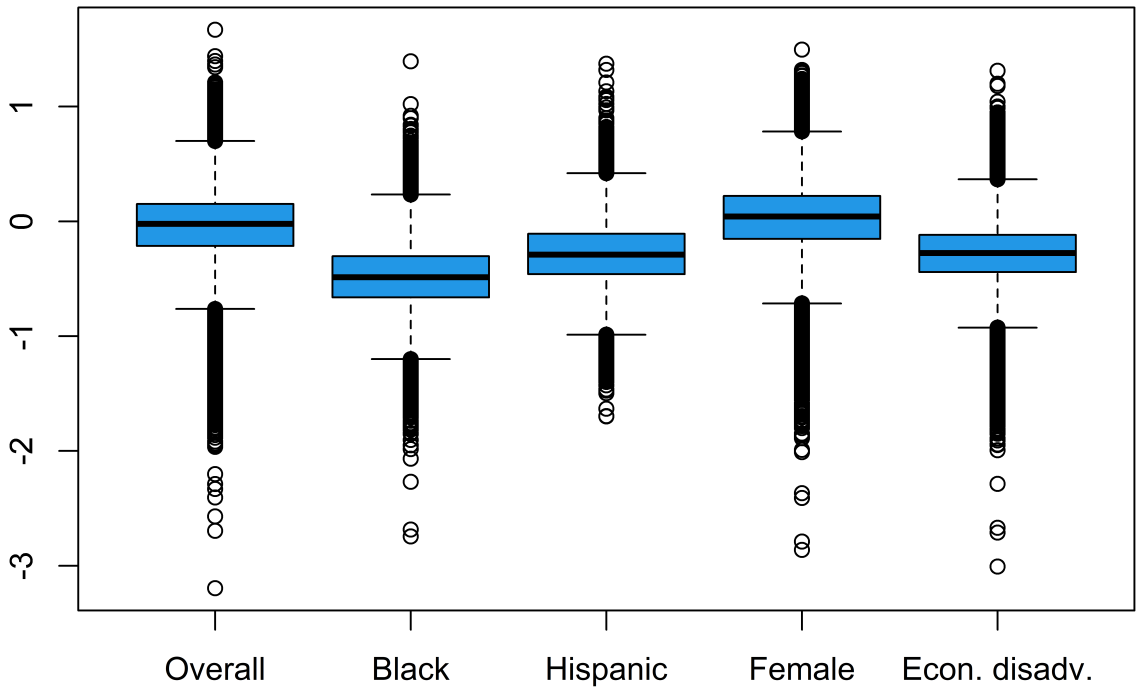


Figure 5: Boxplots of the outcomes of interest

# 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 $(D_{i,t})_{t=1}^{T}$ 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

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

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 $\alpha_i$, $\lambda_t$, and $\zeta_g$ respectively and $X_{i,t}$ is a row vector of additional control variables. $D_{i,t-l}$ 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

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

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.

### 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 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.

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).

**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.

## 4 Results

Figure 6 shows the dynamic treatment effects.

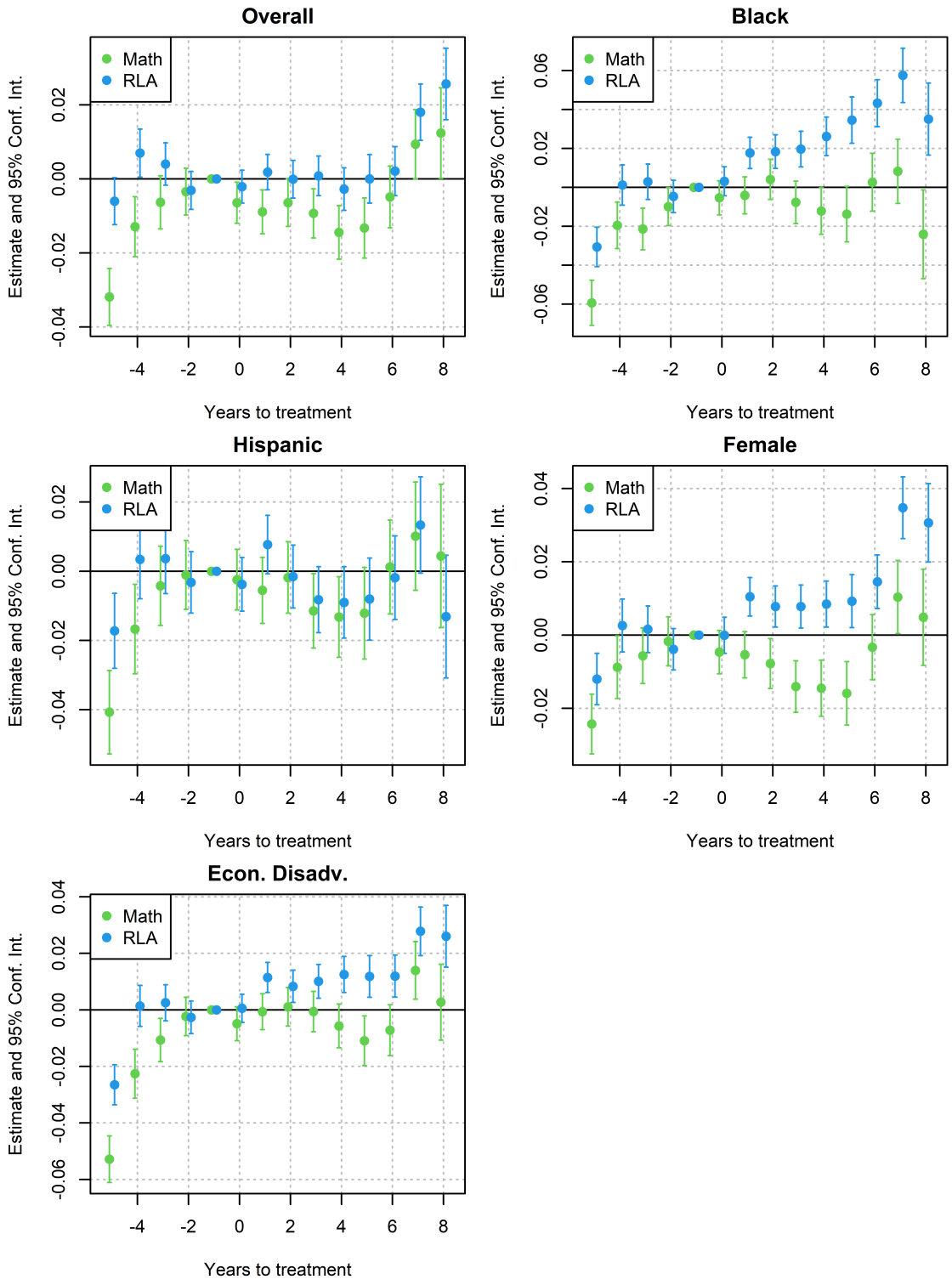


Figure 6: Dynamic Treatment effects by dependent variable and subject

## 5 Conclusion

## References


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 preprint 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.

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.

Park, R. J., Behrer, A. P., and Goodman, J. (2020a). Learning is inhibited by heat exposure, both internationally and within the united states. *Nature Human Behaviour*, 5(1):19–27.

Park, R. J., Goodman, J., Hurwitz, M., and Smith, J. (2020b). 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*.

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

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).