Teachers Helping Teachers? Peer Effects of Elementary School Faculty on Coworker Performance

Christopher Monjaras

Abstract

Teachers influence student achievement, not only through direct instruction but also via between-teacher peer effects which are understudied. Using data from Indiana public elementary schools, I estimate between-teacher spillovers through standardized test score value-added measures by leveraging idiosyncratic employment changes of high- and low-performing teachers. High-quality teachers generate positive spillovers of 0.05–0.15 SD, while low-quality teachers create negative spillovers of 0.05–0.1 SD. These effects are comparable to major education policies like incentive pay programs, highlighting the importance of teacher spillovers in student learning. Understanding these indirect effects is crucial for optimizing teacher workforce policies and improving educational outcomes.

Measuring the contributions of educators to student achievement is critical to understanding how students accumulate skills and, in turn, a significant fraction of their labor market outcomes and economic mobility. The typical models used to calculate teacher quality are based on a function of measurable short-term student outcomes like test scores, graduation rates, and disciplinary records as primary inputs (Koedel, Mihaly, & Rockoff, 2015; Jackson, 2018; Bacher-Hicks, Chin, Kane, & Staiger, 2019). Such measures of teacher quality, referred to in the economic literature collectively as value-added estimation, are a convenient tool to calculate teacher quality based on direct teacher-to-student interactions but often fail to fully capture the totality of ways teachers can influence student outcomes. Indeed, it is an implicit assumption of these models that, conditional on random or quasi-random assignment of students to instructors, teachers act as fully independent spheres of influence over their students such that the entirety of student achievement observed in the data is presumed to be attributable to classroom interaction. Perhaps the sole exception to this is the recent developments in the value-added literature studying the spillover effects between students (Koedel et al., 2015). This simplifying assumption is useful for model tractability, especially in settings with small sample sizes, but is unlikely to yield an accurate measure of teacher quality simply because there are so many sources of potential bias unaccounted for.

The goal of the researcher then is to determine, given their particular setting, which forms of potential bias must be adjusted for and which can safely be ignored. With this paper I seek to add to the existing literature on bias in value-added modeling by presenting a measure and method of general estimation for an understudied form of bias, namely between teacher spillover effects which I define below. I then make the case that between teacher spillover effects ought not be ignored in an unbiased value-added model specification.

A teacher's professional responsibilities go well beyond the instruction of their students. While this is certainly core to their job, teachers also are tasked to work with each other to complete peer professional development programs, joint curriculum planning, and unified management of disciplinary issues to name a few (Darling-Hammond, 2015). These kinds of interactions are what I refer to as between teacher interactions or spillover effects and it is ex ante unclear how these actions impact student achievement. If there is any non-zero effect however, it is certainly being attributed to the wrong teacher in a traditional value added model given that between teacher effects are not accounted for in any model typically used in the literature. In this paper, I attempt to separate out these spillover effects by building on the quasi-experimental measure of bias developed by Chetty et al. (2014a) to estimate of the impact of between-teacher interactions on teacher quality using a novel estimation approach in the context of Indiana Public 4th and 5th grade elementary school teachers. In Chetty et al. (2014a) the authors use teacher mobility to observe changes in effect of teachers of differing quality on similar student cohorts to test for general estimation bias. I modify this estimation strategy to exploit teacher turnover from schools in the form of moves from one school to another and leaves of absence instead as shocks to the network of teachers who are left behind. I then estimate a difference-in-difference framework to track the quality of the teachers left behind in the subsequent years. Importantly, I only consider a school as treated in this setting if the teacher who generates an exit event is among the pool of highest quality individuals in the sample, or is among the worst preforming teachers in the sample. Focusing on the extreme tails of the distribution affords the best chance of detecting potential spillover effects and naturally sets upper and lower bounds on the magnitude of effect.

My results show that after controlling for teacher, student, and school characteristics, when top performing teachers exit a school the teachers they leave behind are 0.05-0.15 of a standard deviation worse in quality then they were before. Additionally, when poor performing teachers exit a school the teachers that remain become 0.05-0.1 of a standard deviation better in quality. Relative to the magnitudes of effect in the value-added literature in other contexts, these estimates are large in comparison to other policy interventions intended to

influence teacher quality but not unreasonably so. Taken together, these results suggest that teachers do produce spillover effects on their peers and that the direction of those spillover effects is dependent on the quality of teachers participating in a network. I go on to show that these results are robust to reduced form specification, but there is some evidence to believe that subject area specialization may impact the magnitude of the results.

The interpretation of these results as causal is dependent on two critical assumptions. The first is that teacher movements between schools and exits are plausibly exogenous. This is equivalent to Assumption 3 from Chetty et al. (2014a) which holds that "changes in teacher [value-added] across cohorts within a school grade are orthogonal to changes in other determinants of student scores." I work in a similar context as Chetty et al. and likewise find that changes in teacher staffing decisions are sufficiently idiosyncratic to suggest that this assumption is reasonable, even in lieu of non-random student sorting. The second is that the time trends for teacher quality among the teachers at untreated schools are an accurate counterfactual for time trends for teachers in schools that experience an exit event from a superstar or dud individual. While I cannot rule out all possible endogeniety stories, nor can I directly test for parallel trends, I provide evidence from event studies, placebo tests, alternative specifications, and other robustness checks to help lend some evidence to their reasonability. That notwithstanding, every effort is made to control for heterogeneous trends among subgroups of teachers and students, and results are tested against alternative treatment intensities and durations to lend support to their believability.

Few scholars to date have dealt with the direct measurement of teacher spillover effects. The paper by Jackson and Bruegmann (2009) is, to my knowledge, the first to consider between-teacher spillovers. The authors use a fairly standard value added model to directly estimate the fraction of student achievement attributable to between-teacher spillovers. In this paper I improve on this method by incorporating the advances in the value added literature spurred by Kane and Staiger (2008) and Rothstein (2009, 2010). Furthermore, I use a method of estimation that blends traditional value-added modeling with difference-in-difference techniques which has a stronger claim to causality. Other related works include Koedel et al. (2009) which focused on the question of joint production of educational outcomes between academic departments at the secondary level. This answers a fundamentally different question than my work in that Koedel is concerned with productivity gains related to overlapping interests between academic specialties. Furthermore, the author makes no claims to causality and admits that the structure of the data used may produced a bias result. Kho et. al. (2022) touched on the issue tangentially when they studied teacher exits related to an incentive pay program to encourage good teachers to work in poor preforming schools

in the state of Tennessee. Though primarily focused on the evaluation of this incentive pay program, their method is similar in structure to my approach and their results are in line with mine. I believe that my work improves upon theirs both in terms of generality and robustness. Not only can I make claims about the spillover effects of good teachers, but I can also consider the impact of poor quality teachers. Furthermore, my method can be used in the context of other data settings allowing future researchers to more accurately calibrate value-added models using the idiosyncrasies of their needs and data concerns directly.

My paper also makes contributions to the literature on eliminating bias and improving stability in the measurement of value-added estimates (Rothstein, 2010; Chetty et al., 2014a; Chetty, Friedman, & Rockoff, 2014b; Guarino, Reckase, & Wooldridge, 2015; Guarino, Reckase, Stacy, & Wooldridge, 2015; Backes et al., 2018; Stacy, Guarino, & Wooldridge, 2018), the broader literature of the role of spillovers in education (Papay, 2011; Opper, 2019; Gershenson, Lindsay, Papageorge, Campbell, & Rendon, 2023; Gilraine & McCarthy, 2024), teacher labor markets (Steele, Pepper, Springer, & Lockwood, 2015; Adnot, Dee, Katz, & Wyckoff, 2016; Bruno & Strunk, 2019; Henry & Redding, 2020; Cullen, Koedel, & Parsons, 2021), and the role of teacher quality in student outcomes (Aaronson, Barrow, & Sander, 2007; Boyd, Grossman, Lankford, Loeb, & Wyckoff, 2009; Hanushek, 2011; Darling-Hammond, 2015; Ronfeldt, Farmer, McQueen, & Grissom, 2015; Adnot et al., 2016; Henry & Redding, 2020) among others.

1 Background

In this section I provide a brief primer to value-added modeling in education. Much of what is presented here is a summary of the excellent literature review by Koedel et al. (2015). Value-added models in education are tools used to separate out the individual contributions of teachers towards the achievement of their students. In most cases, these models are linear, estimated by OLS, and use easily measurable student metrics like test scores, graduation rates, discipline reports and the like as the outcome of interest. There is much disagreement within the value-added literature over what a proper specification for a value-added model looks like and there are very important policy implication for getting it right. Today, almost all states in the US use some form of a value-added model to evaluate teacher performance, to make hiring and firing decisions, and in allocating funding to schools (Darling-Hammond, 2015). Despite this, the most frequently used forms of the value-added model have common roots in the early days of economic theory on education, and in particular, the cumulative student achievement function.

$$A_{it} = A_t[X_i(t), F_i(t), S_i(t), \alpha_{i0}, \epsilon_{it}] \tag{1}$$

Equation 1 represents a generalized form of the students achievement function first proposed by (Ben-Porath, 1967; Hanushek, 1979) and using notation from (Todd & Wolpin, 2003). Here A_{it} describes the achievement level in the chosen metric (test scores, graduation rates etc.) for student i in time t. Here we see that current achievement is a function A_t of cumulative inputs where $X_i(t)$, $F_i(t)$, and $S_i(t)$ are histories of individual, family and school inputs respectively. The α_{i0} term represents a student's initial ability, and an idiosyncratic error ϵ_{it} is also included. Value-added models attempt to estimate the impact of school inputs and, in particular, the role of teachers on student achievement by making the assumption that prior achievement can stand in as a rough approximation for the histories of prior inputs.

As such, most linear value-added models rely on a panel of linked student-teacher data. Typical models control for school and student characteristics and include a fixed effects term for each teacher that represents the value-added contributions by teachers to whatever metric is chosen as the outcome of interest. Practically speaking, the actual values that are returned as value-added measures are often not easily interpreted. Rather, they are mostly used for direct comparison or for ranking between teachers much like how standardized test score results have no inherent meaning except within the context of their larger distribution.

In recent years, much of the new research in this area has focused on the proper specification of the value-added model. Chief among these concerns are how to go about addressing issues of bias, stability in measurement of teacher quality, and what estimating procedure to use. A complete discussion of these topics is well beyond the scope of this paper, so for the purposes of this project I want to focus in on the issue of bias as it is where my work makes its most substantial contribution.

It is well established in the value-added literature that the simple approach of generating OLS estimates on aggregate data with basic student and school controls will produce a biased measure of teacher value-added. There are two key criticisms. First, Rothstein (2009, 2010) points to how issues of non-random selection between teacher-student links can lead to under- or overestimating teacher impact. Second, Goldhaber & Chaplin (2015) and Guarino et al. (2015) show that when student-to-student and teacher-to-teacher interactions outside of the classroom are ignored, teacher value-added is often underestimated. Thus far, no tests exist to determine the scope, direction, or existence of bias related to these issues. Because of that fact, I believe that this paper could represent an important contribution to understanding

bias of the second form, namely between teacher spillover effects.

There is however, an inherent internal inconsistency in my approach as I am using value-added estimation in a paper that challenges biased value-added models. This is a valid criticism, however other recent developments using large administrative data sets seem to suggest a way forward. In their two related papers, Chetty et al. (2014a, 2014b) show that using data over a long period of observation will minimize the possibility of a biased value-added estimation provided that in addition to the traditional student and school controls, a lagged measure of achievement is also used. This is further established by Koedel et. al (2011) and Kinsler (2012) who both challenge the validity of the Rothstein critique and argue a large sample size and sufficiently complex model can overcome bias of this sort. Taken together, these optimizations allow for value-added estimates that are accurate at a 95% confidence level and alleviate concerns that sorting on unobservables are biasing results. In the next section, I will show that the data gathered from the Indiana Department of Education and used for this paper meets this standard. And, in the methods section, I adapt my value-added specification to the Chetty et al. critique. In this way, I hope to alleviate concerns that issues of bias undermine my results.

2 Data and Descriptive Statistics

Data for this paper are student-teacher-year observations from two primary sources. The bulk of the data used for this project comes from the Indiana Department of Education (IDOE) with a small amount of supplemental data collected from the National Center for Educational Statistics (NCES). Unless otherwise stated, all data is collected across the years 2011-2018. Throughout the remainder of this paper, I will use a single year to denote an entire school year using the convention of referring to a school year by its spring semester year (i.e. 2012-2013 becomes 2013).

IDOE

Data from the IDOE was collected as part of a data sharing agreement between the University of Notre Dame's Center for Educational Research (CREO) and the state government of Indiana. Twice a year (once each semester), the IDOE requests each public school and private school that accepts state school choice vouchers to provide administrative information on their student body, faculty, and staff. Schools must report counts of enrollment and employment along with detailed demographic information for all members of their communities. In addition to this, the IDOE also collects information on key educational metrics

including special education status, disciplinary records, free or reduced price lunch (FRPL) eligibility, attendance, mobility, and test scores at the end of each academic year for all students they oversee. Although this is de-identified for use in research, sensitive individual level data is reported and use of this data is restricted.

The CREO database has access to data from as early as 2006 to 2022 however, the limiting factor in determining an appropriate time frame and student age range for the data used in this paper was student test scores. As mentioned in the previous section, a value-added model requires an educational outcome like test scores to estimate teacher quality on. Prior to the 2018-2019 school year, the state of Indiana used the Indiana Statewide Testing for Educational Progress-Plus (ISTEP+) exam to evaluate student mastery of core concepts in both English Language Arts (ELA) and math. This exam was administered to all public school students and some private school students in grades 3-8. This was an adaptive standardized exam meaning all students were shown questions from the same pool and the better they preformed the harder and more advanced the questions became. As such, all students in grades 3-8 were graded on the same scale making comparisons of students both within and across grade levels possible. Starting in 2019, Indiana switched to a new form of standardized testing that uses a different grading scheme and scale. In order to ensure consistency I focus on test scores gathered prior to the change in exam type and include data only from public schools and public charter schools due to the incomplete nature of private school data.

I further restrict my sample of interest to students and teachers between 2011-2018 because the rich demographic data necessary for proper specification of a value-added model was first collected in 2011. The only exception to this is that testing data from 2010 is pulled to serve as a lag to students observed in 2011 in order to account for the Chetty et al. critique. Note also that I focus my analysis on 4th and 5th grade teachers and students only. The reason for this is twofold. First, I want the best chance to capture spillovers and other economic contexts suggest I should look for cases where individuals are working closely together and on similar problems. Logically, this would imply that two elementary school teachers are more likely to generate spillover effects between themselves than an elementary school teacher and a high school teacher for example. I leave the exploration of cases like that for possible future research. Secondly, teacher and student tracking is greatly simplified if students do not move between classes as is common in US middle and high schools and, instead, students are linked to a single teacher in a given year. Not only does this prevent errors in the linking process, but it also simplifies the identification strategy by clearly defining which teachers are responsible for which sets of test scores. As testing only

Table 1: IDOE Descriptive Statistics for 4th and 5th grade Elementary School Students

Variable	Value
Total Number of Unique Student-Year Observations	10,070,247
Total Number of Unique Students	701,947
Total Number of Unique 4th Graders	594,230
Total number of Unique 5th Graders	594,219
% of All Students Who are Female	48.96
% of All Students Who are White	69.35
% of All Students Who are Black	12.10
% of All Students Who are Hispanic	11.07
% of All Students Who are Asian	2.15
% of All Students With Free or Reduced Price Lunch Status	51.22
% of All Students Who Claim a Physical or Mental Disability	14.69
% of All Students Who Have Special Education Status	14.68
% of All Students With Individualized Education Plan for Math	16.71
% of All Students With Individualized Education Plan for ELA	16.59

Note: These descriptive statistics are generated from the 4th and 5th grade public elementary school students included in my analysis from the IDOE data for the 2011-2018 school years. All %s are calculated on the base of unique students, both 4th and 5th graders, rather than student-year observations.

begins in 3rd grade and at least one year of lagged test scores is required, only 4th and 5th grade students satisfy all of these restrictions.

Table 1 presents summary statistics for the students included in the sample. Once pooled, my sample contains 10,070,247 student-year observations for 701,947 unique 4th and 5th grade students. Table 2 contains a similar set of descriptive statistics for the 24,983 unique teachers included in my sample. Compared to a national sample, these tables suggest that Indiana students are more white, less likely to be in special education programs, and are about average in terms of access to the FRPL program. Likewise, comparing the teacher data to a national sample suggests that Indiana teachers are more white, less well educated, but are nearly identical to their national colleagues in terms of gender spread and years of experience.

NCES

All data on school level characteristics was collected from the annual NCES Common Core of Data (CCD). This is publicly available data on students and faculty aggregated, in my case, to the school level. The data are easily linked to the IDOE panels through a

Table 2: IDOE Descriptive Statistics for 4th and 5th grade Elementary School Teachers

Variable	Value	Standard Deviation
Total Number of Unique Teachers	24,983	
Mean Age	44.42	(13.13)
Mean Years of Experience	12.48	(11.06)
% of All Teachers Who are Female	81.77	
% of All Teachers Who are White	92.99	
% of All Teachers Who are Black	4.79	
% of All Teachers Who are Hispanic	0.96	
% of All Teachers Who are Asian	0.06	
% of All Teachers With an Advanced Degree	44.43	
% of All Teachers Classified as Highly Qualified	69.72	

Note: These descriptive statistics are generated from the 4th and 5th grade public elementary school teachers included in my analysis from the IDOE data for the 2011-2018 school years. All %s are calculated on the base of total unique teachers at both 4th and 5th grade levels who appeared in the data at any point in time. The term "Advanced Degrees" refers to teachers with a masters degree, advanced professional degree, or PhD. Since a bachelors degree is require to teach in the state of Indiana, this value is not reported. The term "Highly Qualified" refers to a designation that the IDOE awards to teachers if they meet any of the following criteria: (1) Has High Objective Uniform State Standard of Evaluation (HOUSSE) certification, (2) Has National Board Certification (NBCT) (3) Has passed the PRAXIS II Test demonstrating core subject area mastery (4) Has more than 24 college credit hours in core subject area (5) Has a PhD or equivalent degree in core subject area or general education.

common ID. The NCES CCD data was also collected for school years 2011-2018 and contains information on all 1,364 public elementary schools in Indiana of which 1,115 are used in my analysis. The discrepancy between these values is accounted for after eliminating the 249 elementary schools that do not have 3rd 4th and/or 5th grade students or have only one teacher at either the 4th or 5th grade level. Relevant for my analysis are CCD variables for total school population, percents of the student body by race, gender, ethnicity, and FRPL status, as well as student faculty ratio.

3 Methodology

The empirical specification I use to estimate teacher spillover effects was inspired by the work of Azoula et al. (2010) who use the deaths of top researchers as shocks to co-authorship networks to measure spillover effects in the academic research space. The method I present here to measure the impact of between-teacher spillovers proceeds in three steps.

First, I identify top preforming teachers and poor preforming teachers in my sample by estimating a literature standard value-added model across a teacher's entire observable career data. Second, I generate a measure of time variant value-added by adapting the value-added

model used previously to allow teacher quality to change from year-to-year. This constructed measure will become the the left-hand side dependent variable in the last stage of estimation where I use a simple two-way fixed effects difference-in-difference framework to estimate the impact of an exit event. What follows is a detailed description of each step. The results of each step of this procedure in the context of the IDOE data are presented the results section.

Teacher Value-Added Estimation

Using high and low quality teachers as the source of treatment necessarily implies the need for a measure of teacher quality. In this paper I follow the common approach of using student test scores as the outcome by which I measure teacher quality. However, care must be taken to ensure that the specification avoids the issues of bias and selection laid out in the background section. For this reason, I turn to a general model structure proposed by Kodel et al. (2015) that has become standard within the literature since 2014. The exact specification I estimate is given in the equation below.

$$Y_{isjt} = \beta_0 + Y_{isjt-1}\beta_1 + X_{ist}\beta_2 + S_{ijst}\beta_3 + F_j\theta + \epsilon_{isjt} \tag{2}$$

In this model, observations are organized at the individual student i school s teacher j and year t level. On the left-hand side, Y_{isjt} represents a student's standardized score on the ISTEP+ exam taken in year t. A one year lagged test score term Y_{isjt-1} is included on the right-hand side to account for a student's prior ability and in deference to the Chetty et al. (2014a, 2014b) critique. X_{ist} is a vector of student characteristics. Included as controls here are dummy variables for a student's gender, free or reduced-price lunch status, special education status, whether or not they receive an individualized education plan (IEP), an indicator identifying if a student switched schools mid-year, and a categorical variable for student race. The S_{ijst} term represents a vector of school characteristics. Here I include indicators for school title I status, magnet school status, and charter school status. I also add continuous measures of student population, student-teacher ratio, fraction of the student population that is non-white, and fraction of the student body that is female. F_j is a vector of indicator variables for individual teachers such that θ is the fixed effects contribution of each teacher to individual student achievement and represents their value-added measure. Lastly, the ϵ_{isjt} is the error term.

Two refinements are made to the sample of teachers for which value-added is estimated. First, as mentioned earlier, the IDOE data set tracks ISTEP+ test score data from both

math and ELA subject areas. In every value-added specification I present, I run separate models for math or ELA test scores. When possible, I use the state classifications for subject area accountability to refine the sample of teachers by only including those responsible for the math or ELA test scores under consideration. Because state data is missing on accountability status for a minority of teachers, I make the assumption that these teachers are responsible for both their student's math and ELA test scores. This turns out to not be that restrictive of an assumption because the vast majority of 4th and 5th grade teachers in Indiana are generalists who teach all subjects and would be accountable for both sets of scores by the state's own evaluation rules.

Second, I restrict the pool of teachers in all specifications to only those who work at schools with at least one other 4th or 5th grade teacher. This restriction is only intended to focus the sample on teachers who are part of a larger network of teachers at their school and could feasibly produce spillover effects. In practice, the teachers eliminated from the sample by this restriction would always appear in the never treated control group and, I would argue, would not be a fair comparison to treated teachers. In any case this restriction also does not impact the sample size or demographics substantially.

For the second stage of my estimation procedure, I use an identical specification and covariate set, save for the substitution of F_j for F_{jt}. This transforms the value-added measures into their time variant forms instead of the time invariant forms derived in stage I. In both the first and second stage value-added estimations, I follow the convention in the literature to standardize teacher fixed effects to have mean zero and standard deviation one. This ensures both that the choice of the teacher left out of estimation to prevent colinearity does not impact teacher comparisons, and it allows for easier interpretation of the difference-in-difference model in terms of standard deviations from the mean.

Defining Treatment and Control Groups

From the time invariant teacher value-added measures estimated in the first step of the procedure, I next generate the sets of high and low performing teachers. To do this, I use value-added cutoffs in terms of a number of standard deviations above and below the mean where teachers in the left tail of the distribution with value-added more extreme then the cutoff are designated low performers and high performers are identified in a similar manner from the right tail. I make no claim as to the "correct" definition of a high or low performing teacher and instead opt here to test a wide variety of cutoffs between 0.5-2.5 times the standard deviation above or below the mean in steps of 0.5.

With high and low type teachers defined, I now turn to the topic of distinguishing between treatment and control groups for the difference-in-difference structure. I consider a teacher treated by a high performing teacher in year t if in year $t-1$ at least one high performing teacher exited a school they both worked at by moving to a new school or by leaving the public school education system for any length of time. An identical treatment structure is used for low performing teachers as well and, for simplicity, I impose all treatments as binary. From the IDOE data, I observe the school at which any teacher in my sample works at in a given school year. In applying this treatment structure, I define a mover in my sample as a teacher that is observed in two different schools in two consecutive years while a leaver is a teacher who is observed at a school in year $t-1$ but is not present in the sample in year t. I take both movers and leavers together as the treatment events that shock the intra-school "networks" of teachers left behind in order to measure spillover effects. Note here that it is the network of teachers, defined as the teachers who worked together at a school in a given school year, that is being treated and not the school itself. Thus, a teacher need not remain at a school in year t or beyond to be considered treated. I am interested in capturing the effect of exposure to good or bad teachers and this does not depend on current employment location.

It is not ex ante clear how many years after a treatment generating event a teacher from a treated network ought to be considered treated. Studies of spillovers in other contexts suggest that over time spillovers effects may be non-linear in their decay or growth and the choice of observation length may impact the magnitude of the measured results (Azoulay et al., 2010). In order to account for this here, I test and present results using all possible lengths of treatment from only one year of treatment after a network shock before they return to an untreated state up to the maximum observable length in my data of seven years. The last of these cases is equivalent to a difference-in-difference structure with multiple treatment periods but with the typical absorbing treatment state requirement common in applied work. In any case using less than the maximum treatment length, the specification becomes a difference-in-difference still with multiple treatment periods but allowing a previously treated individual to become untreated again. This specification is also flexible enough to allow for the fact that a high or low performing teacher may generate more than one treatment event if they happen to move or leave from multiple schools. In any case, using this treatment structure means that I consider teachers who are not employed at a school that experiences a move or leave from a high performing teacher when using high performers as the source of treatment, for example, as my control group.

The Difference-in-Difference Estimating Equation

The final phase of estimation is to apply the treatment and control groups in a two-way fixed effects difference-in-difference structure to estimate the between-teacher fixed effects. To do this, I use the stage two time-variant value-added measures for teachers as the outcome variable on the left-hand side. Observations are reorganized at the teacher-year level and fixed effects for school year and teacher ID are the two-way components. Letting $\hat{\theta_2}$ be the time variant fixed effects measured in stage two, I use the following estimating equation to measure spillover effects.

$$\hat{\theta_{2j}} = \gamma_0 + T_t\gamma_1 + F_j\gamma_2 + D_{tj}\gamma_3 + \delta_{tj} \tag{3}$$

Here, T is a time fixed effect, F is a teacher fixed effect, δ is the error term, and D is a treatment indicator set to 1 when a teacher is at a school in the period after a high or low performer exits as described above. When specified in this manner, the γ_3 coefficient reflects the degree to which spillovers exist between-teachers measured in terms of a factor of the standard deviation of the second stage distribution of value-added estimates. In estimating this model, I calculate clustered robust standard errors at the school level. This not only helps to ensure that standard errors are accurate given the usage of an estimated left hand side variable, but it also accounts for correlation between treated and untreated teachers at the same school. Also of note is that treatment by high performers and treatment by low performers are estimated separately as are value-added estimates based on math tests scores versus ELA test scores. This means that 4 separate specifications using combinations of these items are estimated using the stage three model structure. The key identifying assumption here is that the group of untreated teachers is a reasonable comparison for the treated group such that the parallel trends condition holds. Evidence to support this assumption is in the context of the IDOE data is presented in the following sections.

4 Results

I begin by presenting the results of the stage I and stage II value-added estimates. Estimating equation 2 on the data as specified in the previous section, and standardizing the value-added results to have mean zero and a standard deviation of one yields the distributions presented in Figure 1. The value-added estimates from the Indiana sample are well behaved and roughly normally distributed and, there is no reason to suspect discontinuities in the distribution or that a multi-modal pattern of value-added estimates could effect

stage III estimates. Although value-added estimates using ELA ISTEP+ test scores tend to produce a distribution with a long left tail, this should not impact the results as the actual counts of teachers who fall into the high and low performance categories turn out to be fairly symmetric.

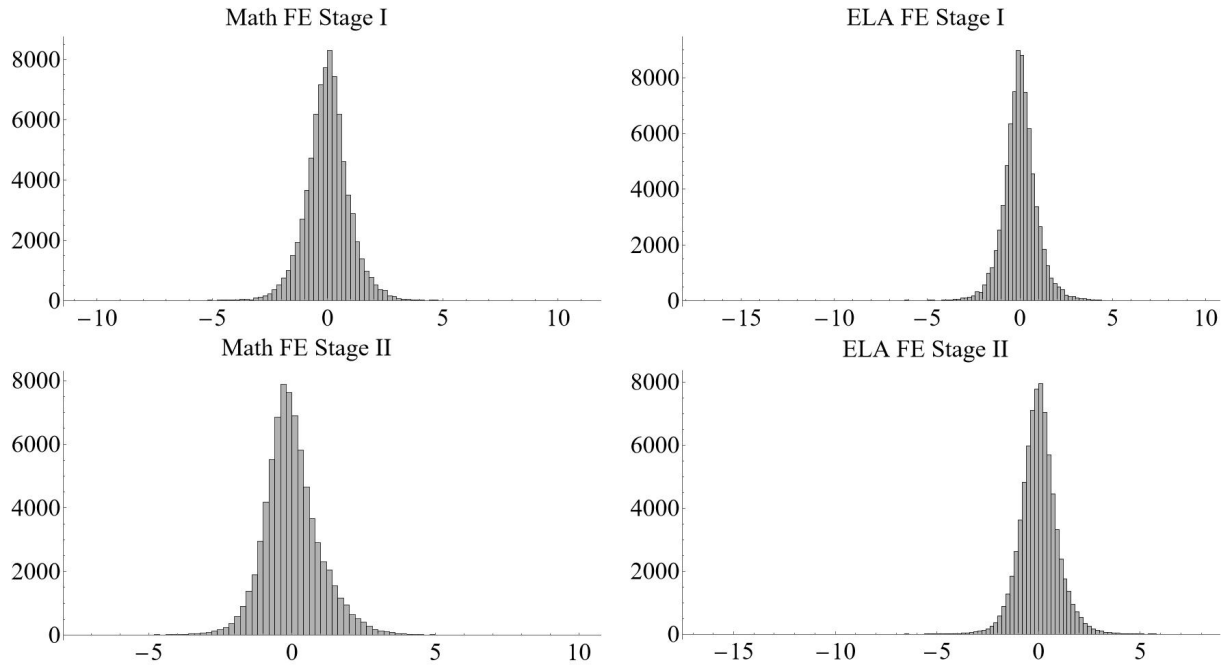

Figure 1: Histograms of Stage I and II Teacher Value-Added Estimates

Table 3 serves to verify this fact by presenting the counts of high and low performers at the various standard deviation cutoffs that will be used in my analysis. The decline in the number of potential treatment inducing teachers as cutoff strength becomes more strict is symmetric and consistent across subject areas and distribution tails. Furthermore, the number of teachers designated as high and low performing teachers remains sufficiently large even at the most extreme of cutoff values suggesting that results derived using cutoff factors of 2 or 2.5 standard deviations retain sufficient power to lend credibility to the results generated from them.

In figure 2, I produce the event studies for a representative set of estimates. This figure was generated using the 7-year treatment specification which is the standard absorbing state design common within the applied literature. I show the results for the 2 standard deviation cutoff which is reasonably representative of the other specifications. The evidence presented here suggest a muddled picture of possible pre-trends such that a story of selection on trend cannot be completely ruled out. It seems plausible that the specifications for math and ELA low performance teachers may hold up to scrutiny. However there appears to be some

Table 3: Counts of High and Low Performing Teachers at Various Value-Added cutoffs by Subject Area

Standard Deviation Cutoff	0.5	1.0	1.5	2.0	2.5
Math High Perf.	5119	2345	1097	529	254
ELA High Perf.	4752	2054	932	414	207
Math Low Perf.	6356	2911	1194	518	261
ELA Low Perf.	6087	2610	1082	500	277

Note: This table contains the counts of high performance (high perf.) teachers and low performance (low perf.) teachers by standard deviation cutoff and subject area. These values were generated by taking stage I value-added estimates and counting the number of teachers at least x standard deviations above the mean for high performance teachers or at least x below the mean for low performers where x is the cutoff given in the column headers. In all cases, value-added estimates have been standardized to have mean 0 and standard deviation 1.

upward pre-trend in the math high performers case and a bit of a saw-tooth pattern for ELA high performers. Despite this, in both of these cases the upward pretend present is followed by a negative effect after treatment. This might suggest that if the results that follow are biased, they are biased downward in magnitude.

Tables 4, 5, 6, and 7 contain the main results of this paper. Each of these tables presents the coefficients of interest only from the stage III two-way fixed effects regression specification given in equation 3. Coefficients are grouped together in tables by the type of treatment event that generated them. As such, four tables are included representing the four possible combinations of high and low performance teachers as measured by their math or ELA test scores. The columns of each table represent a different length of treatment before a teacher is returned to the untreated state. Since an individual can only be treated for a maximum of 7 years when observed over the school years between 2011 and 2018, the right most column is equivalent to the absorbing state specification. Along the rows of the table are the different standard deviation cutoffs tested between 0.5 and 2.5 of a standard deviation above or below the mean. Each table is organized such that coefficients that are closer to the lower right-hand corner represent the longest length of treatment time using the strictest definitions of high and low performers. While estimates in the upper left-hand corner use the shortest and weakest of treatment specifications. The correct interpretation of the coefficients, assuming the identifying assumption is accepted, is the change in value-added in terms of percent change on standard deviation after a high or low quality teacher exits a school for the teachers that remain.

One would expect that if spillovers are present between teachers, then three facts should be observable in the data. First, exits by good teachers would leave their colleagues worse off

Table 4: Stage III Regression Estimates of High Performing Math Teacher Exits Using Math Test Scores

Treat Length	1 Yr	2 Yrs	3 Yrs	4 Yrs	5 Yrs	6 Yrs	7 Yrs
0.5 SD + Mean	-.027*	-.022	-.01	-.008	-.004	-.007	-.005
	(.0157)	(.0164)	(.0182)	(.0194)	(.0204)	(.0213)	(.0218)
1.0 SD + Mean	-.020	-.033**	-.044**	-.056***	-.065***	-.073***	-.077***
	(.0176)	(.0166)	(.0182)	(.0193)	(.0208)	(.0223)	(.0230)
1.5 SD + Mean	-.024	-.054***	-.077***	-.094***	-.106***	-.121***	-.123***
	(.0230)	(.0194)	(.0205)	(.0209)	(.0221)	(.0239)	(.025)
2.0 SD + Mean	-.028	-.075***	-.100***	-.120***	-.131***	-.150***	-.148***
	(.032)	(.0258)	(.0273)	(.0271)	(.0290)	(.0314)	(.0330)
2.5 SD + Mean	-.080*	-.068**	-.091***	-.124***	-.134***	-.151***	-.150***
	(.0413)	(.0312)	(.035)	(.0349)	(.0372)	(.0395)	(.0420)

Note: This table contains two-way fixed effects estimates where coefficients are derived from γ_3 in equation 3. In particular, presented here are the results for math teachers using math test scores to estimate the underlying value-added. The column headers indicate the length of treatment after they experience an exit event from a high performance teacher before other teachers are returned to the untreated pool. Because the maximum length of treatment is 7 years the right most column represents the traditional multiple treatment group absorbing state difference-in-difference specification. The rows indicate the value-added cutoff used to define a high performance teacher. Standard errors for each estimate are presented below the coefficient value in parentheses. Stars follow the conventional pattern: $* = p < 0.1, ** = p < 0.05, *** = p < 0.01$.

Table 5: Stage III Regression Estimates of High Performing ELA Teacher Exits Using ELA Test Scores

Treat Length	1 Yr	2 Yrs	3 Yrs	4 Yrs	5 Yrs	6 Yrs	7 Yrs
0.5 SD + Mean	-.021	-.012	-.031*	-.020	-.036*	-.026	-.026
	(.0160)	(.0163)	(.0173)	(.0191)	(.0208)	(.0223)	(.0232)
1.0 SD + Mean	-.026	-.020	-.041**	-.045**	-.058***	-.057**	-.057**
	(.0188)	(.0167)	(.0178)	(.0194)	(.0209)	(.0231)	(.0241)
1.5 SD + Mean	-.009	-.004	-.015	-.038	-.054**	-.064**	-.063**
	(.0240)	(.0220)	(.0226)	(.0246)	(.0257)	(.0279)	(.0290)
2.0 SD + Mean	-.030	-.021	-.021	-.033	-.056*	-.057*	-.058*
	(.0329)	(.0296)	(.0290)	(.0297)	(.0314)	(.0333)	(.0354)
2.5 SD + Mean	-.042	-.036	-.011	-.025	-.044	-.040	-.051
	(.0504)	(.0426)	(.0405)	(.0399)	(.0411)	(.0428)	(.0456)

Note: This table contains two-way fixed effects estimates where coefficients are derived from γ_3 in equation 3. In particular, presented here are the results for ELA teachers using ELA test scores to estimate the underlying value-added. The column headers indicate the length of treatment after they experience an exit event from a high performance teacher before other teachers are returned to the untreated pool. Because the maximum length of treatment is 7 years the right most column represents the traditional multiple treatment group absorbing state difference-in-difference specification. The rows indicate the value-added cutoff used to define a high performance teacher. Standard errors for each estimate are presented below the coefficient value in parentheses. Stars follow the conventional pattern: $* = p < 0.1, ** = p < 0.05, *** = p < 0.01$.

Table 6: Stage III Regression Estimates of Low Performing Math Teacher Exits Using Math Test Scores

Treat Length	1 Yr	2 Yrs	3 Yrs	4 Yrs	5 Yrs	6 Yrs	7 Yrs
Mean - 0.5 SD	.013	.004	.017	.013	-.001	-.009	-.007
	(.0162)	(.0160)	(.0175)	(.0203)	(.0224)	(.0237)	(.0242)
Mean - 1.0 SD	.011	.015	.028	.021	.010	.009	.003
	(.0187)	(.0159)	(.0171)	(.0190)	(.0213)	(.0228)	(.0241)
Mean - 1.5 SD	.021	.016	.020	.035	.036	.036	.035
	(.0236)	(.0200)	(.0204)	(.0223)	(.0244)	(.0262)	(.0281)
Mean - 2.0 SD	.010	.017	.010	.012	.003	.016	.020
	(.0315)	(.0259)	(.0262)	(.0288)	(.0320)	(.0348)	(.0373)
Mean - 2.5 SD	.048	.052	.043	.044	.035	.041	.052
	(.0463)	(.0349)	(.0360)	(.0394)	(.0436)	(.0478)	(.0510)

Note: This table contains two-way fixed effects estimates where coefficients are derived from γ_3 in equation 3. In particular, presented here are the results for math teachers using math test scores to estimate the underlying value-added. The column headers indicate the length of treatment after they experience an exit event from a low performing teacher before other teachers are returned to the untreated pool. Because the maximum length of treatment is 7 years the right most column represents the traditional multiple treatment group absorbing state difference-in-difference specification. The rows indicate the value-added cutoff used to define a low performing teacher. Standard errors for each estimate are presented below the coefficient value in parentheses. Stars follow the conventional pattern: $* = p < 0.1, ** = p < 0.05, *** = p < 0.01$.

Table 7: Stage III Regression Estimates of Low Performing ELA Teacher Exits Using ELA Test Scores

Treat Length	1 Yr	2 Yrs	3 Yrs	4 Yrs	5 Yrs	6 Yrs	7 Yrs
Mean - 0.5 SD	.053***	.018	.014	.009	-.002	-.001	.002
	(.0162)	(.0164)	(.0181)	(.0203)	(.0218)	(.0231)	(.0239)
Mean - 1.0 SD	.054***	.044***	.052***	.051***	.049**	.059***	.064***
	(.0183)	(.0160)	(.0170)	(.0188)	(.0206)	(.0228)	(.0243)
Mean - 1.5 SD	.05**	.036*	.052***	.073***	.083***	.101***	.100***
	(.0237)	(.0197)	(.0202)	(.0211)	(.0233)	(.0254)	(.0277)
Mean - 2.0 SD	.057*	.048*	.059**	.068***	.077***	.092***	.101***
	(.0310)	(.0265)	(.0260)	(.0271)	(.0291)	(.0319)	(.0347)
Mean - 2.5 SD	.015	.035	.052	.072**	.072**	.091**	.100**
	(.0412)	(.0340)	(.0343)	(.0354)	(.0366)	(.0398)	(.0428)

Note: This table contains two-way fixed effects estimates where coefficients are derived from γ_3 in equation 3. In particular, presented here are the results for ELA teachers using ELA test scores to estimate the underlying value-added. The column headers indicate the length of treatment after they experience an exit event from a low performing teacher before other teachers are returned to the untreated pool. Because the maximum length of treatment is 7 years the right most column represents the traditional multiple treatment group absorbing state difference-in-difference specification. The rows indicate the value-added cutoff used to define a low performing teacher. Standard errors for each estimate are presented below the coefficient value in parentheses. Stars follow the conventional pattern: $* = p < 0.1, ** = p < 0.05, *** = p < 0.01$.

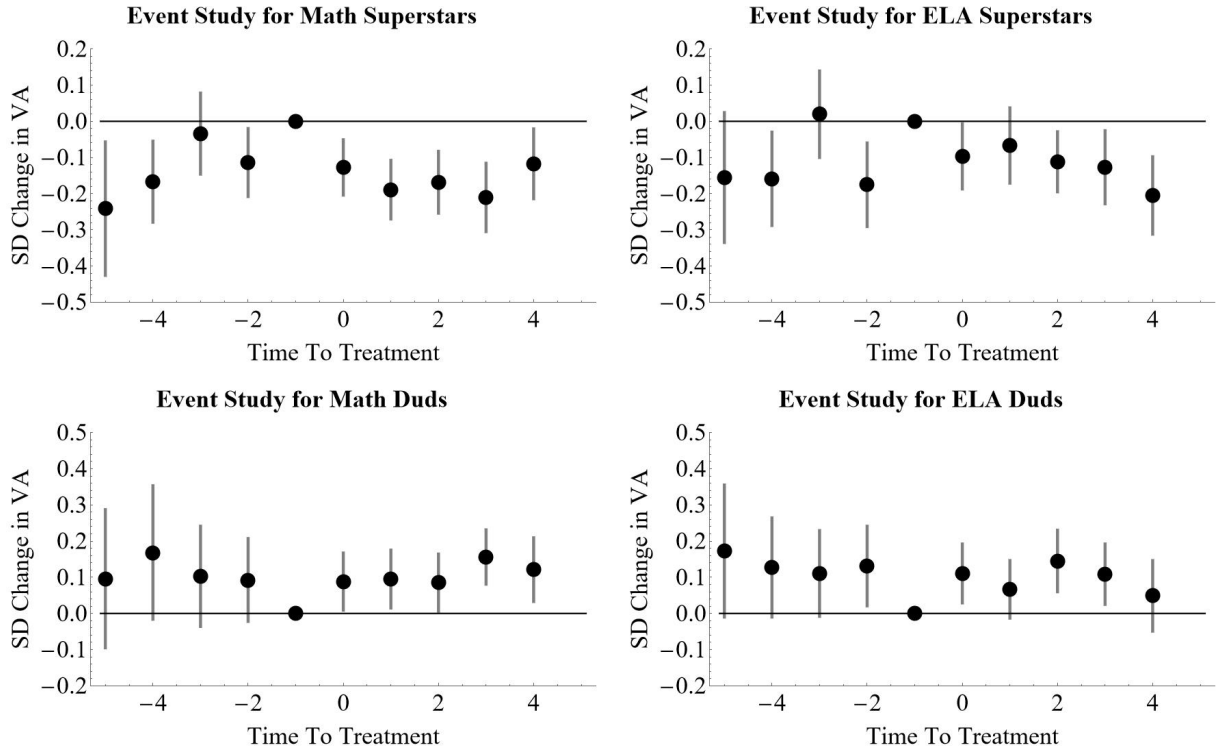

Figure 2: Event Studies of Absorbing State Treatment Specification at 2 SD Cutoff

and exits by bad teachers would improve the quality of the teacher who stay behind. Second, as the treatment length and strictness of high and low performance teachers increases, so too would the magnitude and significance of the effect. This makes sense because decay of any spillover effects would occur gradually. Teachers do not immediately shed the impacts of high or low performance teachers after they exit a school. Instead, the knowledge effects increase or decrease gradually the further they are removed from the treatment inducing teacher in time. Finally, as I increase the cutoff defining high and low performance teachers, the number of treated teachers in the sample declines which would increase standard errors. This is exactly what is observed in the data. In table 4, we see strong negative and significant effects with magnitudes between -0.025 and -0.15 of a standard deviation which implies that when good math teachers leave a school, their colleagues are worse off without them. In other words, a good math teacher produces positive and significant peer effects on the teachers they work with. Similarly, table 7 reports a symmetric story for teachers treated by the exit of a bad English teacher. Here I estimate that upon the exit of a poor quality colleague, other ELA teachers are better off by one tenth of a standard deviation in the strictest specification. Taken together, these results suggest that spillover effects between-teachers can have a substantial impact on their quality and that these effects should not be ignored when applying a value-added model.

Interestingly, the results I highlight above are not as evident in the other specifications. Indeed, I find a weak statistically significant negative effect from high performer ELA exits and a null effect from low performing math teacher exits in tables 5 and 6. It is unclear from the results alone why this might be the case, but it may suggest that some asymmetries are present between the specific skills required to be a successful math teacher versus a high quality ELA teacher. This is not an unusual finding and is supported by the work of Nye et al. (2004) which shows that subject specialization and differences in magnitude on value-added measure using subject specific test scores is more the norm than the exception.

5 Robustness

Alternative Specification - Treatment Via Teacher Entry

In my main specification, I measure teacher spillover effects from exit events, that is cases where teachers take a leave of absence or move from one school to another. Naturally, it is possible to consider the alternative case where treatment events are generated from teacher entries into the public school system. I define "entry events" as any case where a high or low performing teacher moves from one school to another in consecutive school years generating a treatment event on the school they move to, when a teacher enters the public schooling system as a new teacher, or returns to work after a leave of absence. A teacher is considered treated in year t in this setting if they were at a school that experienced an entry event from a high or low performing teacher in year t as well.

I elect to focus on teacher exit events as my main specification and leave the entry case as a robustness check because the identification story is much cleaner for the exit case. Consider that an exit event is a one time shock that has an immediate impact in the year after it occurs. Teachers treated by an exit event are exposed to treatment once, and the treatment event is not dependent on future changes in the quality of spillover generating teachers. Entry events, on the other hand, have both an immediate shock effect in the year of treatment followed by a long term effect from continued exposure in the subsequent years. Disentangling the continuous exposure effect from the initial treatment effect is impossible to do without additional identifying assumptions and so only exits were considered in the main results of this paper. That said, however, the entry case is certainly worth looking at if only to verify that the directions and magnitudes of my main specification estimates are reasonable and accurate.

That being said, there is still value in studying the entry event specification results if only

Table 8: Stage III Regression Estimates of High Performing Math Teacher Entries Using Math Test Scores

Treat Length	1 Yr	2 Yrs	3 Yrs	4 Yrs	5 Yrs	6 Yrs	7 Yrs
0.5 SD + Mean	.161***	.125***	.12***	.135***	.141***	.155***	.161***
	(.0165)	(.0168)	(.018)	(.0199)	(.0219)	(.0233)	(.0244)
1.0 SD + Mean	.203***	.146***	.11***	.118***	.126***	.137***	.143***
	(.0195)	(.0170)	(.0174)	(.0196)	(.0214)	(.0229)	(.0244)
1.5 SD + Mean	.24***	.169***	.13***	.114***	.115***	.114***	.12***
	(.0247)	(.0198)	(.0205)	(.0230)	(.0246)	(.0259)	(.0275)
2.0 SD + Mean	.256***	.154***	.115***	.081***	.074**	.073**	.071**
	(.0354)	(.0272)	(.0269)	(.0295)	(.0310)	(.0321)	(.0338)
2.5 SD + Mean	.339***	.171***	.113***	.064	.052	.05	.057
	(.0489)	(.0377)	(.0343)	(.0394)	(.0411)	(.0433)	(.0450)

Note: This table contains two-way fixed effects estimates where coefficients are derived from γ_3 in equation 3. In particular, presented here are the results for math teachers using math test scores to estimate the underlying value added. The column headers indicate the length of treatment after they experience an entry event from a high performer before other teachers are returned to the untreated pool. Because the maximum length of treatment is 7 years the right most column represents the traditional multiple treatment group absorbing state difference-in-difference specification. The rows indicate the VA cutoff used to define a high performance teacher. Standard errors for each estimate are presented below the coefficient value in parentheses. Stars follow the conventional pattern: $* = p < 0.1, ** = p < 0.05, *** = p < 0.01$.

Table 9: Stage III Regression Estimates of High Performing ELA Teacher Entries Using ELA Test Scores

Treat Length	1 Yr	2 Yrs	3 Yrs	4 Yrs	5 Yrs	6 Yrs	7 Yrs
0.5 SD + Mean	.179***	.146***	.138***	.141***	.144***	.157***	.169***
	(.0166)	(.0163)	(.0176)	(.0197)	(.0213)	(.0233)	(.0246)
1.0 SD + Mean	.177***	.133***	.113***	.107***	.118***	.128***	.137***
	(.0211)	(.0176)	(.0181)	(.0198)	(.0217)	(.0239)	(.0258)
1.5 SD + Mean	.199***	.151***	.12***	.122***	.124***	.138***	.156***
	(.0275)	(.0223)	(.0219)	(.0228)	(.0248)	(.0272)	(.0297)
2.0 SD + Mean	.215***	.141***	.123***	.109***	.1***	.115***	.134***
	(.0413)	(.0323)	(.0299)	(.0301)	(.0333)	(.0362)	(.0391)
2.5 SD + Mean	.254***	.169***	.132***	.113**	.104**	.119**	.138**
	(.0564)	(.0484)	(.0451)	(.0442)	(.0471)	(.0519)	(.0559)

Note: This table contains two-way fixed effects estimates where coefficients are derived from γ_3 in equation 3. In particular, presented here are the results for ELA teachers using ELA test scores to estimate the underlying value added. The column headers indicate the length of treatment after they experience an entry event from a high performer before other teachers are returned to the untreated pool. Because the maximum length of treatment is 7 years the right most column represents the traditional multiple treatment group absorbing state difference-in-difference specification. The rows indicate the VA cutoff used to define a high performance teacher. Standard errors for each estimate are presented below the coefficient vale in parentheses. Stars follow the conventional pattern: $* = p < 0.1, ** = p < 0.05, *** = p < 0.01$.

Table 10: Stage III Regression Estimates of Low Performing Math Teacher Entries Using Math Test Scores

Treat Length	1 Yr	2 Yrs	3 Yrs	4 Yrs	5 Yrs	6 Yrs	7 Yrs
Mean - 0.5 SD	-.081***	-.075***	-.068***	-.076***	-.075***	-.083***	-.089***
	(.0162)	(.0164)	(.0186)	(.0195)	(.0209)	(.0217)	(.0218)
Mean - 1.0 SD	-.065***	-.056***	-.048***	-.053***	-.051**	-.062***	-.064***
	(.0184)	(.0167)	(.0179)	(.0192)	(.0210)	(.0226)	(.0232)
Mean - 1.5 SD	-.085***	-.075***	-.064***	-.064***	-.066**	-.062**	-.07**
	(.0260)	(.0222)	(.0233)	(.0238)	(.0254)	(.0275)	(.0289)
Mean - 2.0 SD	-.073*	-.072**	-.061**	-.064**	-.054	-.057	-.061
	(.0385)	(.0325)	(.0307)	(.0316)	(.0334)	(.0361)	(.0373)
Mean - 2.5 SD	-.141***	-.089**	-.057	-.067*	-.067	-.085*	-.089*
	(.0491)	(.0431)	(.0378)	(.0402)	(.0433)	(.0474)	(.0490)

Note: This table contains two-way fixed effects estimates where coefficients are derived from γ_3 in equation 3. In particular, presented here are the results for math teachers using math test scores to estimate the underlying value added. The column headers indicate the length of treatment after they experience an entry event from a low performer before other teachers are returned to the untreated pool. Because the maximum length of treatment is 7 years the right most column represents the traditional multiple treatment group absorbing state difference-in-difference specification. The rows indicate the VA cutoff used to define a low performance teacher. Standard errors for each estimate are presented below the coefficient value in parentheses. Stars follow the conventional pattern: $* = p < 0.1, ** = p < 0.05, *** = p < 0.01$.

Table 11: Stage III Regression Estimates of Low Performing ELA Teacher Entries Using ELA Test Scores

Treat Length	1 Yr	2 Yrs	3 Yrs	4 Yrs	5 Yrs	6 Yrs	7 Yrs
Mean - 0.5 SD	-.156***	-.133***	-.138***	-.159***	-.168***	-.177***	-.182***
	(.0167)	(.017)	(.0188)	(.0204)	(.0215)	(.0227)	(.0231)
Mean - 1.0 SD	-.165***	-.129***	-.116***	-.125***	-.134***	-.143***	-.154
	(.0175)	(.0173)	(.0181)	(.0198)	(.0215)	(.0234)	(.0248)
Mean - 1.5 SD	-.174***	-.135***	-.106***	-.095***	-.094***	-.091***	-.094***
	(.0252)	(.0211)	(.0222)	(.0239)	(.0257)	(.0278)	(.0295)
Mean - 2.0 SD	-.164***	-.13***	-.098***	-.100***	-.103***	-.102***	-.103**
	(.0367)	(.0291)	(.029)	(.0302)	(.0339)	(.0360)	(.0379)
Mean - 2.5 SD	-.12**	-.125***	-.066*	-.061*	-.064	-.053	-.047
	(.0469)	(.0355)	(.0364)	(.0362)	(.0414)	(.0443)	(.0465)

Note: This table contains two-way fixed effects estimates where coefficients are derived from γ_3 in equation 3. In particular, presented here are the results for ELA teachers using ELA test scores to estimate the underlying value added. The column headers indicate the length of treatment after they experience an entry event from a low performer before other teachers are returned to the untreated pool. Because the maximum length of treatment is 7 years the right most column represents the traditional multiple treatment group absorbing state difference-in-difference specification. The rows indicate the VA cutoff used to define a low performance teacher. Standard errors for each estimate are presented below the coefficient value in parentheses. Stars follow the conventional pattern: $* = p < 0.1, ** = p < 0.05, *** = p < 0.01$.

to confirm the direction and general magnitudes of effect from the main results. Tables 8, 9, 10, and 11 present the results of these regressions. Treatment length and influential teacher cutoff definitions remain the same despite the concerns laid out above in order to facilitate easy comparison with the main results. I find that strong positive and significant spillover effects in the rage of 0.1-0.7 of a standard deviation are present from high performance teachers while low performance teachers produce -0.5 to -0.18 of a standard deviation in spillover effects. These estimates are largely in line with the main results though slightly larger in effect size. They also exhibit the same patterns of standard errors and effect sizes as treatment length and cutoff strictness vary. Interestingly, I do not observe the asymmetry between subject areas that was on display in the main results as statistically significant spillover effects are present in both the math and ELA cases.

Placebo Test - Exits by Middling Teachers

I can further support my results by considering the case of exits by teachers from the middle of the distribution as a placebo test. I have argued that it is proper to focus on high and low performers as they are the most likely to generate spillover effects, and that the further into the tails of the distribution a teacher is, the stronger the effects I ought to observe. We can put these two statements to the test and lend evidence to dismiss the idea the results of this paper are explained by noise in teacher quality related to when they are treated by running the main specification models on middling teachers. For this test, I define a middling teacher (a teacher in the middle of the distribution) as a teacher with a value-added between -0.5 and +0.5 of a standard deviation away from the mean. These teachers are then tracked over time for exit events which generate the treatment events on other teachers just like the estimation procedure from the main specification. Equation 3 is then re-estimated using these treatment events, the results of which are presented in table 12. If the quality of a teacher exiting a school is driving the results of this paper rather than the exit event itself or some other policy modifying event like changes in teacher compensation, accountability measurement, or statewide demographic shift, I should expect to see a null effect from this placebo test.

As predicted, I do find a null effect from middling teachers in all treatment length specifications and regardless of which test scores are used to determine value-added. This helps to rule out the possibility of a spurious relationship between teacher quality post exit event. These results also help to defend the exogeneity assumption set out in the introduction of this paper.

Table 12: Stage III Regression Estimates of Math and ELA Middling Teacher Exit Events

Treat Length	1 Yr	2 Yrs	3 Yrs	4 Yrs	5 Yrs	6 Yrs	7 Yrs
Math Middling Teachers	.024	.023	.012	.002	-.019	-.019	-.023
	(.0188)	(.0231)	(.0271)	(.0269)	(.0317)	(.0324)	(.0327)
ELA Middling Teachers	.006	.005	.032	.033	.035	.046	.044
	(.0220)	(.0262)	(.0300)	(.0314)	(.0331)	(.0338)	(.0340)

Note: This table contains two-way fixed effects estimates where coefficients are derived from γ_3 in equation 3. In particular, presented here are the results based on value-added estimates from both math and ELA test scores. The column headers indicate the length of treatment after a teacher network experiences an exit event from a middling teacher before other teachers are returned to the untreated pool. Because the maximum length of treatment is 7 years the right most column represents the traditional multiple treatment group absorbing state difference-in-difference specification. The rows indicate the VA cutoff used to define a low performing teacher. Standard errors for each estimate are presented below the coefficient value in parentheses. Stars follow the conventional pattern: $* = p < 0.1, ** = p < 0.05, *** = p < 0.01$.

6 Discussion

This paper makes two important contributions to the value-added modeling literature in education. First, I propose a new method of identifying teacher spillover effects which, if not accounted for in value-added models, could lead to biased estimates of teacher quality. The method I propose exploits the movement of top and bottom preforming teachers between schools and exits from employment in public education entirely as observable shocks to the networks of teachers that work together in schools. Doing so allows for the estimation of the impact of that teacher on their fellow peers via a difference-in-difference structure. The result is a bounded estimate of the magnitude to which teacher-to-teacher spillover effects exist within the context of the data set it is applied to. Naturally, questions of validity would arise and so, to lend some support to this method, in my second contribution I offer a first of its kind estimate of between-teacher spillover effects using administrative data from the Indiana Department of Education. In doing so, I find that high quality math teachers impart a quality boost of 0.05 to 0.15 of a standard deviation on their colleagues. Likewise, poor quality English instructors leave their fellow teachers worse off by a factor of 0.05 to 0.1 of a standard deviation. I go on to show that these estimates are robust to regression specification and treatment structure.

Comparing these results to the rest of the value added literature reveals that these are relatively large effects suggesting teacher spillovers represent a substantial source of bias which scales with the quality of the teacher. To put these results into context it is first useful to compare the magnitudes of my results to other policy interventions intended to impact value added. My estimates suggest that the magnitude of between-teacher spillovers is larger than previously estimated by value-added modeling alone at 0.08 of a standard

deviation (Jackson & Bruegmann, 2009), larger than firing the bottom 5% of teachers at 0.04 of a standard deviation (Hanushek, 2011), and about as large as an incentive pay program for teachers at 0.1 of a standard deviation (Kho et al., 2022) or increasing the mix of english language learning students in a classroom at -0.1 of a standard deviation (Hanushek & Rivkin, 2010). But to understand the policy implications of my results its important to note that the kind of bias I point out here with between-teacher spillovers does not impact the ranking of teachers relative to each other as long as the composition of the pool of teachers remains the same. This means that using percentile ordering of teachers remains a valid way to sort and compare groups of faculty members. Instead, what these results suggest is that when these network effects are created or broken due to compositional changes, the effects observed on the average teacher are either incorrect or miss-attributed. Essentially, my results suggest that good teachers are being relatively undervalued because their positive network effects are not being taken into account while poor teachers are even more detrimental to a school as a whole due to their negative network effects. What this might suggest is that policies which incentivize the retention of high quality teachers and the removal of low quality teachers have larger effects than we would have previously expected. Certainly, a natural followup to this paper would be to re-evaluate these types of policies in lieu of the between-teacher spillover effects I reveal to be present here.

I can offer something along these lines with a quick back of the envelope calculation using number from Hanushek (2011) Scaling his measure effect size on student earning to a mid-range significant effect of 0.1 standard deviations above the mean which I find for high performing math teachers, for example, suggests that between-teacher spillovers may account for approximately an additional $4000 in present value student earnings annually per 20 students. Furthermore, my results would suggest that replacing low quality teachers would have a larger effect size than previously expected. Hanusheck (2011) estimates that a dismissal of 5-8% of the worst performing teachers and replacement with average teachers could move the US education system to world leading test scores in math and reading. Accounting for my spillover measures would put the range or removal and replacement in the 3-6% range implying much less turnover of the worst performing teachers would be required to make a sizable impact on the US education system then previously thought. Additional research is needed to confirm these results, however.

References

Aaronson, D., Barrow, L., & Sander, W. (2007, January). Teachers and student achievement in the chicago public high schools. *Journal of Labor Economics*, *25*(1), 95-135. Retrieved from http://dx.doi.org/10.1086/508733 doi: 10.1086/508733

Adnot, M., Dee, T., Katz, V., & Wyckoff, J. (2016, September). Teacher turnover, teacher quality, and student achievement in dcps. *Educational Evaluation and Policy Analysis*, *39*(1), 54-76. Retrieved from http://dx.doi.org/10.3102/0162373716663646 doi: 10.3102/0162373716663646

Azoulay, P., Zivin, J. S. G., & Wang, J. (2010). Superstar Extinction. *The Quarterly Journal of Economics*, *125*(2), 549–589. Retrieved 2022-08-31, from http://www.jstor.org/stable/27867490

Bacher-Hicks, A., Chin, M. J., Kane, T. J., & Staiger, D. O. (2019). An experimental evaluation of three teacher quality measures: Value-added, classroom observations, and student surveys. *Economics of Education Review*, *73*, 101919. Retrieved from https://www.sciencedirect.com/science/article/pii/S0272775719302717 doi: https://doi.org/10.1016/j.econedurev.2019.101919

Backes, B., Cowan, J., Goldhaber, D., Koedel, C., Miller, L. C., & Xu, Z. (2018). The common core conundrum: To what extent should we worry that changes to assessments will affect test-based measures of teacher performance? *Economics of Education Review*, *62*, 48-65. Retrieved from https://www.sciencedirect.com/science/article/pii/S0272775717301759 doi: https://doi.org/10.1016/j.econedurev.2017.10.004

Ben-Porath, Y. (1967). The Production of Human Capital and the Life Cycle of Earnings. *Journal of Political Economy*, *75*(4), 352–365. Retrieved 2022-11-10, from http://www.jstor.org/stable/1828596

Boyd, D. J., Grossman, P. L., Lankford, H., Loeb, S., & Wyckoff, J. (2009, December). Teacher Preparation and Student Achievement. *Educational Evaluation and Policy Analysis*, *31*(4), 416-440. Retrieved from http://dx.doi.org/10.3102/0162373709353129 doi: 10.3102/0162373709353129

Bruno, P., & Strunk, K. O. (2019, August). Making the Cut: The Effectiveness of Teacher Screening and Hiring in the Los Angeles Unified School District. *Educational Evaluation and Policy Analysis*, *41*(4), 426-460. Retrieved from http://dx.doi.org/10.3102/0162373719865561 doi: 10.3102/0162373719865561

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014a). Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates. *The American Economic Review*, *104*(9), 2593–2632. Retrieved 2022-11-10, from http://www.jstor.org/

stable/43495327

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014b). Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood. *The American Economic Review*, *104*(9), 2633–2679. Retrieved 2022-11-10, from http://www.jstor.org/stable/43495328

Cullen, J. B., Koedel, C., & Parsons, E. (2021, 01). The Compositional Effect of Rigorous Teacher Evaluation on Workforce Quality. *Education Finance and Policy*, *16*(1), 7-41. Retrieved from https://doi.org/10.1162/edfp_a_00292 doi: 10.1162/edfp_a_00292

Darling-Hammond, L. (2015). Can Value Added Add Value to Teacher Evaluation? *Educational Researcher*, *44*(2), 132–137. Retrieved 2022-11-10, from http://www.jstor.org/stable/24571540

Gershenson, S., Lindsay, C., Papageorge, N., Campbell, R., & Rendon, J. (2023, November). Spillover Effects at School: How Black Teachers affect their White Peers' Racial Competency. *NBER Working Paper*(31847). Retrieved from http://dx.doi.org/10.3386/w31847 doi: 10.3386/w31847

Gilraine, M., & McCarthy, O. (2024, 02). The Effect of the Prior Teacher on Value-Added. *The Review of Economics and Statistics*, 1-45. Retrieved from https://doi.org/10.1162/rest_a_01409 doi: 10.1162/rest_a_01409

Goldhaber, D., & Chaplin, D. D. (2015). Assessing the "Rothstein Falsification Test": Does It Really Show Teacher Value-Added Models Are Biased? *Journal of Research on Educational Effectiveness*, *8*(1), 8-34. Retrieved from https://doi.org/10.1080/19345747.2014.978059 doi: 10.1080/19345747.2014.978059

Guarino, C. M., Reckase, M. D., Stacy, B. W., & Wooldridge, J. M. (2015, January). Evaluating specification tests in the context of value-added estimation. *Journal of Research on Educational Effectiveness*, *8*(1), 35-59. Retrieved from http://dx.doi.org/10.1080/19345747.2014.981905 doi: 10.1080/19345747.2014.981905

Guarino, C. M., Reckase, M. D., & Wooldridge, J. M. (2015). Can Value-Added Measures of Teacher Performance Be Trusted? *Education Finance and Policy*, *10*(1), 117-156.

Hanushek, E. A. (1979). Conceptual and Empirical Issues in the Estimation of Educational Production Functions. *The Journal of Human Resources*, *14*(3), 351–388. Retrieved 2022-11-10, from http://www.jstor.org/stable/145575

Hanushek, E. A. (2011). The economic value of higher teacher quality. *Economics of Education Review*, *30*(3), 466-479. Retrieved from https://www.sciencedirect.com/science/article/pii/S0272775710001718 doi: https://doi.org/10.1016/j.econedurev.2010.12.006

Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *The American Economic Review*, *100*(2), 267–271. Retrieved 2024-10-04, from http://www.jstor.org/stable/27805002

Henry, G. T., & Redding, C. (2020). The Consequences of Leaving School Early: The Effects of Within-Year and End-of-Year Teacher Turnover. *Education Finance and Policy*, *15*(2), 332-356.

Jackson, C. K. (2018, October). What Do Test Scores Miss? The Importance of Teacher Effects on Non-Test Score Outcomes. *Journal of Political Economy*, *126*(5), 2072-2107. Retrieved from http://dx.doi.org/10.1086/699018 doi: 10.1086/699018

Jackson, C. K., & Bruegmann, E. (2009). Teaching students and teaching each other: The importance of peer learning for teachers. *American Economic Journal: Applied Economics*, *1*(4), 85–108. Retrieved 2024-10-04, from http://www.jstor.org/stable/25760183

Kane, T., & Staiger, D. (2008). Estimating teacher impacts on student achievement: An experimental evaluation. *NBER Working Paper Series*(14607). Retrieved from http://dx.doi.org/10.3386/w14607 doi: 10.3386/w14607

Kho, A., Henry, G. T., Pham, L. D., & Zimmer, R. (2022). Spillover Effects of Recruiting Teachers for School Turnaround: Evidence From Tennessee. *Evaluation and Policy Analysis*, 1-17.

Kinsler, J. (2012). Assessing Rothstein's critique of teacher value-added models. *Quantitative Economics*, *3*(2), 333-362. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.3982/QE132 doi: https://doi.org/10.3982/QE132

Koedel, C. (2009). An empirical analysis of teacher spillover effects in secondary school. *Economics of Education Review*, *28*(6), 682-692. Retrieved from https://www.sciencedirect.com/science/article/pii/S0272775709000430 doi: https://doi.org/10.1016/j.econedurev.2009.02.003

Koedel, C., & Betts, J. R. (2011). Does Student Sorting Invalidate Value-Added Models of Teacher Effectiveness? An Extended Analysis of the Rothstein Critique. *Education Finance and Policy*, *6*(1), 18–42. Retrieved 2024-08-08, from http://www.jstor.org/stable/educfinapoli.6.1.18

Koedel, C., Mihaly, K., & Rockoff, J. E. (2015). Value-Added Modeling: A Review. *Economics of Education Review*, *47*, 180-195. Retrieved from https://www.sciencedirect.com/science/article/pii/S0272775715000072 doi: https://doi.org/10.1016/j.econedurev.2015.01.006

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004, September). How large are teacher effects? *Educational Evaluation and Policy Analysis*, *26*(3), 237–257.

Retrieved from http://dx.doi.org/10.3102/01623737026003237 doi: 10.3102/01623737026003237

Opper, I. M. (2019). Does Helping John Help Sue? Evidence of Spillovers in Education. *The American Economic Review*, *109*(3), 1080–1115. Retrieved 2022-08-31, from https://www.jstor.org/stable/26602977

Papay, J. P. (2011, February). Different Tests, Different Answers: The Stability of Teacher Value-Added Estimates Across Outcome Measures. *American Educational Research Journal*, *48*(1), 163–193. Retrieved from http://dx.doi.org/10.3102/0002831210362589 doi: 10.3102/0002831210362589

Ronfeldt, M., Farmer, S. O., McQueen, K., & Grissom, J. A. (2015). Teacher Collaboration in Instructional Teams and Student Achievement. *American Educational Research Journal*, *52*(3), 475–514. Retrieved 2022-11-10, from http://www.jstor.org/stable/24546739

Rothstein, J. (2009, 10). Student Sorting and Bias in Value-Added Estimation: Selection on Observables and Unobservables. *Education Finance and Policy*, *4*(4), 537-571. Retrieved from https://doi.org/10.1162/edfp.2009.4.4.537 doi: 10.1162/edfp.2009.4.4.537

Rothstein, J. (2010). Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement. *The Quarterly Journal of Economics*, *125*(1), 175–214. Retrieved 2022-11-10, from http://www.jstor.org/stable/40506280

Stacy, B., Guarino, C., & Wooldridge, J. (2018). Does the precision and stability of value-added estimates of teacher performance depend on the types of students they serve? *Economics of Education Review*, *64*, 50-74. Retrieved from https://www.sciencedirect.com/science/article/pii/S0272775717300596 doi: https://doi.org/10.1016/j.econedurev.2018.04.001

Steele, J. L., Pepper, M. J., Springer, M. G., & Lockwood, J. (2015). The Distribution and Mobility of Effective Teachers: Evidence from a Large, Urban School District. *Economics of Education Review*, *48*, 86-101. Retrieved from https://www.sciencedirect.com/science/article/pii/S0272775715000758 doi: https://doi.org/10.1016/j.econedurev.2015.05.009

Todd, P. E., & Wolpin, K. I. (2003). On the Specification and Estimation of the Production Function for Cognitive Achievement. *The Economic Journal*, *113*(485), F3–F33.