Problem set 4: The ETIC and diff-in-diff

Answer key - PMAP 8521, Spring 2021

March 8, 2021

Contents

In 1996, Nada Eissa and Jeffrey B. Liebman published a now-classic study on the effect of the Earned Income Tax Credit (EITC) on employment. The EITC is a special tax credit for low income workers that changes depending on (1) how much a family earns (the lowest earners and highest earners don't receive a huge credit, as the amount received phases in and out), and (2) the number of children a family has (more kids = higher credit). See this brief explanation for an interactive summary of how the EITC works.

Eissa and Liebman's study looked at the effects of the EITC on women's employment and wages after it was initially substantially expanded in 1986. The credit was expanded substantially agin in 1993. For this problem set, you'll measure the causal effect of this 1993 expansion on the employment levels and annual income for women.

A family must have children in order to quality for the EITC, which means the presence of 1 or more kids in a family assigns low-income families to the EITC program (or "treatment"). We have annual data on earnings from 1991–1996, and because the expansion of EITC occurred in 1993, we also have data both before and after the expansion. This treatment/control before/after situation allows us to use a difference-in-differences approach to identify the causal effect of the EITC.

The dataset I've provided (`eitc.dta`) is a Stata data file containing more than 13,000 observations. This is non-experimental data—the data comes from the US Census's Current Population Survey (CPS) and

includes all women in the CPS sample between the ages of 20–54 with less than a high school education between 1991–1996. There are 11 variables:

- `state`: The woman's state of residence. The numbers are Census/CPS state numbers: http://unionstats.gsu.edu/State_Code.htm
- `year`: The tax year
- `urate`: The unemployment rate in the woman's state of residence
- `children`: The number of children the woman has
- `nonwhite`: Binary variable indicating if the woman is not white (1 = Hispanic/Black)
- `finc`: The woman's family income in 1997 dollars
- `earn`: The woman's personal income in 1997 dollars
- `age`: The woman's age
- `ed`: The number of years of education the woman has
- `unearn`: The woman's family income minus her personal income, in *thousands* of 1997 dollars

```
library(tidyverse)  # For ggplot, %>%, mutate, filter, group_by, and friends
library(haven)      # For loading data from Stata
library(broom)      # For showing models as data frames

# This turns off this message that appears whenever you use summarize():
# `summarise()` ungrouping output (override with `.groups` argument)
options(dplyr.summarise.inform = FALSE)

# Load EITC data
eitc <- read_stata("data/eitc.dta") %>%
  # case_when() is a fancy version of ifelse() that takes multiple conditions
  # and outcomes. Here, we make a new variable named children_cat(egorical)
  # with three different levels: 0, 1, and 2+
  mutate(children_cat = case_when(
    children == 0 ~ "0",
    children == 1 ~ "1",
    children >= 2 ~ "2+"
  ))
```

1. Exploratory data analysis

Create a new variable that shows if women have 0 children, 1 child, or 2+ children (I did this for you already above).

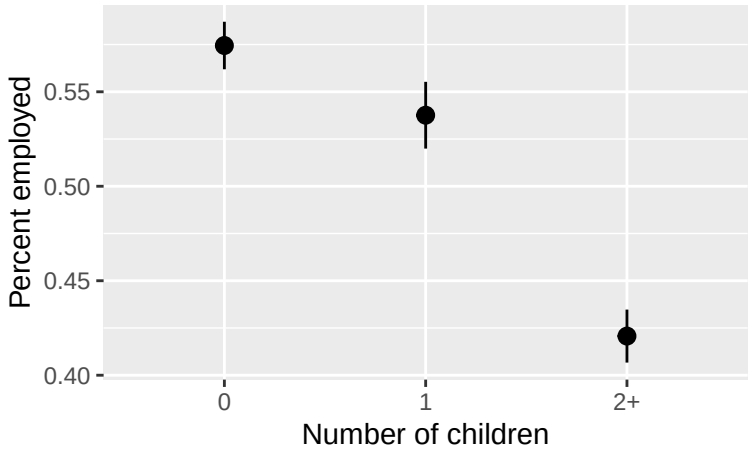
Visualize the average of `work`, `finc`, `earn`, `nonwhite`, `ed`, and `age` across each of these different levels of children.

Work

```
# Work
eitc %>%
  group_by(children_cat) %>%
  summarize(avg_work = mean(work))
```

children_cat	avg_work
0	0.574
1	0.538
2+	0.421

```
# stat_summary() here is a little different from the geom_*() layers you've seen
# in the past. stat_summary() takes a function (here mean_se()) and runs it on
# each of the children_cat groups to get the average and standard error. It then
# plots those with geom_pointrange. The fun.args part of this lets us pass an
# argument to mean_se() so that we can multiply the standard error by 1.96,
# giving us the 95% confidence interval
ggplot(eitc, aes(x = children_cat, y = work)) +
  stat_summary(geom = "pointrange", fun.data = "mean_se", fun.args = list(mult = 1.96)) +
  labs(x = "Number of children", y = "Percent employed")
```

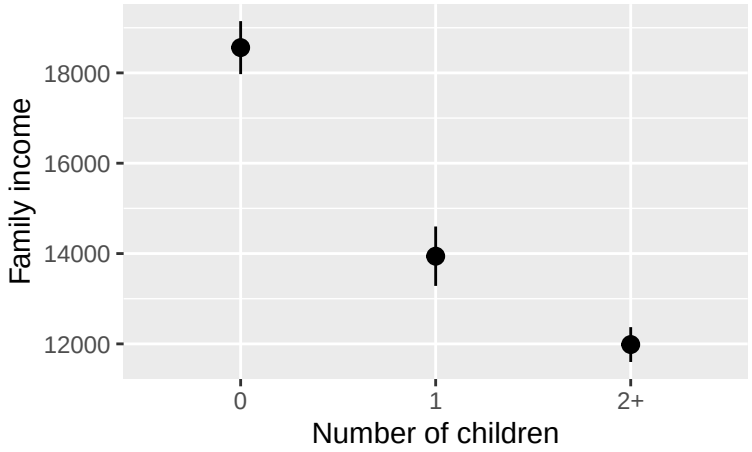


Family income

```
eitc %>%
  group_by(children_cat) %>%
  summarize(avg_work = mean(finc))
```

children_cat	avg_work
0	18560
1	13942
2+	11985

```
ggplot(eitc, aes(x = children_cat, y = finc)) +
  stat_summary(geom = "pointrange", fun.data = "mean_se", fun.args = list(mult = 1.96)) +
  labs(x = "Number of children", y = "Family income")
```

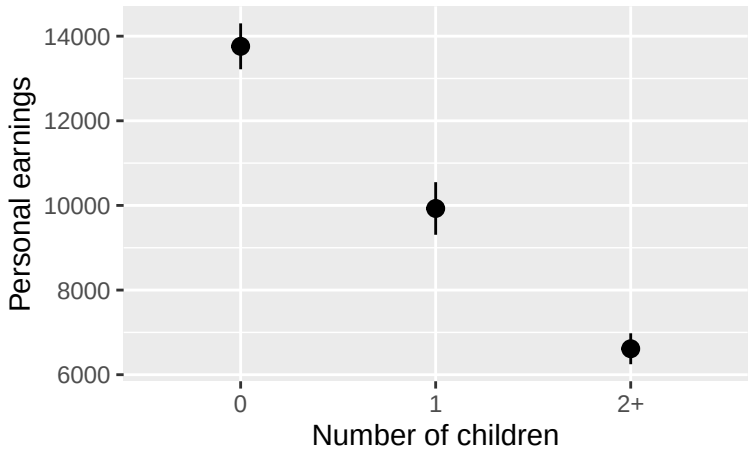


Earnings

```
eitc %>%
  group_by(children_cat) %>%
  summarize(avg_work = mean(earn))
```

children_cat	avg_work
0	13760
1	9928
2+	6614

```
ggplot(eitc, aes(x = children_cat, y = earn)) +
  stat_summary(geom = "pointrange", fun.data = "mean_se", fun.args = list(mult = 1.96)) +
  labs(x = "Number of children", y = "Personal earnings")
```

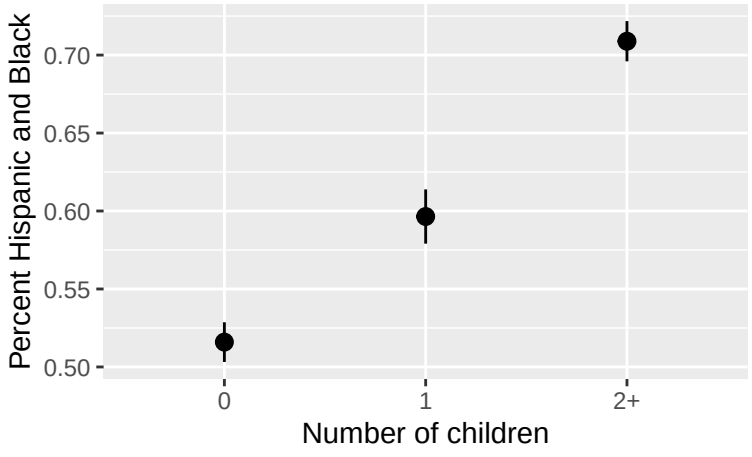


Race

```
eitc %>%
  group_by(children_cat) %>%
  summarize(avg_work = mean(nonwhite))
```

children_cat	avg_work
0	0.516
1	0.596
2+	0.709

```
ggplot(eitc, aes(x = children_cat, y = nonwhite)) +
  stat_summary(geom = "pointrange", fun.data = "mean_se", fun.args = list(mult = 1.96)) +
  labs(x = "Number of children", y = "Percent Hispanic and Black")
```

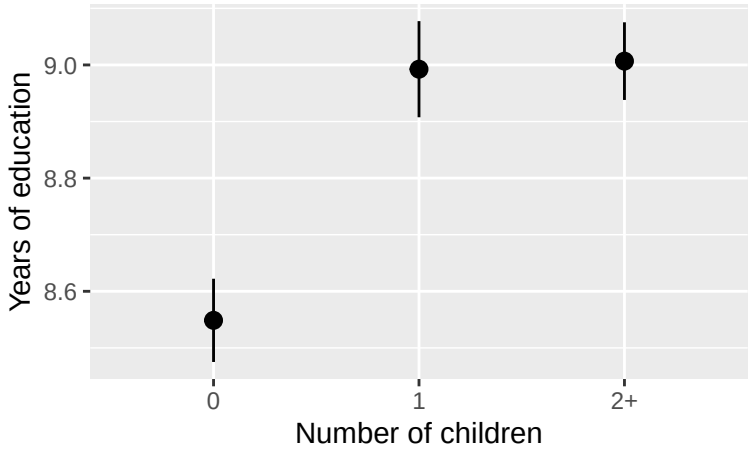


Education

```
eitc %>%
  group_by(children_cat) %>%
  summarize(avg_work = mean(ed))
```

children_cat	avg_work
0	8.55
1	8.99
2+	9.01

```
ggplot(eitc, aes(x = children_cat, y = ed)) +
  stat_summary(geom = "pointrange", fun.data = "mean_se", fun.args = list(mult = 1.96)) +
  labs(x = "Number of children", y = "Years of education")
```

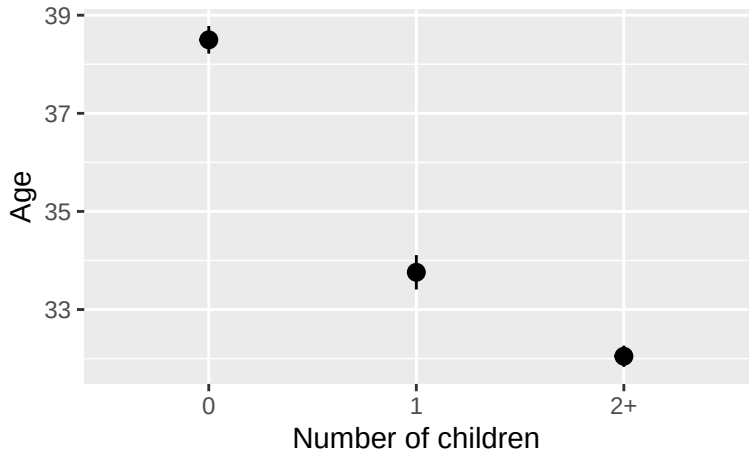


Age

```
eitc %>%
  group_by(children_cat) %>%
  summarize(avg_work = mean(age))
```

children_cat	avg_work
0	38.5
1	33.8
2+	32.0

```
ggplot(eitc, aes(x = children_cat, y = age)) +
  stat_summary(geom = "pointrange", fun.data = "mean_se", fun.args = list(mult = 1.96)) +
  labs(x = "Number of children", y = "Age")
```



General summary

Describe your findings in a paragraph. How do these women differ depending on the number of kids they have?

There are substantial differences between the demographics and outcomes of the women in this data, depending on how many kids they have. Those without children are more likely to be employed, have higher family and personal incomes, more likely to be white, and be older. Curiously, those without children have slightly less education than those that do, but only a difference of 0.4 years (or 5 months). Among those with kids, those who have two kids tend to be far less likely to be employed, have less family and personal income, are far more likely to be Black or Hispanic, and are slightly younger than those who have one child.

2. Create treatment variables

Create a new variable for treatment named `any_kids` (should be TRUE or 1 if `children` > 0) and a variable for the timing named `after_1993` (should be TRUE or 1 if `year` > 1993).

```
eitc <- eitc %>%
  mutate(any_kids = children >= 1,
         after_1993 = year > 1993)
```

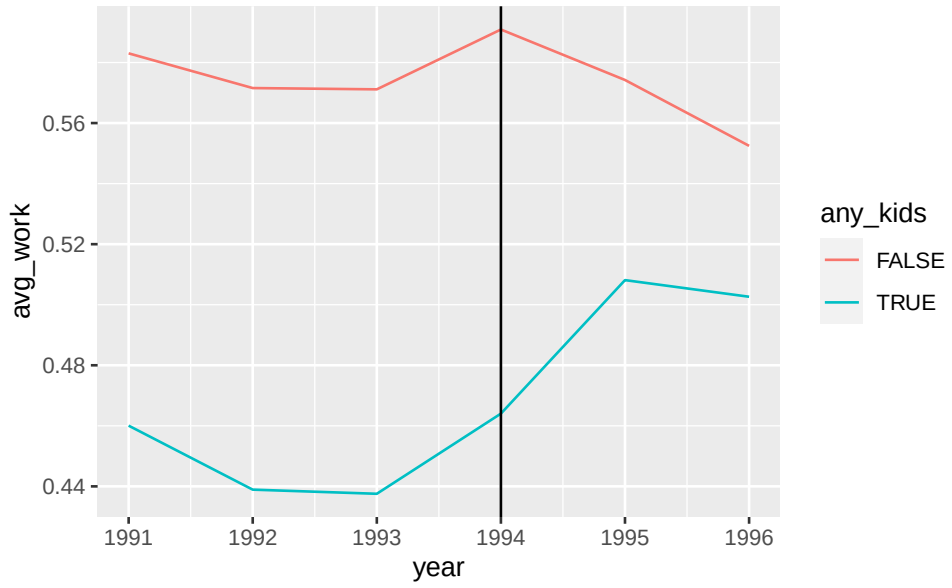
3. Check pre- and post-treatment trends

Create a new dataset that shows the average proportion of employed women (`work`) for every year in both the treatment and control groups (i.e. both with and without kids). (Hint: use `group_by()` and `summarize()`, and group by both `year` and `any_kids`.)

```
eitc_by_year <- eitc %>%
  group_by(year, any_kids) %>%
  summarize(avg_work = mean(work))
```

Plot these trends using colored lines and points, with year on the x-axis, average employment on the y-axis. Add a vertical line in at 1994 (hint: use `geom_vline(xintercept = SOMETHING)`.

```
ggplot(eitc_by_year, aes(x = year, y = avg_work, color = any_kids)) +
  geom_line() +
  geom_vline(xintercept = 1994)
```



Do the pre-treatment trends appear to be similar?

The pre-treatment trends appear to be similar. Those without kids are more likely to work, but both groups saw a decrease in the proportion employed in 1992 and 1993, and a corresponding increase in 1994. The two groups only diverge after 1994, ostensibly because of the expansion of the EITC.

4. Difference-in-difference by hand-ish

Calculate the average proportion of employed women in the treatment and control groups before and after the EITC expansion. (Hint: group by `any_kids` and `after_1993` and find the average of `work`.)

```
eitc_diff_diff_means <- eitc %>%
  group_by(any_kids, after_1993) %>%
  summarize(avg_work = mean(work))

cell_A <- eitc_diff_diff_means %>%
  filter(any_kids == FALSE, after_1993 == FALSE) %>%
  pull(avg_work)

cell_B <- eitc_diff_diff_means %>%
  filter(any_kids == FALSE, after_1993 == TRUE) %>%
  pull(avg_work)

cell_C <- eitc_diff_diff_means %>%
  filter(any_kids == TRUE, after_1993 == FALSE) %>%
  pull(avg_work)

cell_D <- eitc_diff_diff_means %>%
  filter(any_kids == TRUE, after_1993 == TRUE) %>%
  pull(avg_work)

diff_in_diff <- (cell_D - cell_C) - (cell_B - cell_A)
```

Calculate the difference-in-difference estimate given these numbers. (Recall from class that each cell has a letter (A, B, C, and D), and that the diff-in-diff estimate represents a special combination of these cells.)

	Before 1993	After 1993	Difference
Women with no kids	0.576	0.573	-0.002
Women with kids	0.446	0.491	0.045
Difference			0.047

What is the difference-in-difference estimate? Discuss the result. (Hint, these numbers are percents, so you can multiply them by 100 to make it easier to interpret. For instance, if the diff-in-diff number is 0.15 (it's not), you could say that the EITC caused the the proportion of mothers in the workplace to increase 15 percentage points.)

The diff-in-diff estimate here shows that the EITC had a causal effect of increasing the labor force participation of mothers with children by nearly 5 percentage points.

5. Difference-in-difference with regression

Run a regression model to find the diff-in-diff estimate of the effect of the EITC on employment (`work`) (hint: remember that you'll be using an interaction term).

```
model_simple <- lm(work ~ any_kids + after_1993 + any_kids * after_1993, data = eitc)
tidy(model_simple)
```

term	estimate	std.error	statistic	p.value
(Intercept)	0.575	0.009	65.06	0.000
any_kidsTRUE	-0.129	0.012	-11.09	0.000
after_1993TRUE	-0.002	0.013	-0.16	0.873
any_kidsTRUE:after_1993TRUE	0.047	0.017	2.73	0.006

```
glance(model_simple)
```

r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC	deviance	df.residual	nobs
0.013	0.012	0.497	58.5	0	3	-9885	19780	19817	3391	13742	13746

How does this value compare with what you found in part 4 earlier? What is the advantage of doing this instead of making a 2x2 table?

The coefficient for the `any_kidsTRUE:after_1993TRUE` term should be identical to the number you got from the 2x2 table: 0.047 The advantages of calculating this value through regression are (1) it's easier and faster, (2) you can more easily get measures of uncertainty and significance, and (3) you can include other control variables.

6. Difference-in-difference with regression and controls

Run a new regression model with demographic controls. Eissa and Liebman used the following in their original study: non-labor income (family income minus personal earnings, or the `unearn` column), number of children, race, age, age squared, education, and education squared. You'll need to make new variables for age squared and education squared. (These are squared because higher values of age and education might have a greater effect: someone with 4 years of education would have 16 squared years, while someone with 8 years (twice as much) would have 64 squared years (way more than twice as much).)

```
eitc <- eitc %>%
  mutate(age_2 = age^2,
         ed_2 = ed^2)

model_complex <- lm(work ~ any_kids + after_1993 + any_kids * after_1993 +
                      unearn + children + nonwhite + age + age_2 + ed + ed_2,
                    data = eitc)
tidy(model_complex)
```

term	estimate	std.error	statistic	p.value
(Intercept)	0.061	0.060	1.019	0.308
any_kidsTRUE	-0.021	0.015	-1.433	0.152
after_1993TRUE	-0.009	0.012	-0.704	0.482
unearn	-0.018	0.001	-30.841	0.000
children	-0.052	0.005	-11.458	0.000
nonwhite	-0.063	0.008	-7.404	0.000
age	0.030	0.003	9.217	0.000
age_2	0.000	0.000	-8.392	0.000
ed	-0.004	0.006	-0.669	0.503
ed_2	0.001	0.000	3.262	0.001
any_kidsTRUE:after_1993TRUE	0.058	0.016	3.554	0.000

```
glance(model_complex)
```

r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC	deviance	df.residual	nobs
0.104	0.103	0.473	160	0	10	-9216	18456	18547	3077	13735	13746

Does the treatment effect change? Interpret these findings.

After controlling for demographics, the diff-in-diff estimate is now 0.058, which is a little larger than it was in the simple model, since the control variables helped filter out or isolate some of the variation in `work`.

7. Varying treatment effects

Make two new binary indicator variables showing if the woman has one child or not and two children or not. Name them `one_kid` and `two_plus_kids` (hint: use `mutate(BLAH = children == SOMETHING)`.

```
eitc <- eitc %>%
  mutate(one_kid = children == 1,
         two_plus_kids = children >= 2)
```

Rerun the regression model from part 6 (i.e. with all the demographic controls), but remove the `any_kids` and `any_kids * after_1993` terms and replace them with two new interaction terms: `one_kid * after_1993` and `two_plus_kids * after_1993`.

```
model_different_kids <- lm(work ~ one_kid + two_plus_kids + after_1993 +
                             one_kid * after_1993 + two_plus_kids * after_1993 +
                             unearn + children + nonwhite + age + age_2 + ed + ed_2,
                           data = eitc)
tidy(model_different_kids)
```

term	estimate	std.error	statistic	p.value
(Intercept)	0.061	0.060	1.017	0.309
one_kidTRUE	-0.014	0.016	-0.904	0.366
two_plus_kidsTRUE	-0.022	0.022	-1.003	0.316
after_1993TRUE	-0.009	0.012	-0.704	0.481
unearn	-0.018	0.001	-30.837	0.000
children	-0.053	0.006	-8.118	0.000
nonwhite	-0.063	0.008	-7.389	0.000
age	0.030	0.003	9.213	0.000
age_2	0.000	0.000	-8.386	0.000
ed	-0.004	0.006	-0.681	0.496
ed_2	0.001	0.000	3.276	0.001
one_kidTRUE:after_1993TRUE	0.044	0.021	2.071	0.038
two_plus_kidsTRUE:after_1993TRUE	0.067	0.018	3.643	0.000

```
glance(model_different_kids)
```

r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC	deviance	df.residual	nobs
0.104	0.103	0.473	133	0	12	-9216	18459	18565	3076	13733	13746

For which group of women is the EITC treatment the strongest for (i.e. which group sees the greatest change in employment)? Why do you think that is?

The EITC effect is largest for those with two kids, raising workforce participation by 6.7 percentage points. Those with only one kid saw a 4.4 percentage point increase. This is probably because the EITC is larger for families with more kids.

8. Check parallel trends with fake treatment

To make sure this effect isn't driven by any pre-treatment trends, we can pretend that the EITC was expanded in 1991 (starting in 1992) instead of 1993.

Create a new dataset that only includes data from 1991–1993 (hint: use `filter()`). Create a new binary before/after indicator named `after_1991` (hint: `year >= 1992`). Use regression to find the diff-in-diff estimate of the EITC on `work` (don't worry about adding demographic controls).

```
eitc_fake_treatment <- eitc %>%
  filter(year < 1994) %>%
  mutate(after_1991 = year >= 1992)

model_fake_treatment <- lm(work ~ any_kids + after_1991 + any_kids * after_1991,
                           data = eitc_fake_treatment)
tidy(model_fake_treatment)
```

term	estimate	std.error	statistic	p.value
(Intercept)	0.583	0.015	39.132	0.000
any_kidsTRUE	-0.123	0.020	-6.262	0.000
after_1991TRUE	-0.012	0.018	-0.631	0.528
any_kidsTRUE:after_1991TRUE	-0.010	0.024	-0.415	0.678

```
glance(model_fake_treatment)
```

r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC	deviance	df.residual	nobs
0.017	0.016	0.496	41.9	0	3	-5309	10628	10663	1819	7397	7401

Is there a significant diff-in-diff effect? What does this mean for pre-treatment trends?

If we pretend the EITC expansion happened in 1991 instead of 1993, we don't see any significant diff-in-diff effect. The coefficient for the interaction term is -0.01, which seems to *decrease* workforce participation a tiny bit, but the p-value is 0.678 (which is really high), so there's likely no significant effect. There are no preexisting trends that could be driving the results.