Registry of Efficacy and Effectiveness Studies

Study Title:
The Effects of High-Stakes Teacher Evaluation Policies on Student Disciplinary Outcomes

Registry ID: #1748.1v3

Version History

Changes were published on September 25, 2019 7:40 PM ET.
Currently viewing this version.
Description of changes published:
Updated subscript indexing notation in DD estimates and included correct interaction terms in DDD estimates.

Changes were published on April 4, 2019 1:34 PM ET.
Review this version.
Description of changes published:
Previous submission (1 day earlier) only include difference-in-differences analytic model, but it did not include difference-in-difference-in-differences model. Updated analytic model and additional materials to include triple differences.

The first version of this entry was published on April 3, 2019 4:25 PM ET.
Review this version.

Section I: General Study Information

PI name: David Liebowitz
PI affiliation: University of Oregon

Co-PI name: Lorna Porter
Co-PI affiliation: University of Oregon

Primary funding source:
No Funder

Award number(s):

IRB name: Committee for the Protection of Human Subjects, the University of Oregon Institutional Review Board
IRB approval date: 2018-11-02
IRB approval number: 10192018.029

Other registration name:
Other registration date:
Other registration number:

Study start date: 2019-04-03
Study end date:

Intervention start date:

Timing of entry: Prior to analysis of outcome data

Brief abstract:
In a span of seven years between 2009 and 2016, nearly all U.S. states adopted high-stakes teacher evaluation policies. These policies have as a goal increasing the quality of teacher performance through frequent observation and feedback. We investigate the potential unintended consequences of the introduction of these policies on teachers’ response to students’ class behavior.

We estimate the causal impact of the implementation of high-stakes teacher evaluation policies on the frequency with which students are the subject of an Office Disciplinary Referral (ODR) from their classroom teacher. We hypothesize that, in the 44 states and the District of Columbia that had implemented major reforms to their evaluation systems between 2009 and 2016, the increased scrutiny experienced by teachers may have led some to be more likely to remove students from their classrooms as a result of perceived misbehavior. Our primary data source is the School-Wide Information System (SWIS) data. These data include records of each educator-recorded behavioral infraction approximately 6,000 schools from 2006/07 to 2017/18. We leverage Steinberg and Donaldson’s (2016) tally of evaluation reforms, extended by Kraft et al. (2019), to fit a two-way fixed effect difference-in-difference model that estimates the impact of high-stakes evaluation policy reform on ODRs. Our first difference is the change in the rate of classroom-based subjective ODRs in locales that experienced the teacher evaluation policy reform. Our second difference is the change in the rate of ODRs in locales that did not (or had not yet) experienced the change. As a critical improvement over standard state policy variation difference-in-difference estimates that struggle to capture endogenous differences across states, we employ triple-difference estimates in which our third difference is the change in the rate of objective and/or non-classroom based ODRs. Since these types of infractions occur within the same schools and presumably are not influenced by changes in teacher evaluation policy (i.e., students are no more/less likely to bring a knife to school under pre- or post-treatment conditions), we argue that our triple-difference estimates are unbiased by state- or district- policy differences. To further test this approach, we estimate models in which our third difference comes from restricted-use Civil Rights Data Collection measures of suspension rates, which we similarly argue should not be influenced by changes in teacher evaluation policies. We conduct robustness checks for differences in disciplinary referral trends pre-policy implementation and for Goodman-Bacon (2018) early- and late-timing variation weights.

We conclude with exploratory analysis of the potential for school leadership actions to moderate the effect of high-stakes evaluation on discipline outcomes. SWIS schools receive externally-validated ratings on the quality of implementation of their Positive Behavioral Intervention and Supports (PBIS) systems. We model the extent to which teachers in schools with effective systems of behavior support, as captured by these ratings, are or are not impacted by the introduction of high-stakes teacher evaluation policies.

Keywords: teacher evaluation, student behavior, discipline, school leadership, PBIS, policy analysis

Comments:

Section II starts on the next page.
Section II: Description of Study

Type of intervention:
Policy, Practice

Topic area of intervention:
Social and Behavioral Context for Academic Learning, Educational Leadership, Teacher Evaluation

Number of intervention arms:
1

Target school level of intervention:
Kindergarten, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12

Target school type:
Rural, Urban, Suburban

Location of implementation:
*United States*: West, Midwest, Northeast, South

Further description of location:
Teacher evaluation reforms passed in 44 states + DC between 2009 and 2016:
AK
AL
AR
AZ
CO
CT
DC
DE
FL
GA
HI
ID
IL
IN
KS
KY
LA
MA
MD
ME
MI
MN
MO
MS
NC
ND
NH
NJ
NM
NV
NY
OH
OK
OR
Remaining states used as part of comparison condition.

**Brief description of intervention arm:**
In response to strong incentives from the federal government’s Race to the Top (RTT) program, state legislatures across the country enacted laws aimed at increasing accountability for public school teachers. These efforts have primarily focused on implementing high-stakes teacher evaluation systems intended to increase the quality of teacher performance through frequent observation and feedback. Between 2009 and 2016, 44 states and the District of Columbia had implemented major reforms to their teacher evaluation systems (Steinberg & Donaldson, 2016; Kraft et al. 2019).

**Brief description of comparison condition:**
Comparison condition is states that had not (or had not yet) implemented high-stakes teacher evaluation policies.

**Comparison condition:** Business-as-usual

**Comments:** In some models in Analytic Model II (see below), we intend to estimate results only for grades 3-11. We do this on the assumption that evaluators may have been more responsive to high-stakes teacher evaluation policy changes in grades in which high-stakes tests occur. During the period of our analysis, all states mandated yearly tests in grades 3-8. High school tests occur in either grades 9, 10, 11 or some combination of these years. While in some states the years of high-stakes tests are clearly defined, in others some students take them in different years. We model results in grades 3-11, therefore, and assume these are downward-biased estimates of the true effects.

**Section III: Research Questions**

**Confirmatory research question:**

*Question 1*
Did the introduction of high-stakes teacher evaluation policies that rely on regular observations increase the rate of office disciplinary referrals (ODRs)?

**Exploratory research question:**

*Question 1*
Do changes in rates of office disciplinary referrals (ODRs) stemming from the introduction of high-stakes teacher evaluation policies vary by the measured quality of behavioral supports in schools?

**Comments:**
Section IV-A: Study Design (Selection)

**Study Design:**
Quasi-experimental Design with comparison group (QED)

**Comments:**

Section IV-B: Study Design (Input)

**Design**

**Unit of intervention implementation:**
State

**Assignment within blocks or selected strata:**
Yes

**Please define the natural blocks or purposefully selected strata.**
Time period

**Unit outcome data measured:**
School

**Intermediate clusters between unit of implementation and unit of measurement:**
Yes

**Description:**
District

**Matching Procedures**

**Comparison units will be selected at:**

**Design Classification**

Based on the responses above, this study has been classified as:
QED: Multisite (Blocked) Nested

**Comments**

We will fit a two-way fixed effect (school and year fixed effects) difference-in-difference model. Our first difference is the change in the per-student rate of classroom-based subjective ODRs in schools that experienced the teacher evaluation policy reform. Our second difference is the change in the per-student rate of ODRs in locales that did not (or had not yet) experienced the change. We employ triple-difference estimates in which our third difference is the change in the per-student rate of objective or non-classroom based ODRs. To further test this approach, we estimate models in which our third difference comes from restricted-use Civil Rights Data Collection measures of district, out-of-school suspension rates, which we similarly argue should not be influenced by changes in teacher evaluation policies.

Thus, note that the radio buttons do not exactly capture the unit of implementation. This occurs at both the
Section V: Sample Characteristics

Approximate number of schools per intermediate cluster: 13

Approximate number of intermediate clusters per assignment unit: 10

Approximate number of assignment units in the comparison condition within each time period: 51

Approximate number of assignment units in the intervention condition within each time period: 51

Number of time periods: 12

Certain schools that were targeted for the study:
Yes - Schools that use the SchoolWide Information System (SWIS) database to record behavioral data

Certain schools that were excluded from the study:
Yes - Juvenile justice schools excluded due to alternate disciplinary systems and potential non-participation in statewide evaluation policies Private and charter schools because not subject to statewide evaluation policies

Certain intermediate clusters that were targeted for the study:
Yes - Districts that contained schools that used the SWIS system.

Certain intermediate clusters that were excluded from the study:
Yes - Districts located in non-US states and BIA districts excluded due to not being subject to teacher evaluation policy reform

Certain assignment units that were targeted for the study:
No

Certain assignment units that were excluded from the study:
No

Certain time periods that were targeted for the study:
Yes - 2006-07 to 2017-18 to have three years prior to first change in teacher evaluation, through most recent year of available data. However, demographic data is only available through 2015-16. We impute demographic data for these two years, and we also test for robustness to excluding these years. Additionally, panels will be balanced in certain analyses, resulting in some schools in states with censored data being dropped.

Certain time periods that were excluded from the study:
No

Comments

Section VI-A starts on the next page.
Section VI-A: Outcomes (Selection)

Confirmatory question 1 - number of outcome measures: 2

Comments:

Confirmatory Question 1, Outcome Measure 1

Outcome domain: Student Social, Emotional, & Behavior - Behavior

Minimum detectable effect size: 0.075 SD

Outcome measure: Per-student classroom-based ODR rate

Scale of outcome measure: Continuous

Normed or state test: No

Test-retest reliability: N/A

Internal consistency: N/A

Inter-rater reliability: N/A

Same outcome measure in treatment and comparison groups: Yes

Confirmatory Question 1, Outcome Measure 2

Outcome domain: Student Social, Emotional, & Behavior - Behavior

Minimum detectable effect size: 0.075 SD

Outcome measure: Per-student classroom-based subjective ODR

Scale of outcome measure: Continuous

Normed or state test: No

Test-retest reliability: N/A

Internal consistency: N/A

Inter-rater reliability: N/A

Same outcome measure in treatment and comparison groups: Yes

Comments:
Section VII: Analysis Plan

Baseline data collected prior to start of intervention: No

Covariates to include at the school level in the model:
Gender, Free/reduced lunch status, Race, Grade

Covariates to include at the intermediate cluster level in the model:
Aggregate of Individual Characteristics

Covariates to include at the assignment unit level in the model:

Analytic model:

$$\text{ODR}_{jst} = \Gamma_j + \Pi_t + \beta_1 EV A L_{st} + \Theta(X_{jt}) + \varepsilon_{st}$$

In simplified form, this represents the per-student rate of Office Disciplinary Referrals (ODRijst) in school i, in district j, state s and time t, regressed on the indicator EVAL if the school is in a state that is in a year with a high-stakes evaluation system. Beta_1 is the causal parameter of interest in the Diff-in-Diffs regression framework. The two-way fixed effect model includes school- (gamma) and year- (pi) fixed effects and vectors of school-level (j) background characteristics. Time-varying characteristics (e.g., % free/reduced lunch) are included as main effects. Standard errors are clustered at the state level.

NOTE UPDATE on APRIL 4: the above model only includes the analytic strategy for our difference-in-differences estimates. We present here our simplified triple difference estimates and have added an updated memo that spells them out fully in the additional materials section.

NOTE UPDATE on SEPTEMBER 25: we have made some slight corrections to the indexing of our DD models and the correct interactions in our DDD models.

$$\text{ODR}_{jst} = \Delta_y + \Gamma_j + \Pi_t + \beta_1 EV A L_{x A F F E C T_{jst}} + \beta_2 EV A L_{s t} + \beta_3 A F F E C T + (A F F E C T_{st} \cdot \Gamma_j)\phi + (A F F E C T_{st} \cdot \Pi_t)\delta + (X_{jt})\theta + \varepsilon_{st}$$

For the triple difference estimates, we define the same terms as above and add an indicator variable (AFFECT) which takes the value of 1 if the observed school-level ODR outcome is for a “subjective” disciplinary category and 0 for a “less subjective” disciplinary category. These categories are defined in Greflund, McIntosh, Mercer and May (2014) from 21 problem behavior types in the SWIS management system. The main effect of AFFECT captures differences in the within-school rates of office referrals for subjective and “less subjective” behaviors. The causal parameter of interest is Beta_1 in the triple difference regression framework: it is the interaction between the timing of the introduction of the high-stakes evaluation policy and the subjective reason for the ODR. We estimate analogous triple differences with the indicator variable CLASS which takes the value of 1 when the office behavioral referral occurs from a classroom location and 0 when the ODR originates from any other physical location (e.g., cafeteria, stairwell, playground, etc.).

Plan to handle cases with missing outcome data:
Delete cases with missing data for the outcome being analyzed

Planned multiple comparisons adjustment, confirmatory question 1 (Student Social, Emotional, & Behavior - Behavior):
No

Comments:
**Section VIII: Additional Materials**
Right click to open files in a new window.

**Links**

No links have been added yet.

**Files**

File Name: REES_analytic_method_1748.pdf
Description:

- File Name: REES_analytic_method_1748_DDDupdate.pdf
Description: Updated 1 day after original submission to include triple difference estimates

- File Name: REES_analytic_method_1748_DDDupdate_0925.pdf
Description: Updated to correct indexing subscript mistakes in DD notation and correct interactions in DDD estimates.

**Comments**