What Do Classroom Observation Scores Tell Us About Student Success? Capturing the Impact of Teachers Using At-Scale Classroom Observation Scores

By

Sy Doan

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

K-12 Leadership & Policy Studies

09 August 2019

Nashville, Tennessee

Approved:

R. Dale Ballou, Ph.D.

Matthew G. Springer, Ph.D.

Christopher A. Candelaria, Ph.D.

C. Kirabo Jackson, Ph.D.

ACKNOWLEDGMENTS

First, I would like to acknowledge Dale Ballou, whose thoughtfulness and dedication have guided me throughout my time at Vanderbilt. My time and conversations with Dale serve as my north star in understanding how to be a better researcher, teacher, and mentor as I continue my scholarly career. Matt and Chris, thank you for your attention, perspective, and collaboration over the course of my PhD. I am indebted to Kirabo and many others who have provided valuable feedback throughout the dissertation process.

Next, I would like to thank the Tennessee Education Research Alliance (TERA). Erin, Susan, Jessica, and Matthew, among many TERA staff, have supported me as both a researcher and a human being from day one and I could not have asked for a better research home during my time at Vanderbilt. I was blessed to spend five years with so many kind and talented PhD students. To my cohort, Sarah, Susan, Laura, Jenna, CJ, and Tuan, I've learned so much from all of you about how to lead lives of purpose and compassion, all while throwing in a little bit of Stata on the side. Walker, thank you for being a constant source of support and inspiration.

Lastly, to my family, I am driven every day by your love and dedication. Mom and Dad, not a day goes by without me thinking about the sacrifices you made to get us to where we are today. I am very happy that you will be able to say that you now have a doctor in the family, even if it's not quite the type of doctor you envisioned. Bo and Mio, your unconditional support keeps me going during those times when everything gets a bit tougher.

For me to complete this program, there had to be many. I am grateful beyond measure.

TABLE OF CONTENTS

Page

ACKNOWLEDGMENTS iii

LIST OF TABLES vi

LIST OF FIGURES vii

Chapter

1 Introduction 1

1.1 Objective 1
1.2 Research Questions 2
1.3 Contributions 4

2 Literature Review 6

2.1 Teacher Evaluation in the United States 6
2.2 Validating Teacher Quality Measures 8
2.2.1 Value-Added Effects on Test Scores 9
2.2.2 Value-Added Effects on Non-Test Score Outcomes 11
2.2.3 Classroom Observation Scores 13

3 Background & Data 16

3.1 The TEAM Evaluation System 16
3.1.1 Classroom Observations 17
3.1.2 Student Growth 19
3.1.3 Achievement 22
3.1.4 Levels of Effectiveness 22
3.2 Data 22
3.2.1 Student-Teacher Rosters 23
3.2.2 Student and Teacher Demographics 24

3.2.3 Teacher Quality Measures . . . 24
3.2.4 Student Outcomes . . . 25
3.2.5 Relationships Between K-12 and Long-Run Student Outcomes . . . 28

4 Methods . . . 36

4.1 Linking Teacher Quality Measures and Teacher Effects . . . 36
4.2 Measurement Error and Classroom Observation Scores . . . 39
4.2.1 Producing Adjusted Observation Scores . . . 43
4.3 Identifying Teacher Effects Using Observation Scores . . . 52
4.3.1 Potential Threats to Validity in the Teacher Switching Design . . . 56
4.3.2 Comparing and Combining Observation and Value-Added Scores . . . 59

5 Results . . . 67

5.1 Time-Varying K-12 Outcomes . . . 67
5.1.1 Using Zero-Imputation to Account for Missing Teacher Data . . . 70
5.1.2 Comparing Methods for Accounting for Rater Error . . . 71
5.1.3 K-12 Observation Score Effects by Grade Level . . . 75
5.1.4 Domain-Specific Effects . . . 77
5.2 Long-Run Student Outcomes . . . 78
5.2.1 Using "Back-Projected" scores . . . 78
5.2.2 Teacher Effects on Long-Run Outcomes . . . 80
5.2.3 Long-Run Observation Score Effects by Grade Level . . . 83
5.3 Comparing and Combining Observation and Value-Added Effects . . . 85

6 Conclusion . . . 124

REFERENCES . . . 128

LIST OF TABLES

Table Page

3.1 Sample Descriptive Statistics 35

4.1 Year-Specific Autocorrelations of Observation Scores 63
4.2 Observation Scores and Classroom Characteristics (Item Level) 64
4.3 Leave Rater(s) Out Data for Mr. Johnson 65
4.4 Correlations Across Observation and Value-Added Scores 66

5.1 LYO Observation Score Effects on Achievement 100
5.2 LYO Observation Score Effects on Absences 101
5.3 LYO Observation Score Effects on Suspensions 102
5.4 Zero-Imputed LYO Effects on K-12 Outcomes 103
5.5 Rater Fixed Effect Adjusted (RFE) Effects on K-12 Outcomes 104
5.6 Leave-Rater-Out (LRO) Effects on K-12 Outcomes 105
5.7 Unadjusted Observation Score Effects on K-12 Outcomes 106
5.8 Correlations Across Domain Scores 107
5.9 Domain-Specific Effects on Achievement 108
5.10 Domain-Specific Effects on Absences 109
5.11 Domain-Specific Effects on Suspensions 110
5.12 Back-Projected Observation Score Effects on K-12 Outcomes 111
5.13 LYO Observation Score Effects on On-Time HS Graduation 112
5.14 LYO Observation Score Effects on Post-Secondary Enrollment at Age 19 113
5.15 LYO Observation Score Effects on Post-Secondary Completion at Age 23 114
5.16 LYO Observation Score Effects on In-State Wages 115
5.17 Single and Multiple Measure Effects on Achievement (Matched Sample) 116
5.18 Single and Multiple Measure Effects on Suspensions (Matched Sample) 117
5.19 Single and Multiple Measure Effects on On-Time HS Graduation (Matched Sample) 118
5.20 Single and Multiple Measure Effects on Absences (Matched Sample) 119
5.21 Single and Multiple Measure Effects on PS Enrollment at 19 (Matched Sample) . . 120
5.22 Single and Multiple Measure Effects on PS Award at 23 (Matched Sample) 121
5.23 Single and Multiple Measure Effects on Annual Wages at 25 (Matched Sample) . . 122
5.24 Teacher VA Effects on In-State Wages at Age 25 123

LIST OF FIGURES

Figure Page

3.1 TEAM Observation Rubric (General Educator) 31
3.2 Associations between K-12 Student Achievement and Long-Run Outcomes 32
3.3 Associations between K-12 Student Absences and Long-Run Outcomes 33
3.4 Associations between K-12 Student Suspensions and Long-Run Outcomes 34

4.1 Autocorrelations of Observation Scores 61
4.2 Changes in Student Characteristics and Teacher Observations 62

5.1 LYO and Zero-Imputed K-12 Effects 92
5.2 K-12 Observation Score Effects by Adjustment Method 93
5.3 LYO Effects on K-12 Outcomes, by Grade Level 94
5.4 K-12 Observation Score Effects by Domain 95
5.5 LYO and Back-Projected K-12 Effects 96
5.6 LYO Effects on Long-Run Outcomes, by Grade Level 97
5.7 Observation and VA Effects on K-12 Outcomes, Matched Sample 98
5.8 Observation and VA Effects on Long-Run Outcomes, Matched Sample 99

Chapter 1

Introduction

1.1 Objective

Over the past ten years, policymakers across the United States have substantially reformed the processes and systems used in their states and districts to evaluate public school K-12 teachers (Steinberg & Donaldson, 2016). Once consisting of little more than brief administrator "walk-throughs", modern teacher evaluation systems now use multiple modes for measuring teacher effectiveness, often linking teachers' performance on these measures to hiring, compensation, and tenure decisions (Goe, 2010; Grissom & Youngs, 2015; Martínez, Schweig, & Goldschmidt, 2016; Steinberg & Kraft, 2017). As intended by stakeholders supporting its adoption, the reformation of teacher evaluation systems was envisioned as a critical lever in preparing students for success not only within the K-12 classroom, but also, in their future colleges and careers (C. Jackson & Cowan, 2018). Thus, the efficacy of this new wave of teacher evaluation is intimately tied to whether its adoption can be used to promote student flourishing across a range of outcomes.

Recent research provides promising evidence that teachers can affect student outcomes of interest and that these effects are, in part, captured by commonly-used teacher evaluation measures (Chetty, Friedman, & Rockoff, 2014a; Chetty, Friedman, & Rockoff, 2014b; C. Jackson, 2018; Gershenson, 2016; Kane, McCaffrey, Miller, & Staiger, 2013). In particular, considerable research has been conducted on the properties of teacher value-added (VA) estimates, measures of teacher effectiveness formed using student test score data (Koedel, Mihaly, & Rockoff, 2015). Prior work finds that teachers' VA scores, on average, adequately capture their contribution to the test score performance of their students (Rivkin, Hanushek, & Kain, 2005) and, to a lesser extent, a range of non-test score outcomes of interest, such as attendance (Gershenson, 2016), discipline (C. Jackson, 2018), college attendance, and early career wages (Chetty et al., 2014b).

However, amidst the attention surrounding the use of value-added models for teacher evalu-

ation, administrator-led classroom observations remain the most commonly-used component in K-12 teacher evaluation systems (Kraft & Gilmour, 2017; Steinberg & Kraft, 2017; Whitehurst, Chingos, & Lindquist, 2014). Often positioned as the "subjective" measure accompanying "objective" teacher VA estimates, less is known about the relationship between teachers' classroom observation scores and the outcomes of the students they teach (Hansen & Goldhaber, 2015), particularly of observation scores captured within the context of an at-scale evaluation system. Given the lack of standardized assessments in all grades and subjects, evidence on the validity of classroom observation scores is critically important for ensuring that the evaluation of both "tested" and "non-tested" teachers is well-aligned with the goal of promoting student success. Additionally, as policymakers increasingly call attention to the importance of teachers and schools in promoting outcomes beyond test scores (Brighouse, Ladd, Loeb, & Swift, 2016; Ladd, 2017), classroom observations can potentially capture information about teacher effectiveness unnoticed by test-based measures such as teacher VA. If skills important to students' well-being are not also reflected in their tested achievement, an exclusive focus on value-added may miss several important ways in which teachers promote student success. Classroom observations, by focusing on multiple teacher practices rather than a single student outcome, may offer a wider view of the multiple dimensions of teacher quality.

1.2 Research Questions

In my dissertation, I study whether teacher observation scores collected as part of an at-scale teacher evaluation system validly capture teachers' impacts on multiple student outcomes. Specifically, I address the following research questions:

1. To what extent do at-scale classroom observation scores capture K-12 teachers' impacts on student:
    - K-12 outcomes (e.g., test scores, attendance, discipline, high school graduation)
    - Post-secondary attendance and completion

- Early career wages

2. How do the effects of changes to observation scores compare to the effects of changes to teacher value-added on these same outcomes?

I examine these questions using administrative education and labor data from the state of Tennessee that spans the 2006-07 to 2017-18 academic years. The core of these data are teachers' classroom observation scores, collected as part of Tennessee's statewide teacher evaluation system that was implemented during the 2011-12 academic year. These observation data are matched to student K-12, post-secondary, and labor market records, allowing me to connect teachers' measured effectiveness to the outcomes of their students from childhood to early adulthood.

Identifying the effects of differences in students' exposure to teacher quality, as measured by observation scores, poses several challenges due to both patterns of non-random assignment between students and teachers and the presence of measurement error in the observation scores themselves. Several promising strategies for addressing these threats to identification have been used in prior studies assessing the validity of teacher value-added estimates (Chetty et al., 2014a; Chetty et al., 2014b; Kane & Staiger, 2008). However, differences in the measurement processes of value-added estimates and observation scores, specifically the use of human raters in the latter, necessitate further modification of these strategies to address my research questions.

In my dissertation, I use a quasi-experimental "teacher switching" design that leverages changes in teacher quality at the department level due to teacher mobility to identify the effects of changes in observation score-measured teacher quality on student outcomes (Chetty et al., 2014a; Chetty et al., 2014b). The robustness of these estimates is tested using multiple forms of observation score "adjustment" designed to remove or mitigate different types of measurement error. Using the "teaching switching" design across a host of observation measures, I find considerable evidence that the information captured by teachers' classroom observations validly captures portions of their impacts on a range of K-12, post-secondary, and early career labor market outcomes. Specifically, among K-12 outcomes, I estimate that a standard deviation increase in teacher quality, as measured by observation scores, would result in an increase of .09 standard deviations of student test scores,

.15 fewer absences, and .02 fewer student suspensions in a given year. Observation score effects remain broadly statistically significant under different measurement adjustments and modifications to the identification strategy, though the extent to which these adjustments affect the estimated parameters differs by student outcome. Additionally, I find that changes in observation scores also capture teacher effects on longer-run outcomes. Increases to observation scores are estimated to result in a .04, 1.6, and 1.7 percentage point increase in the probability that student graduates high school in four years, attends a post-secondary institution by age 19, and earns a post-secondary award or degree by age 23, respectively. I find positive, significant effects on students' wages at age 24 onward, though point estimates are imprecise and vary widely across specification. Relative to the strength of teacher VA "effects" on the same student outcomes, the estimated observation score effects are comparable, if not slightly larger, across all outcomes.

This dissertation uses a "traditional" five chapter structure. Chapters 2 and 3 provide motivation and context for my analysis, situating this dissertation within both the extant literature (Chapter 2) and the specific Tennessee context of teacher evaluation reform (Chapter 3). In Chapter 4, I describe a theoretical model connecting teacher quality and student outcomes, explicate the parameters of interest in the dissertation, and detail the measurement and identification strategies I use to obtain clean estimates of these parameters. I provide results for my primary research questions in Chapter 5, in addition to findings from supplemental robustness analyses. Lastly, Chapter 6 connects findings from this dissertation to recommendations for current teacher evaluation practice and outlines future topics of study.

1.3 Contributions

My dissertation contributes to an already substantial literature on teacher quality and teacher evaluation in several ways. First, my dissertation will be among the first analyses to rigorously assess the validity of classroom observations obtained from an at-scale evaluation system. Previous validations of classroom observations (e.g., Kane et al., 2013; Ho & Kane, 2013; Garrett & Steinberg, 2015; Hamre, Pianta, Mashburn, & Downer, 2007; Blazar & Kraft, 2017) have primarily

used data collected from pilot or low-stakes settings, such as the Measures of Effective Teaching (MET) Projec). While the experimental conditions from these studies confer several benefits to the researcher, the context in which observation scores are captured may not reflect conditions seen within at-scale settings such as Tennessee. By using observation scores captured in midst of an active teacher evaluation system, results in this dissertation arguably generalize to other evaluation systems more readily than those obtained from prior experimental studies.

Second, the relatively long existence of the TEAM evaluation system, coupled with access to rich administrative data provided by both the Tennessee Education Research Alliance (TERA) and the MeasureTN P-20 data system, allow me to conduct novel analyses that connect observation scores to students' post-secondary and labor market outcomes. As education policy continues to shift attention onto supporting student progress throughout the P-20 pipeline, results from this dissertation represent a valuable contribution to the empirical evidence on how current teacher evaluation practice can support policymakers' ambitious and far-reaching goals for student success both in and beyond the K-12 classroom.

Lastly, this dissertation will provide direct comparisons of the predictive power of observation scores and teacher VA estimates on a range of student outcomes, with analyses involving only teacher VA serving as replications of prior validation work conducted by Bacher-Hicks, Kane, and Staiger (2014), Chetty et al. (2014a), Chetty et al. (2014b), and Rothstein (2017), among others. The ability to compare the two most common modes of teacher evaluation within the same sample should provide valuable evidence as states and districts consider how to structure their multiple measure teacher evaluation systems during the ESSA era.

Chapter 2

Literature Review

2.1 Teacher Evaluation in the United States

While adoption of high-stakes teacher evaluation systems at the district and state levels did not occur *en masse* until the late 2000s, efforts to measure teacher effectiveness have occurred in American schools since the early 20th century (Callahan, 1962; Tyack & Cuban, 1995). In the early 1900s, school administrators regularly rated staff on a wide range of dimensions, producing "efficiency records" that could be used to justify public expenditure on their school (Callahan, 1962). Items from these early efficiency records range from those that would be familiar to contemporary observation rubrics, such as "Academic preparation" and "Choice in questioning", to others likely too abstract or inappropriate in modern contexts, such as "General appearance" and "Moral influence" (Boyce, 1912; Ruediger & Strayer, 1910). Teacher evaluation practice generally evolved throughout the 20th century to explicitly include more formative processes that allowed increased interaction between teachers and their supervisors. Most notably, the development of the "clinical supervision" (Cogan, 1973; Goldhammer, 1969) and "differentiated supervision" models (McGreal, 1983) introduced evaluation practices, such as the "pre-conference, observation, post-conference" structure, that are commonplace in modern teacher evaluation systems.

Despite the presence of some similar practices, the scale of teacher evaluation in the current era far exceeds that seen in systems prior to the 21st century. While most school districts had some form of teacher evaluation in place, local implementation varied wildly (Wise, Darling-Hammond, McLaughlin, & Bernstein, 1984). A 1984 RAND survey of evaluation practices in 32 school districts found that the surveyed districts had broad agreement with regard to evaluated dimensions (e.g., teaching procedures, classroom management) and the use of pre- and post-evaluation conferences but differed over training provided to evaluators, the number of evaluations, and the extent to which evaluations informed subsequent professional development activity (Wise et al., 1984).

Differences in implementation are unsurprising given a lack of national standards and limited state standards for evaluation during this era. A 2002 review of state statutes determined that while 42 of 50 states included some language referencing teacher evaluation, only 7 states explicitly required that evaluations be held (Veir & Dagley, 2002).

The increased standardization of teacher evaluation practice in the United States seen over the 2010s is often attributed to several Obama Administration policies that encouraged or outright required states and districts to implement systems using multiple measures of teacher performance and to tie performance on these measures to meaningful consequences. The $5 billion Race to the Top (RttT) program was appropriated through the 2009 American Recovery and Reinvestment Act and competitively awarded grants to states promising to implement USDOE preferred policies in six core policy areas, including teacher evaluation (Hallgren, James-Burdumy, & Perez-Johnson, 2014). Specifically, the RttT criteria in teacher evaluation specifies that states implement a multiple measure teacher evaluation system that includes a student growth component and annually provides differentiated ratings. In total, 19 states won RttT awards, with individual grants ranging from $17 to $700 million. Relatedly, the USDOE Teacher Incentive Fund (TIF) grant competition rewarded states for developing multiple measure evaluation systems that informed teacher compensation, providing an additional funding-centered incentive for states to adopt more rigorous evaluation systems. A second Obama Administration policy that encouraged states to adopt multiple measure systems was the No Child Left Behind (NCLB) waiver process. Waivers of NCLB's "100% proficiency" provision were offered to states in exchange for adopting multiple measure teacher and principal evaluation systems in addition to other administration-preferred policy reforms (Polikoff, McEachin, Wrabel, & Duque, 2014). A minority of states promised the adoption of a statewide teacher evaluation system, with the majority opting to grant districts local control over the design of their systems (Pennington, 2014).

While states have varied with regard to implementing promised changes (Howell, 2015), the federal initiatives that induced them have clearly changed the taste and capacity for adopting outcomes-based accountability previously applied only at the school level to the evaluation of

individual teachers (Dee, Jacob, & Schwartz, 2013; Hamilton, Stecher, & Yuan, 2008; Minnici & Hill, 2007). From 2010-11 to 2016-17, 46 states enacted reforms of their teacher evaluation systems, with roughly 80 percent of these systems requiring that measures of student test score performance be incorporated alongisde classroom observation measures (Steinberg & Donaldson, 2016; Kraft & Gilmour, 2017). In addition to more robust evaluation criteria, the consequences attached to teacher evaluation have also increased. An NCTQ review of state policies in 2017 found that teachers' evaluation scores were used as determinants for assigning professional improvement plans (35 states), compensation levels (19 states), dismissal (23 states), and licensure (8 states).

2.2 Validating Teacher Quality Measures

Particularly given the increased stakes attached to evaluation scores, a first order concern for any measure of teacher quality is the extent to which the measure validly captures its intended constructs. Simple in the abstract, the notion of validating a teacher quality measure is difficult in practice due to the complexity of defining what it means to be an effective teacher (Borko, 2004). For this purpose of this dissertation, I define "teacher quality" in a narrow, but important, manner: the extent to which teachers affect student outcomes of interest. There are important questions regarding validity that are addressed only indirectly, if at all, within this frame. Do classroom observation scores truly capture the constructs described in their rubrics? Do the different domains of an observation rubric genuinely represent separate constructs? Ultimately, what does it mean for a teacher to be "effective"? These answers are important particularly if observation scores are used for formative purposes. Literature on the development of specific measures (e.g., Danielson, 1996; Hamre et al., 2007; Grossman, Loeb, Cohen, & Wyckoff, 2013), mixed methods studies connecting measured scores to specific classroom behavior (e.g., Hill et al., 2008; Hill, Charalambous, Blazar, et al., 2012, and factor analytic studies can speak more readily to these issues. However, given the emphasis placed on student success, broadly defined, as motivation for the adoption of teacher evaluation systems, focusing on the linkages between evaluation measures and student outcomes is a policy-relevant, albeit narrow, exploration of the validity of observation scores and other teacher

evaluation measures. Below, I review existing empirical literature examining whether these linkages exist for teacher value-added estimates and observation scores, the two evaluation measures examined in my dissertation.

2.2.1 Value-Added Effects on Test Scores

Research on teacher value-added (VA) effects on student achievement preceded their adoption into teacher evaluation systems by roughly 30 years, beginning in the 1970s with pioneering work from Hanushek (1970; 1971), Levin (1970), and Murnane (1975). Early work on this topic used limited non-experimental data, finding that student achievement varied significantly depending on the teachers to whom they were assigned and that simple versions of teacher effects on achievement were estimable; Green (2014) credits Hanushek (1981) as the first use of the term "value-added" to describe teacher effects on student achievement. Building from this initial work, several other studies continued to find evidence of non-experimental teacher effects in mostly K-8 settings (Hanushek, 1992; Rockoff, 2004; Rowan, Correnti, & Miller, 2002; Sanders & Rivers, 1996).

While this early work in value-added demonstrated that teacher assignment was a meaningful predictor of student achievement, the non-experimental nature of these studies made it unclear the extent to which the estimated teacher "effects" reflected genuine impacts on student learning. Patterns of non-random sorting between students, schools, and teachers were empirically documented within education research as early as the 1966 Coleman Report. Unaccounted for differences in outcomes due to the types of students a teacher is typically assigned could be misattributed to estimates of teacher effectiveness. Findings from these non-experimental studies were later corroborated by Nye, Konstantopoulos, and Hedges (2004), who use data from the Tennessee STAR classroom size experiment to identify teacher effects on student achievement under experimental settings, finding that these effects had comparable variance to teacher effects calculated in prior non-experimental studies.[1] Results from the Nye study serve as a "proof of concept" for teacher

[1] Chetty, Friedman, Hilger, et al. (2011) would later use the the Tennessee STAR data to identify experimental

value-added estimates: under randomized conditions, students' test scores varied meaningfully depending on the teacher to which they were assigned. However, additional evidence was needed to support the claim that VA calculated under non-experimental conditions provided reasonable approximations of the VA estimates that teachers would have obtained under random assignment.

Kane and Staiger (2008) and Kane et al. (2013) use experimental designs to provide evidence toward this end. In Kane and Staiger (2008), the first validation of teacher VA of this type, the authors use randomization from an impact study of NBPTS teacher certification in Los Angeles Unified School District to see whether the differences in teachers' VA estimates, calculated under "business-as-usual" settings, accurately predicted differences in the test scores of students to whom they had been randomly assigned. Specifically, among their experimental sample of roughly 150 classrooms, the authors find, albeit with substantial standard errors, that differences in teachers' VA estimates calculated during non-experimental years predicted differences in of student achievement of the same magnitude the following year when rosters where randomly assigned, i.e., teacher VA and student achievement shared a "one-to-one" relationship on average. The later Kane et al. (2013), part of the Measures of Effective Teaching (MET) Project, expands this line of inquiry to a broader sample of 3000 teachers in six urban districts. Importantly, the scope of the MET Project extends past value-added scores, and also considers the validity of classroom observation scores and student survey responses, in addition to composite measures of teacher effectiveness. Results from the study focusing on teacher VA largely corroborate with Kane and Staiger (2008), finding that these measures provide unbiased predictions of student test score achievement. Recent experimental work in this vein includes a recent study by Bacher-Hicks, Chin, Kane, and Staiger (2017) that uses a sample of 66 teachers who were randomly assigned classrooms to validate teacher VA estimates, classroom observation scores (MQI and CLASS), and Tripod survey responses. Using a design similar to Kane et al. (2013), the authors find that VA estimates, using either a state and alternative assessment, provided unbiased predictions of how students of those teachers would perform under random assignment.

"classroom" effects on test scores and other long-run outcomes

Absent randomization, a number of studies rely on quasi-experimental designs to account for the complex, idiosyncratic nature of student-teacher assignment (Clotfelter, Ladd, & Vigdor, 2006; Dieterle, Guarino, Reckase, & Wooldridge, 2015; Rothstein, 2009). Specifically, the most promising quasi-experimental designs sweep away the concerns with within-school sorting by aggregating data to the department (i.e. school-grade-subject) level, using changes in department-average teacher quality as a source of identifying variation for estimating the impact of changes to teacher quality (Chetty et al., 2014a; Gershenson, Hart, Lindsay, & Papageorge, 2017; Rivkin et al., 2005). In their pioneering work on teacher value-added, Chetty, Friedman, and Rockoff (2014a, 2014b) explicate arguably the most widely-known version of this design, which they refer to as a "teacher switching" quasi-experiment. Since, the teacher switching design has been used to validate Grades 4-8 teacher value-added estimates in Los Angeles (Bacher-Hicks et al., 2014) and North Carolina (Rothstein, 2017), in addition to high school teacher VA estimates in North Carolina (C. K. Jackson, 2014; Mansfield, 2015).

2.2.2 Value-Added Effects on Non-Test Score Outcomes

Recently, there has been interest in whether value-added models can identify teacher effects on student outcomes beyond test scores. Studies of this type fall into one of two related, but distinct, categories. First, a number of studies examine whether teacher's estimated value-added effects on test scores is predictive of their impact on other outcomes. VA estimates, while designed to specifically capture teacher impact on student test scores, may also indirectly capture portions of teacher effects on other outcomes to the extent that the student skills needed for academic achievement also generalize to attainment on other outcomes. Chetty et al. (2014b) demonstrate the teacher VA estimates are predictive of students high school graduation, college attendance, and early career wages. Similarly, C. Jackson (2018) finds that high school teacher's test score value-added is predictive of college graduation and students' intent to attend college.

Secondly, other studies use value-added style models to estimate "direct" teacher effects on non-test score outcomes (Blazar & Kraft, 2017; Gershenson, 2016; C. Jackson, 2018). Non-test

score teacher effects are estimated similarly to test score teacher effects and rely on the use of lagged outcomes as covariates (Chetty et al., 2014b; Gershenson, 2016). As result, non-test score teacher effects are typically estimated only for outcomes where there are annually repeated measures, such as attendance, discipline, grades, and on-track grade progression. A notable exception is Koedel (2008), who estimates high school teacher effects on student dropout using an instrumental variables strategy similar to the teacher switching design described previously.

A number of papers exploring teachers' non-test score effects use survey reports of students' attitudes and behaviors as an outcome variable. Using different survey instruments, Blazar and Kraft (2017), Jennings and DiPrete (2010), and Ruzek, Domina, Conley, Duncan, and Karabenick (2015) all estimate teacher effects on student attitudes/behaviors and test score outcomes, generally finding low correlations between teachers' abilities to improve test scores and attitudes/behaviors. Other studies estimate teacher effects on non-test outcomes obtained through administrative data, such as absences, disciplinary outcomes, and GPA. C. Jackson (2018) estimates teacher effects on student cognitive (i.e. test scores) and non-cognitive skill on a sample of high school teachers in North Carolina. Non-cognitive skills are proxied using a factor constructed from student absences, GPA, suspensions, and on-time grade progression. Gershenson (2016) conducts a similar analysis and compares teachers' test score and absence value-added effects estimated in both North Carolina administrative data and the ECLS-K. Gershenson's results largely corroborate C. Jackson (2018), finding that the variance of teachers' effects on absence, a non-test score outcome, is comparable to teachers' effect on test scores and that teachers' effects on these two outcomes are only weakly correlated.

Several key findings emerge from these studies. In general, the estimated teacher effects on non-test score outcomes have variances comparable in size to teachers' test score effects. Second, consistent across all studies, teachers' effects on test scores and non-test score outcomes are generally very weakly correlated, suggesting that teachers can have distinct abilities in affecting different student outomes. C. Jackson (2018) demonstrates the implications of this by showing that regressions using both test score and non-cognitive VA explain substantially more of a series

of long-run outcomes (i.e. high school graduation, intent to attend college) than regressions using test score VA alone. Interpreted in conjunction with Chetty et al. (2014b), these results suggest that while teacher test score value-added may be predictive of non-test score outcomes, these measures may strongly underestimate teachers' total effects on these outcomes.

2.2.3 Classroom Observation Scores

While both observation scores and teacher value-added (VA) are interpreted as measures of teacher quality, what, specifically, is meant by teacher quality is much less clearly defined for the former than it is for the latter. The notion of teacher quality carried by teacher VA is specifically defined as teachers' impact on student test achievement. This narrow definition readily lends itself to clear criterion measures (student test scores) against which to test the validity of teacher VA. By contrast, observation score rubrics typically encompass a much broader definition of teacher quality, often covering domains as distinct as the quality of teachers' classroom instruction to the degree to which they act professionally in the work place. When observation scores are averaged to form "overall" observation scores, as is often done in evaluation practice, it becomes less clear what construct(s) this composite represents, and subsequently, which measures might serve as appropriate criterion measures in a validation study.

Existing validations of classroom observation measures have typically selected student test scores as a criterion of interest, with prior studies typically consisting of (1) estimating correlations between observation and VA scores in non-experimental settings or (2) experimental designs used in the initial development of the measure. The Charlotte Danielson Framework for Teaching (FFT), likely the most widely used teacher observation rubric in the United States, is the subject of much of this research. Milanowski (2011) calculates descriptive correlations between teachers' FFT and value-added scores in three districts, finding correlations the range of .05-.48 across districts, subjects and years; direct correlations between teachers' prior year observation scores and the achievement of their students similarly ranged from .04 to .21. Kane, Taylor, Tyler, and Wooten (2011) examine the relationship between FFT scores and student achievement using administrative

data from Cincinnati Public Schools, finding that a one point increase in teachers' average observation scores (on a four point scale) is associated with a .171 and .212 standard deviation increase in math and reading achievement, respectively.

Similar analyses have been conducted using alternate observation rubrics, such as the Mathematical Quality of Instruction (MQI) (Hill, Charalambous, Blazar, et al., 2012), PLATO (Grossman et al., 2013; Grossman, Cohen, Ronfeldt, & Brown, 2014) and CLASS (Hamre et al., 2007) consistently find correlations within the range of .1-.4 (Chin & Goldhaber, 2016); Schacter and Thum (2004) stand out among these studies with correlations between .55 and .70 between observation scores and teacher value-added. The TAP Skills, Knowledge, and Responsibilities (SKR) observation rubric, a Danielson Framework "cousin" according to Milanowski (2011), was analyzed using a randomized sample in Schacter and Thum (2004), where the authors find substantial correlations (.55-.70) between a teacher quality factor constructed using SKR scores and teacher value-added across multiple subjects.

Like for other evaluation measures, the MET Project and subsequent papers using MET data have contributed substantially to research on classroom observation scores. Kane and Staiger (2012), an official MET Project study, examines the reliability and validity of observation scores specifically. Within the MET sample, four observation protocols are used: the Charlotte Danielson FFT, PLATO, CLASS, and the UTeach Teacher Observation Protocol (UTOP). Using disattenuation procedures, the authors estimate that observation scores for the various protocols are correlated between .12-.34 to student math gains but also significantly correlated to measures of student effort and "positive emotional attachment".

Recently, two papers, Garrett and Steinberg (2015) and Steinberg and Garrett (2016), use MET Project data investigate the relationships between student achievement, student characteristics, and Danielson FFT observation scores. Garrett and Steinberg (2015) find significant TOT relationships between observation scores and achievement, but ITT and IV estimates accounting for noncompliance to MET randomized rosters were not statistically significant. Steinberg and Garrett (2016) return to these data and find that teachers' observation scores are significantly related to

their students prior achievement. While this is a common finding in non-experimental studies (see Clotfelter et al., 2006; Kalogrides, Loeb, & Béteille, 2013), what is unique about the Steinberg and Garrett finding is this relationship persists even within the MET Project setting where students are randomly assigned to teachers, suggesting that students' incoming achievement either (1) modifies the quality of teachers' instructional practice or (2) biases observers' assessment of teachers' instructional practice.

Chapter 3

Background & Data

3.1 The TEAM Evaluation System

In January 2010, the Tennessee General Assembly passed the "First to the Top" (FTTT) Act (Senate Bill 5), mandating a series of comprehensive education policy reforms including a reformation of Tennessee's teacher evaluation system. The state's reform efforts were further supported with receipt of a $500 million grant through the U.S. Department of Education's "Race to the Top" grant competition. The FTTT Act required that teachers be evaluated using a composite "Level of Effectiveness" that was calculated using performance across multiple measures of effectiveness. As mandated by the FTTT Act, for teachers in tested grades and subjects, 50 percent of the LOE calculation is determined by "student growth" as measured by the Tennessee Value-Added Assessment System (35 percent) and a locally-selected "student achievement" measure (15 percent).[1] Evaluation policies beyond these weights were to be determined by recommendation from a 21 member Teacher Evaluation Advisory Committee (TEAC) and the State Board of Education (SBE).

Chief among the TEAC's responsibilities was selecting the rubrics and protocol to be used for the observation component of the evaluation system. Toward this end, the TEAC commissioned a year-long pilot of four evaluation models during the 2010-11 academic year: (1) the Tennessee Educator Acceleration Model (TEAM), (2) the Project COACH model, (3) the Teacher Effectiveness Measure (TEM) Model, and (4) the Teacher Instructional Growth for Effectiveness and Results (TIGER) Model. The four models differ primarily on the specific observation rubric used, the frequency of observation, and whether student surveys are included as a performance measure. Upon

[1]The SBE approved the use of the TRIPOD student perception survey as an evaluation measure in 2013. This measure is currently used as part of the Shelby County/TEM evaluation model. The state is continuing to pilot other student perception survey instruments such as My Student Survey and Panorama

TEAC's recommendation at the end of the 2010-11 pilot, the SBE adopted TEAM as the state's primary evaluation model but also approved the remaining three models as accepted alternatives for districts wishing to employ them.[2] Between 2011-12 to 2017-18, more than 80 percent of teachers receiving evaluation scores in Tennessee were evaluated under the state's preferred TEAM model, followed by roughly 10 percent evaluated under TEM, the evaluation model used by the state's largest district, Shelby County Schools.

The TEAM evaluation system (and its alternates) became state-wide policy beginning in the 2011-12 academic year. Formally, Tennessee teachers' performance within the evaluation system factors into their tenure status, the frequency of observation in the subsequent academic year, and, in select districts, compensation. I provide additional detail below on the four types of evaluation measures used in Tennessee.

3.1.1 Classroom Observations

Classroom observation scores are the most heavily-weighted performance measure in Tennessee, making up no less than 50 percent of a teacher's summative rating across years, evaluation models, and "tested vs. non-tested" status. Observation policy is broadly consistent across all evaluation models: trained observers, often the school's principal or assistant principal, conduct and score teachers' classroom practice using a pre-determined rubric and review teachers' strengths, weaknesses, and scores following the conclusion of the observation. The observation process across models differs primarily with regard to items included in the observation rubric and frequency/length of each observation.

The TEAM Observation Rubric. The TEAM observation rubric is the most commonly-used rubric in Tennessee. The rubric was developed in conjunction with the the National Institute for Excellence in Teaching (NIET) and is an adaptation of the organization's Skills, Knowledge, and Responsibilities (SKR) observation rubric. According to documentation provided by NIET, items for

[2] A fifth model, the Achievement Framework for Excellent Teaching (AFET), used by the Achievement School District, was approved for the 2012-13 school year.

the SKR were selected through review of existing state teaching standards in Massachusetts, California, and Connecticut, in addition to the Charlotte Danielson Framework For Teaching (NIET, 2010). Schacter and Thum (2004) validate an early version of the SKR rubric using random assignment, finding that a "teacher quality" factor estimated using 12 SKR items was strongly predictive of student achievement gains.

The TEAM observation rubric consists of 23 items nested within 4 domains: Instruction, Environment, Planning, and Professionalism (see Figure 3.1). A comparison of the TEAM rubric with the standard SKR rubric (NIET, 2010, p.17) reveals that the two rubrics are nearly identical, with the most significant differences occurring within the Professionalism domain. TEAM requires that each teacher is observed for each of 23 TEAM items 1-3 times by a trained observer every academic year. The specific number of required observations depends on a teacher's tenure status and prior-year evaluation scores; all other factors held equal, pre-tenure teachers and teachers with lower evaluation scores are required to receive more observations than their peers. At least half of the observations are required to be unannounced (TDOE, 2018).

Following the clinical supervision model, teachers and their observers engage in a pre-conference prior to announced observations. Pre-conferences are used to discuss an intended lesson plan for the observation and to discuss relevant contextual information about the teacher's classroom. After all observations, teachers and observers participate in a post-conference where observers highlight areas of "reinforcement" (i.e. strengths) and "refinement" (i.e. weaknesses) and share scores for each of the TEAM items evaluated during the observations. Teachers receive an integer 1-5 score for each item evaluated, with scores of 1 and 5 indicating that teacher performance was "Significantly Below Expectations" and "Significantly Above Expectations", respectively. Observations can occur at any point during the academic year for all domains besides Professionalism, which is to be scored following the conclusion of state testing. Districts are expected to submit complete observation score records to TDOE by the beginning of the July following the academic year.

Alternative Observation Rubrics. The Shelby County TEM observation rubric is more extensive than TEAM, consisting of 7 domains and 41 items. TEM provides specific items for its

Professionalism Domain but, unlike the TEAM model, teachers' Professionalism scores receive their own weight for the LOE calculation and are not considered a part of a teacher's observation score. Shelby teachers receive between 2-4 observations, with the specific number depending on their prior year evaluation scores and whether they are new to Shelby County schools. The Project COACH districts use the most structurally distinctive observation model where most teachers undergo 8 "mini-observations" throughout the course of the year, with at least one mini-observation conducted by two co-observers. The COACH system is designed to give teachers quicker and more frequent feedback throughout the year, relative to the other observation models. The COACH rubric itself consists of 6 domains and 42 items and is based on the Kim Marshall Model for Teaching. Among the alternate models, TIGER uses an observation rubric most similar to TEAM. This is due to the fact that both rubrics draw from the Charlotte Danielson Framework For Teaching as a common source; TIGER is a first-degree adaptation whereas TEAM is modeled from SKR, which itself is modeled after the Danielson Model.

3.1.2 Student Growth

Teachers are scored on the student growth component of the evaluation system using (1) individual-level TVAAS estimates, (2) growth on a submitted portfolio of student work, or (3) school-level TVAAS estimates.

The TVAAS Model. Unlike other states implementing teacher evaluation reforms in the post-NLCB era, Tennessee has a long tradition of using value-added estimates for accountability purposes. The TVAAS model was developed by William Sanders (Sanders & Horn, 1994) and was formally incorporated into Tennessee's pre-NCLB education accountability system through the 1992 Tennessee Education Improvement Act (EIA) (Morgan, 2004). Under the EIA, schools and districts under-performing with regard to TVAAS or other designated performance measures were subject to a probationary periods and possible forced removal of school/district leadership. In addition to these formal penalties, school and district-level TVAAS estimates were publicly posted as a form of public accountability and reports of individual teacher TVAAS were distributed among

practitioners (Sanders & Horn, 1998). The use of TVAAS as a formal accountability measure ceased in 2003-04 due to NCLB-imposed restrictions on the use of norm-referenced measures for accountability, though the state continued to use TVAAS estimates for diagnostic purposes (Morgan, 2004).

Since its development in Tennessee, the TVAAS model is now the state value-added model for North Carolina, Ohio, Pennsylvania, and South Carolina, in addition to being used for district-level value-added in several other states. Despite its popularity, the TVAAS is among the more complicated value-added models used in education accountability (McCaffrey, Lockwood, Koretz, & Hamilton, 2004). Full explication is beyond the scope of this dissertation proposal though I provide an overview of the model.[3] TVAAS offers two separate value-added models, the Multivariate Response Model (MRM) which is used for Grade 4-8 TVAAS estimates and the Univariate Response Model (URM) which is used for EOC and Early Grades TVAAS estimates. The MRM is a linear mixed model that jointly estimates teacher effects on student achievement across grades, years, and subjects within a five year window. Teacher effects are modeled as random and allowed to vary as time passes; a Grade 3 teacher's effect on her student's Grade 3 achievement is allowed to vary from her effect on those students' Grade 4 achievement. Initially, teachers' TVAAS estimate were annually "re-estimated" as new years of data fell into the model's five year estimation window though this practice was discontinued prior to the 2014-15 academic year. The MRM does not control for covariates beyond fixed dummy indicators for subject-year-grade combinations, relying on information about students' prior testing and teacher assignment history to account for non-random assignment.

For tests where (1) students are not consecutively tested in each grade or (2) prior and current year scores are obtained from different tests, teacher value-added is estimated using the URM. The URM fits separate models for each subject-year-grade and controls for as many past test scores as a student has available. These prior test scores and their estimated coefficients are used to project a student's predicted score in that subject-year-grade, which is subsequently used as a covariate

[3] Ballou, Sanders, and Wright (2004) provides a quality overview of the MRM model

in a teacher random effects model to obtain estimates of teachers' value-added. In addition to teacher-level estimates, TVAAS also estimates district- and school-level TVAAS. These measures can be used by teachers in non-tested grades and subjects as their growth score or can be selected to represent the "Achievement" measure in the LOE calculation. District and school effects are estimated using variants of the URM and MRM models, but unlike teacher effects, are modeled as fixed rather than random.

For accountability purposes, subject-grade-year specific TVAAS estimates are combined into "evaluation composites" by first transforming TVAAS estimates into indexes by dividing estimates by their standard error. Various evaluation composites are calculated by taking student enrollment-weighted averages of the subject-grade-year specific indexes. Typically, a up-to-three year composite index of all a teacher's available TVAAS estimates in that time span is used to calculate her LOE.

Student Growth Portfolios. In effort to increase the number of teachers eligible for individual growth scores, the TEAM system also calculates student growth using submitted portfolios of student work in selected "non-tested" grades and subjects. A 2017 TDOE report indicates that the student growth portfolio option was used by slightly over 2000 teachers during the 2015-16 academic year, with nearly 75 percent of these teachers teaching the fine arts and remaining teachers teaching world languages, physical education, Pre-K/K, or first grade (Stone, 2017). To receive a student growth portfolio score, teachers must submit work from 3-6 students at two points in time within the academic year. Teachers are to select work from academically diverse students, aiming to select students below, at, and above proficiency in tested subjects. Student growth portfolios are rated by trained peer reviewers and assigned a 1-5 score that can be used as that teacher's "student growth" score for LOE calculation. Stone (2017) finds that teachers' student growth portfolio are aligned with their observation scores, with roughly 80 percent of teachers receiving student growth portfolio and observations scores ± 1 level of each other.

3.1.3 Achievement

A teacher-selected "Achievement" measure comprises 15 percent of a teacher's LOE calculation. At the beginning of the school year, teachers and their evaluators are responsible for selecting among a list of TDOE-approved measures. The criteria for these measures are loose, with the official TEAM website noting that selected measures "showed a relationship to student growth" and "could be returned in a timely manner". Broadly, these achievement measures are typically school-level averages of TCAP/TNReady scores, TVAAS scores, ACT scores, or graduation rates. In 2014-15, the three most commonly selected achievement measure was school-wide composite TVAAS score (20.2 percent), followed by graduation rate (9.6 percent) and ACT scores (6.1 percent). Teachers and evaluators are responsible for defining the criteria for earning a 1-5 score for all achievement measures with exception of graduation rate, which has state-defined criteria.

3.1.4 Levels of Effectiveness

Teacher performance on the observation, growth, and achievement components are combined to form a 0-200 composite score. The weights on each component vary by year, evaluation model, and whether a teacher is in a tested subject/grade, though generally, observations comprise no less than 50 percent of a teacher's composite score. A 1-5 Level of Effectiveness (LOE) Rating is assigned based on teachers' cutscores. As is consistent with many other contemporary evaluation systems (see Kraft & Gilmour, 2017), despite the emphasis on more rigorous processes, the distribution of the summative "Level of Effective" measure in Tennessee has been highly left-skewed in all years of the evaluation system, with no fewer than 70 percent of teachers in any given year receiving either a Level 4 or 5 rating.

3.2 Data

To address my research questions, I match data on (1) teacher evaluation scores, (2) K-12 student teacher linkages, and (3) students' K-12, post-secondary, and labor market outcomes. I

obtain these data from two sources. First, the Tennessee Education Research Alliance (TERA) database is a repository for longitudinal "research-ready" K-12 student-, teacher-, and school-level administrative data. TERA-provided data files themselves are constructed from extracts drawn from the Tennessee Department of Education's Education Information System (EIS), Personnel Information Reporting System (PIRS), and TNCompass. Broadly, I use TERA data for all K-12 information including student-teacher assignments, student and teacher demographics, teacher evaluation scores, and K-12 student outcomes. Second, I use the MeasureTN P-20 database to access information on student's post-secondary and labor market outcomes.

3.2.1 Student-Teacher Rosters

I link students to teachers primarily using student-teacher linkage files used by the SAS Institute for estimating teacher TVAAS scores. These linkage files match students to teachers in a particular year-subject and also allow teachers to indicate "percentage claim" over a student's instructional time for that year-subject and a student's "instructional availability", a categorical variable based on a student's attendance rate. A limitation of the provided linkages are that they are available only for teachers in tested grades and subjects. Luckily, Tennessee has a relatively expansive set of "tested grades and subjects" which, in addition to Grades 3-8, includes End-of-Course (EOC) exams in most high school core academic subjects (e.g., English I-IV, Algebra I, Geometry, Biology) and a limited sample of K-2 linkages from schools that elected to administer a voluntary early grades assessment. [4]

Student-teacher rosters serve as the "base" files to which I merge on data on student demographics, student outcomes, and teacher evaluation to construct my analytic sample. A key sample restriction that I impose is that I keep only students who (1) were claimed by a single teacher for a given subject-by-year combination and (2) were labeled as having "full instructional availability". Limiting my analytic sample to students with unambiguous linkages to their teachers for a

[4]English IV rosters are not available through TVAAS linkage files and are constructed using administrative course records.

given subject helps to clarify the attribution of students' outcomes to their exposure to specific teachers, which is central to this analysis; future research may investigate how methods for accounting for co-teaching in value-added estimation, as described by Hock and Isenberg (2012), can be applied to address the issue of split claims. Students who do not meet these restrictions are systematically different from those that do, as the former are more likely to be highly mobile or participate in "intervention" or "pull-out" courses. After retaining only student-subject-year observations with meeting these restrictions, the final analytic data set consists of roughly eight million records. Descriptives for key student and teacher variables are provided in Table 3.1, with additional information on student and teacher data provided below.

3.2.2 Student and Teacher Demographics

I match standard student and teacher demographic information to teacher-student rosters. These data are available from 2005-06 to 2017-18. Student demographic information includes ethnicity, gender, FRPL, SPED, and ELL eligibility. Teacher demographic information includes ethnicity, gender, years of teaching experience, and salary. Roughly 30 percent of Tennessee students in my sample are non-white, with half of all students eligible for free and reduced-price lunch.

3.2.3 Teacher Quality Measures

I use two teacher quality measures in this analysis, (1) classroom observation scores and (2) "leave-year-out" VA estimates, as calculated in Chetty et al. (2014a). Overall classroom observation scores are available from 2011-12, the first year of statewide TEAM implementation, to 2017-18 for all observation models (e.g. TEAM, TEM, TIGER). The availability of item-level observation scores differs by year and model, with item-level scores available for all districts using the TEAM rubric from 2011-12 to 2017-18 and select alternate rubric districts during this period. My primary analyses use the overall observation scores (or measures constructed from overall observation scores) in order to include as large a proportion of districts in the analyses as possible. However, for subanalyses that require item-level scores, my analytic sample will be composed

only of teachers for whom item-level observation scores are available. This item-level sample is comprised mostly of teachers observed under the most popular TEAM rubric.

TVAAS estimates are the state's official VA measure. However, it useful for validation purposes to have control over how each teacher quality measure is estimated. Because the complexity of TVAAS makes re-estimation infeasible, I use teacher VA estimates obtained from "leave-year-out" models that have been validated as minimally biased predictors of student test score gains in several studies (Bacher-Hicks et al., 2014; Backes et al., 2018; Chetty et al., 2014b; Rothstein, 2017). I estimate teacher VA estimates in all four Grade 3-8 subjects (Math, Reading, Science, Social Studies) in addition to a select number of EOC subjects for which I have rosters and test scores (i.e., Algebra I-II, Biology, English I-III) using a specification that follows CFR's as closely as possible. Specifically, the model includes controls for student race/ethnicity, gender, FRPL, SPED, and ELL status, a cubic function of prior-year test scores in all subjects, and classroom (i.e. teacher-by-year), school-by-year, and school-by-grade averages of all student-level covariates. VA estimates are calculated separately by subject and school level (i.e. elementary, middle, high). Drift limits are set to six for all VA models.

3.2.4 Student Outcomes

Student Test Scores. K-12 Student test scores are broadly available from 2006-07 to 2017-18 in Grades 3-8 core subjects (e.g., Math, Reading, Science, Social Studies) and select subjects (Algebra I-II, Biology, English I-III, and History) with valid EOC exams.[5] Importantly, there a number of EOC courses for which rosters are available in a given year but not student test scores such as English IV, Math Foundations, and US History. Prior to 2015-16, Tennessee used the Tennessee Comprehensive Assessment Program (TCAP) as its state standardized assessment. In 2015-16, the state adopted the computer-administered TNReady assessment. Due to incomplete administration in the first year of assessment, standardized test scores in Grades 3-8 are not available in 2015-16; EOC scores in 2015-16 are unaffected. TNReady test scores in grades 3-8 are first available in

[5]Social Studies scores are only available from 2006-07 to 2013-14.

2016-17. Within the years of my sample, Tennessee students have typically achieved proficiency rates on the state assessments between 30-40 percent in math and reading, rates that closely match Tennessee student performance on the NAEP assessment during that period. In my analyses, I use test scores that have been standardized within year, grade, and subject.

Student Attendance. Student attendance records are available from 2005-06 to 2017-18 and identify the number of absences, excused and unexcused, experienced by a student in a given academic year. In any given year, students averaged 7-8 total (excused and unexcused) absences. In all analyses, I operationalize student attendance as the number of absences, both excused and unexcused, in a given year.

Student Discipline. I have access to student discipline records from 2005-06 to 2017-18. These records are at the student x incident level and provide information on the incident type, incident type (e.g., vandalism, bullying, theft), and discipline type (e.g., in-school suspension, out-of-school suspension, expulsion). While this level of detail makes more granular discipline measures possible, I use a student's total number of suspensions (in-school or out-of-school) as a relatively simple measure of student discipline. In my data, roughly 90 students, in a given year, never experience a suspension. Among students experiencing at least 1 suspension, the average number of suspensions incurred in a year was 2.4.[6]

Student High School Graduation. Student graduation records are available from 2005-06 to 2016-17 and identify the type (e.g., high school diploma, special education diploma, GED) and year of completion for each student. I use an indicator of "on-track" high school graduation as my main high school graduation outcome in the analysis. This indicator is defined as one for students who could be matched to a diploma record during their twelfth grade year and as zero for all students who are in cohorts in which high school graduation could potentially be observed but could not be matched to a diploma record. Students from recent cohorts for which high school graduation could not yet be observed are giving missing values for this indicator. The manner in which this indicator is defined conflates students who transfer to private schools or out-of-state

[6]For both attendance and suspension analyses, models using logged number of attendances and suspensions yielded qualitatively similar results. These outcomes are opertaionalized as counts in the manuscript for interpretability.

K-12 schools with those who have dropped out. As a result, the statewide on-time graduation rate I calculate in my sample (77 percent) is roughly 10 percentage points lower than rates reported by TDOE during the period of data.[7]

Student Post-Secondary Outcomes. Student post-secondary attendance and completion are broadly available from fall of the 2006-07 academic year to the fall of the 2018-19 academic year. Student post-secondary enrollment data is provided by MeasureTN, who obtain the data through partnership with the Tennessee Higher Education Commission and the National Student Clearinghouse. These data included student-term specific enrollment status, major, GPA, credit hours, and full/part-time status. Importantly, while data on initial enrollment is available for students in my sample who enroll in any type of institution, data on post-secondary completion is limited to students attending post-secondary institutions that report to the Tennessee Higher Education Commission (THEC), which consists of only public institutions within the state of Tennessee.[8] In my analysis, I measure both post-secondary enrollment (age 19) and completion (age 23) at specific ages, separating out enrollment/completion outcomes according to whether a student enrolled at or completed an (1) Associates Degree/Certificate program or (2) a Bachelor's degree program.

Student Labor Market Outcomes. Lastly, the data set includes individual quarterly wages from the third fiscal quarter of 2006 to the second fiscal quarter of 2018. I sum quarterly wages to the person-year level to calculate annual wages for each individual. Years are defined to best align with the academic year, where Quarters 1 (January-March) and 2 (April-June) in calendar year t are assigned to academic year t but Quarters 3 (July-September) and 4 (October-December) are assigned to academic year $t+1$. In my analysis, I only include wage outcomes for individuals who report 4 quarters of wages within a given year. Annual wages are inflation adjusted using the Consumer Price Index and expressed in 2018 constant dollars. Individual wage data are provided by the University of Tennessee-Knoxville and the Tennessee Department of Labor and Workforce Develop-

[7] I explored defining a graduation indicator that assigned zeroes only for students with explicit enrollment records indicating drop-out; students who were unaccounted for in either the graduation or drop-out data were defined as missing. This produced an estimated on-time graduation rate of 97 percent, exceeding the state reported values.

[8] The lack of data on post-secondary completion from out-of-state and private school attendees can potentially lead to sample selection bias if the likelihood of attending these types of institutions is correlated with observed teacher quality.

ment. These data are collected through Tennessee's Unemployment Insurance (UI) system and are inclusive of all earnings in Tennessee during this period of time that are covered by the UI system. Importantly, this excludes individuals who (1) attended a Tennessee K-12 school but have moved and/or earn wages outside the state of Tennessee and (2) are employed in certain sectors that are not covered by the UI system (e.g., select military and agricultural occupations, self-employment, select federal employment). As result, my calculated annual wages will by systematically lower than individuals' true annual wages in addition to being missing for out-of-state-individuals.

3.2.5 Relationships Between K-12 and Long-Run Student Outcomes

Prior to examining whether teachers impact multiple student outcomes, it may be useful to see how student attainment across multiple outcomes interrelate descriptively (Chetty, Friedman, & Rockoff, 2011). An implied, but critical, component of schools' increasing focus on preparing students for "college and career" is that the methods used to evaluate teachers, schools, and other educational interventions are well-aligned with students long-run outcomes. However, critics of test score-based evaluation claim that placing emphasis on tested achievement may possibly come at the detriment to the promotion of other skills that may be more important for long-term success (Ladd, 2017). For select cohorts of students, I have the ability to directly observe both K-12 and long-run outcomes. However, teachers and school leaders will not be able to peer in the future to determine whether a given teacher or intervention had an effect on students' distal outcomes. Therefore, it may be of interest to see whether student outcomes measured during their K-12 careers can serve as useful proxies for their attainment later in life.

I examine how the three annually-measured K-12 student outcomes, test scores (averaged across all available subjects), absences, and suspensions, measured at different grades, are correlated to four longer-run outcomes: (1) on-time high school graduation, (2) enrollment in any post-secondary institution at age 19, (3) attaining any post-secondary degree/credential at age 23 and (3) annual in-state wages at age 25. Because students' family and community resources are hugely influential on both K-12 and long-run outcomes, I estimate the relationship between K-12

and long-run outcomes by running school-cohort fixed effects models that regress a given long-run outcome on levels of a given K-12 outcome, controlling for student gender, race, and FRPL, ELL, and SPED designation. While strictly descriptive in nature, I opt to use coefficients estimated from these models, rather than raw correlations, in order to better isolate the relationship between student attainment between K-12 and long-run outcomes. These regressions are run separately by grade in order to examine how the strength of these associations changes as the points in time at which student' K-12 and long-run outcomes are measured become closer.

In Figures 3.2, 3.3, and 3.4, I present the estimated associations between students K-12 and long-run outcomes. The use of school-cohort fixed effects entails that all coefficients be interpreted as the relationship between a given K-12 and long-run outcome pair within a given cohort of students who passed through that school in that year.

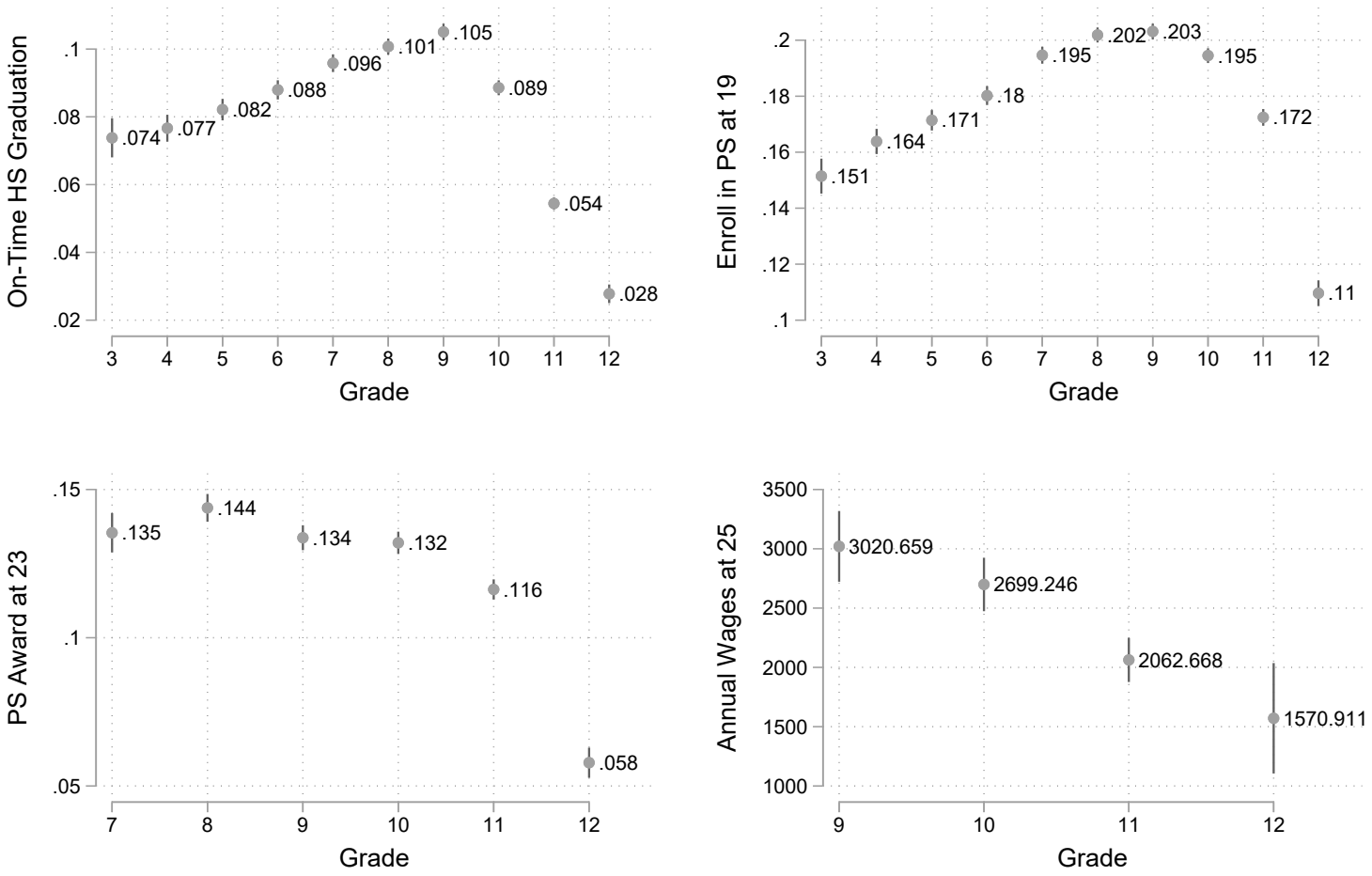
Notably, each K-12 outcome, at all grades, is significantly associated with each long-run student outcome, with the strength of this association varying across grades. Again noting that the estimates presented in the previous figures are strictly associations, the results suggest that student attainment on K-12 outcomes, even in the elementary grades, may foretell their attainment on longer-run outcomes. In grades 3-9, there is a largely negative monotonic relationship between grade and association, with K-12 outcomes in older grades (i.e., grades more proximate to the long-tun outcome of interest) showing stronger associations with long-run outcomes than K-12 outcomes in earlier grades. Exceptions to this pattern are student suspensions in Grades 3 and 4, which are shown to have larger negative associations with high school graduation and post-secondary enrollment than suspensions in later grades. Because suspensions are relatively rare in the early grades, their occurrence may be particularly telling of student propensity toward attaining later outcomes. While the associations between K-12 and long-run outcomes generally become larger in magnitude between Grades 3-9, these associations tend to become smaller as students progress through the high school grades, with particularly small associations in grade 12. The negative relationship between the magnitude of the estimated coefficients and HS grade is likely driven by attrition bias: students with non-missing values of K-12 outcomes in the high school

grades differ systematically from students with missing values (i.e., students who have dropped out), the former group being more prone to higher levels of attainment than the latter. Additionally, the relatively low but still significant associations in twelfth grade outcomes are likely due to "senioritis" type behavior where students' high school graduation and post-secondary attendance status for the following year is largely already determined by the spring of their senior year.

Figure 3.1: TEAM Observation Rubric (General Educator)

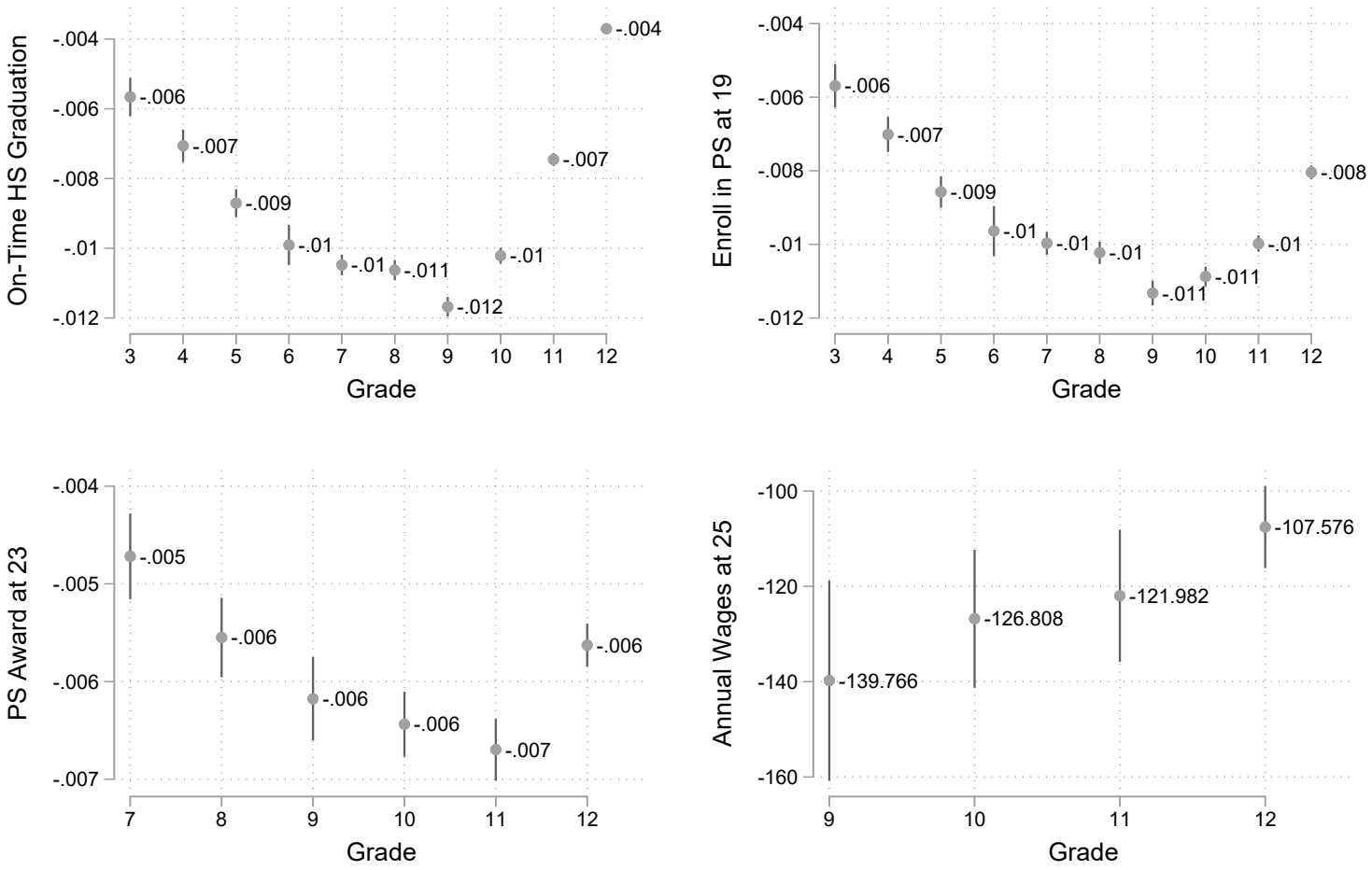
Instruction: Activities and Materials	Instruction: Questioning	Environment: Expectations	Professionalism: Leadership
Instruction: Grouping Students	Instruction: Standards and Objectives	Environment: Managing Student Behavior	Professionalism: Professional Growth
Instruction: Lesson Structure and Pacing	Instruction: Teacher Content and Knowledge	Environment: Respectful Culture	Professionalism: School and Community
Instruction: Motivating Students	Instruction: Teacher Knowledge of Students	Environment: Environment	Professionalism: Use of Data
Instruction: Presenting Instructional Content	Instruction: Thinking	Planning: Instructional Plans	Planning: Assessment
Instruction: Problem Solving	Instruction: Teacher Feedback		Planning: Student Work

Figure 3.2: Associations between K-12 Student Achievement and Long-Run Outcomes



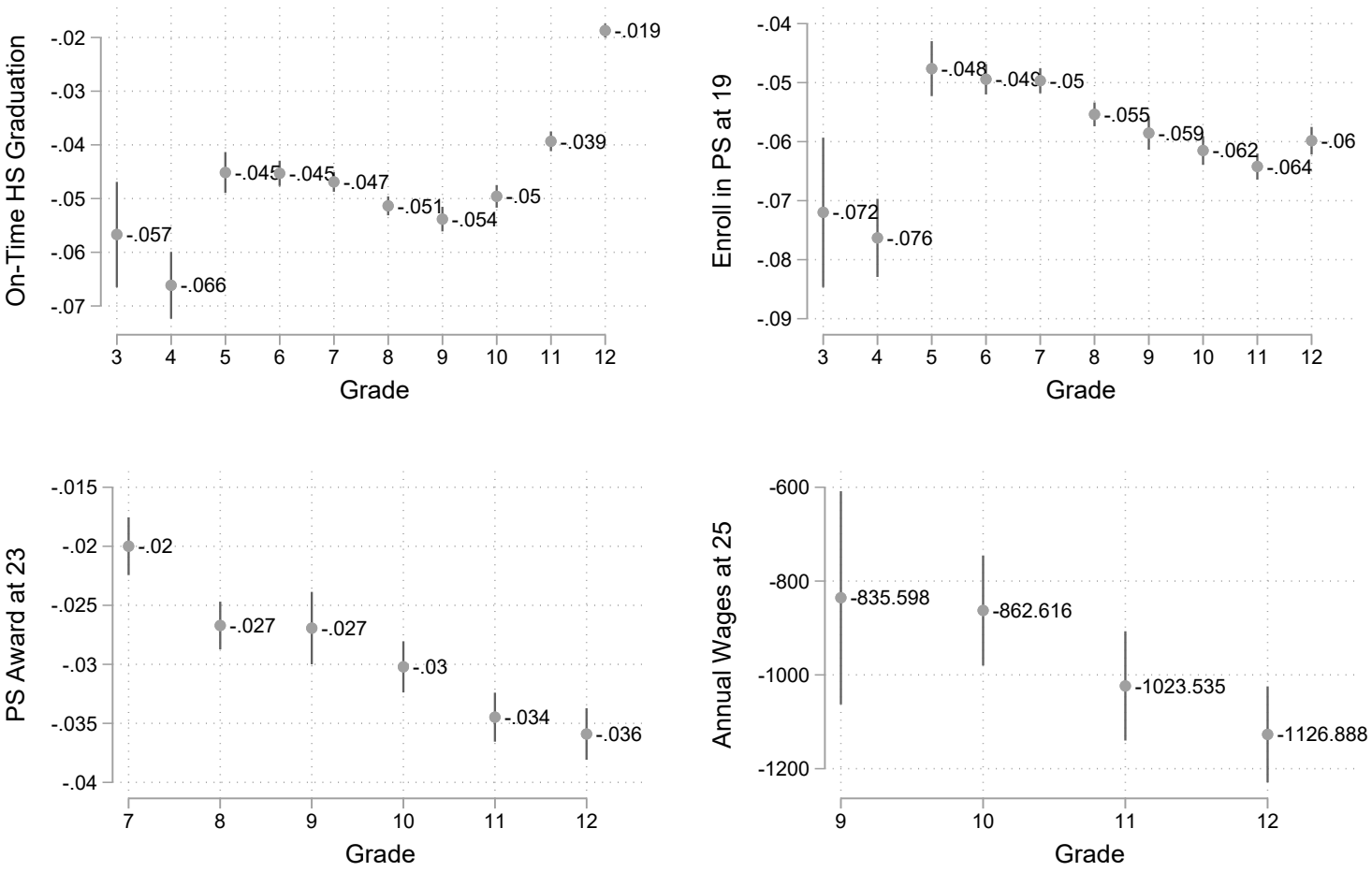
Note: Figure shows the estimated association between standardized student achievement and a given long-run student outcome, by student grade. Each point is the coefficient on student achievement from a separate by-grade school-cohort fixed effects models that regress a given long-run outcome on achievement and a vector of student controls. Graduation and enrollment outcomes are operationalized as binary indicators and wages are expressed in 2018 constant dollars. Vertical lines indicate 95 percent confidence intervals. Standard errors are clustered at the school-cohort level.

Figure 3.3: Associations between K-12 Student Absences and Long-Run Outcomes



Note: Figure shows the estimated association between the number of student absences and a given long-run student outcome, by student grade. Each point is the coefficient on student absences from a separate by-grade school-cohort fixed effects models that regress a given long-run outcome on absences and a vector of student controls. Graduation and enrollment outcomes are operationalized as binary indicators and wages are expressed in 2018 constant dollars. Vertical lines indicate 95 percent confidence intervals. Standard errors are clustered at the school-cohort level.

Figure 3.4: Associations between K-12 Student Suspensions and Long-Run Outcomes



Note: Figure shows the estimated association between the number of student suspensions and a given long-run student outcome, by student grade. Each point is the coefficient on student suspensions from a separate by-grade school-cohort fixed effects models that regress a given long-run outcome on suspensions and a vector of student controls. Graduation and enrollment outcomes are operationalized as binary indicators and wages are expressed in 2018 constant dollars. Vertical lines indicate 95 percent confidence intervals. Standard errors are clustered at the school-cohort level.

Table 3.1: Sample Descriptive Statistics

	Mean	SD
Student Demographics		
Prop. Student FRPL	0.499	0.500
Prop. Student SPED	0.098	0.298
Prop. Student Female	0.499	0.500
Prop. Student Black	0.227	0.419
Prop. Student Hispanic	0.071	0.257
Prop. Student ELL	0.083	0.277
Student Grade Level		
Elementary (Grades 3-5)	0.390	0.488
Middle (Grades 6-8)	0.359	0.480
High (Grades 9-12)	0.251	0.434
Student K-12 Outcomes		
State Test Score (Z)	0.074	0.953
Number of Absences (Excused & Unexcused)	7.914	8.321
Number of Suspensions	0.300	1.117
On-Time HS Graduation	0.770	0.421
Student PS Outcomes		
Enroll in Any PS by Age 19	0.547	0.498
Enroll in BA Granting Institution by Age 19	0.254	0.435
Enroll in Non-BA Granting Institution by Age 19	0.324	0.464
Earn any PS Award by Age 23	0.242	0.428
Earn Bachelors by Age 23	0.151	0.358
Earn Associates/Certificate by Age 23	0.104	0.304
Student Wage Outcomes		
Annual Wages in Age 23	22521.65	13290.75
Annual Wages in Age 24	26245.70	15035.06
Annual Wages in Age 25	29286.17	16477.67

Note: Table shows means and standard deviations for key student and teacher characteristics. Descriptive statistics are drawn from a sample of student-year-subject observations between 2011-12 to 2017-18 that were (1) singularly claimed by a single teacher for a given subject-year combination and (2) were labeled as having "full instructional availability" in the TVAAS linkage files.

Chapter 4

Methods

In this dissertation, I explore the extent to which information collected as part of the classroom observation process captures teachers' effects on student outcomes of interest. Estimating these relationships using administrative data is complicated due to patterns of non-random assignment between students and teacher quality and measurement error within classroom observation scores. In this section, I first present a theoretical model that presents the parameters of interest in my analyses and potential confounds in identifying these parameters. Next, I describe the analytic techniques I apply to mitigate the influence of these confounds.

4.1 Linking Teacher Quality Measures and Teacher Effects

To formalize the parameters of interest in this analysis, consider a stylized decomposition of student attainment of some outcome $y_{i(t)}$ into two parts:

$$y_{i(t)} = \mu_{jt} + \varepsilon_{it} \tag{4.1}$$

Where μ_{jt} represents a teacher's impact on student attainment of $y_{i(t)}$ and ε_{it} representing all non-teacher determinants of $y_{i(t)}$. In the case of this analysis, I am interested in estimating the extent to which teachers' observation scores Obs_{jt} capture their underlying effect on various student outcomes. This can be accomplished by regressing $y_{i(t)}$ on Obs_{jt}:

$$y_{i(t)} = \gamma Obs_{jt} + \varepsilon_{it} \tag{4.2}$$

Obs_{jt}, as with all other teacher quality measures in this analysis, is standardized to be mean zero with unit variance. γ, the parameter of interest, represents the extent to which differences in Obs_{jt}

predict differences in $y_{i(t)}$. In an ideal setting in which students are randomly assigned to teachers, γ can be represented as:

$$\gamma = \frac{cov(Obs_{jt}, y_{i(t)})}{var(Obs_{jt})} = \frac{cov(Obs_{jt}, \mu_{jt} + \varepsilon_{it})}{var(Obs_{jt})} = \frac{cov(Obs_{jt}, \mu_{jt})}{var(Obs_{jt})} \tag{4.3}$$

Where the magnitude of the γ depends on how much the dimensions of teacher quality captured by the observation rubric covary with teachers' underlying ability to affect a given student outcome. Under experimental conditions, γ can be given causal warrant and interpreted as the expected effect on outcome $y_{i(t)}$ of a student being switched from a teacher with an average observation score to a teacher with an observation score 1 SD above the average. If γ is statistically and substantively significant, this would indicate that observation scores serve as meaningful signifiers of differences in teacher quality with respect to student attainment of a particular outcome.

Under business-as-usual settings, estimating the γ terms is complicated by the presence of student-teacher sorting that results in non-negligible covariances between the structural error term, ε_{it}, and components of teacher observation scores, Obs_{jt} (Clotfelter et al., 2006; Kalogrides et al., 2013). Additionally, interpretation of γ is further complicated when acknowledging that Obs_{jt} contains a mixture of both "true score" and error components:

$$Obs_{jt} = \mu^*_{jt} + e^1_{jt} + e^2_{jt} \tag{4.4}$$

where μ^*_{jt} represents a teacher "true score", as defined by what the observation rubric is intended to measure, and e^1_{jt} and e^2_{jt}, which represent random and correlated measurement error, respectively. Measurement error, in the context of observation scores, could arise from numerous sources including idiosyncrasies during the period in which a teacher is observed, rater imprecision, or more systematic forms of rater bias such as favoritism for a specific teacher (Ho & Kane, 2013; Hill, Charalambous, & Kraft, 2012). For the purposes of my analysis, I delineate two different types of measurement error with respect to whether they are (1) "random" or uncorrelated to student outcomes of interest (e^1_{jt}) or (2) are potentially correlated to student outcomes (e^2_{jt}). Distinctions

between these two types are subsequently discussed in further detail.

Put together, allowing for the presence of both measurement error and student sorting results in a formula for γ more complex than what is presented in the ideal case (Equation 4.3):

$$\begin{aligned}\gamma &= \frac{cov(Obs_{jt}, y_{it})}{var(Obs_{jt})} = \frac{cov(\mu^*_{jt} + e^1_{jt} + e^2_{jt}, \mu_{jt} + \varepsilon_{it})}{var(\mu^*_{jt}) + var(e^1_{jt}) + var(e^2_{jt})} \\ &= \frac{cov(\mu^*_{jt}, \mu_{jt}) + cov(\mu^*_{jt}, \varepsilon_{it}) + cov(e^2_{jt}, \mu_{jt}) + cov(e^2_{jt}, \varepsilon_{it})}{var(\mu^*_{jt}) + var(e^1_{jt}) + var(e^2_{jt})}\end{aligned} \tag{4.5}$$

Explaining the (co)variance terms that appear in Equation 4.5 helps motivate my approaches to identification and measurement. In the numerator, there are three additional covariances beyond $cov(\mu^*_{jt}, \mu_{jt})$, the "signal" covariance of interest. Two different types of selection bias are represented in the covariances $cov(\mu^*_{jt}, \varepsilon_{it})$ and $cov(e^2_{jt}, \varepsilon_{it})$, which represent selection on the true score and correlated measurement error components of Obs_{jt}, respectively. Students would select on measurement error, rather than true scores, if families are enticed to select a school due to an exceptionally charismatic principal who also rates her teachers leniently. Lastly, $cov(e^2_{jt}, \mu_{jt})$ represents covariance between measurement error in teachers' observation scores and teachers' true effect on a given outcome. This occurs if observation scores fail to exclusively measure the constructs implied by their rubrics and instead capture "off-construct" aspects of teacher quality. In practice, because my analysis does not make a distinction between what observation scores are "intended" to measure and what they do measure empirically, this covariance is functionally treated as a source of signal. While uncorrelated measurement error, e^1_{jt}, does not appear in the numerator of Equation 4.5 as it is, by definition, unrelated to other terms in the model, it contributes variance in the denominator of the equation, resulting in attenuation bias.

The empirical challenge of this dissertation will be to identify ways to remove or restrict undesirable covariances from Equation 4.5 while preserving true score variance. I address these issues in two ways. First, I describe approaches for purging or mitigating measurement error from observation scores using multiple adjustment procedures. Specifically, I use a mixture of (1) "leave-out"

procedures, which omit observation scores selects years and/or raters and (2) value-added style regression adjustment to remove sources of variation in observation scores that are potentially endogenous with respect to a student outcome of interest. Next, in lieu of random assignment, I implement a "teacher switching" design, as used in Chetty et al. (2014a)) that mitigates bias due to endogenous within- and between-school sorting between students and teachers. Under the identifying assumption that students do not react endogenously to "aggregate changes" in teacher quality, this "teacher switching" design provides quasi-experimental variation in students' exposure to teachers of different observation score levels. Finally, building upon the observation score-specific findings, I describe how my analytic strategy can be applied to compare the estimated effects of changes to observation scores and value-added estimates.

4.2 Measurement Error and Classroom Observation Scores

The previous section describes, in generic terms, how the presence of measurement error may bias attempts to estimate the effect of changes to any teacher quality measure on student outcomes. Understanding the nature of these potential biases, and how to correct for them, requires further explication on how and what types of measurement error manifests in observation scores. Discussions on the presence of measurement error are abundant in the teacher value-added literature, and indeed, the ways these issues are motivated (and eventually addressed) in this dissertation are heavily influenced by this research. However, it is important to note that the measurement processes for teachers observation and value-added scores differ in such ways that require different approaches for accounting for their respective sources of measurement error. In the context of teacher value-added, teacher effectiveness is estimated directly from student test scores. Therefore, factors affecting student test performance beyond teacher effectiveness have a direct path to being forms of measurement error for the resulting value-added estimates. Disparities across teachers in the types of students assigned and school resources available, naturally, are chief concerns when estimating teacher value-added.

While stakeholders generally consider observation scores to be a more transparent measure

than teacher value-added (Goldring et al., 2015), the underlying measurement process used to score the former is arguably more complex, and certainly more idiosyncratic, than the process used to estimate the latter. Observation scores are sometimes described as an "inputs-based" measure, contrasting them with the "outputs-based" scores provided by value-added. This description is not strictly true in the case of the TEAM rubrics, as rubric items measure both teacher inputs (e.g., "Teacher displays extensive content knowledge of all the subjects she or he teachers") and student outputs (e.g., "Students are consistently well-behaved and on task"). As a result, measurement error may function different even for items within the same observation rubric. For observation items that are scored according to student behaviors, many of the "value-added' concerns regarding error due to student sorting can be analogously applied to the observation score case. Like a teacher whose value-added scores are negatively impacted by being assigned students with lower levels of academic skill, a teacher who is either consistently assigned poorer-behaved students or whose students are idiosyncratically poorer-behaved on the day of an observation may suffer from lower scores on the "Managing Student Behavior" item due to the types of students she has been assigned rather than her underlying skill on that item. In contrast, for "input-based" items based on teacher practices, error driven by students functions through causing teachers to modify their practice or through affecting a rater's score independent of teacher practice, i.e., a rater assuming a teacher shows high levels of content knowledge because the teacher has been assigned high-performing students. Additionally, a separate type of measurement error that affects "input"-based items but not teacher VA is day-to-date fluctuations in teacher performance. One would not expect a single "off day" from a teacher to meaningfully affect student test score performance since student achievement is ostensibly the sum result of a students' year-long exposure to a teacher. However, if this "off day" occurs on the day of a classroom observation, this performance dip could have substantial impact on that teacher's observation scores.

Beyond differences in what teacher value-added (student outputs) and observation scores (teacher inputs and student outputs) are intended to measure, the involvement of a human rater in the observation score process substantially changes how measurement error is introduced into both mea-

sures. Raters can impart their own error into observation scores as well as moderate the extent to which endogenous student and teacher characteristics impart error. Fixed rater errors might manifest in the form rater leniency, in which certain raters are more prone to assign higher or lower scores than others. Additionally, raters may harbor biases toward certain types of teachers (e.g., racial or gender bias) or specific teachers. Both types of rater bias are unique forms of measurement error for which there is no equivalent in the value-added case. Additionally, while student background characteristics, as previously described, do have a theoretical pathway through which they affect observation scores, this influence is not direct as it is in the case of teacher value-added, but rather, is moderated by a teacher's rater(s). During planned observations, teachers evaluated under TEAM are encouraged to discuss classroom context with their rater prior to the observation. Certain raters may take these contextual factors into consideration, performing ad-hoc adjustments in the rating process that account for different types of classrooms. Other raters might not account for these factors at all, potentially opening a direct path for student characteristics to affect observation scores.

The adjustment procedures that I consider in this analysis focus specifically on the role that raters and students have in driving measurement error within observation scores. There are infinitely many scenarios through which measurement error can be imparted upon a teacher's observation score. For the purposes of describing how these errors may bias my γ estimates, however, these errors can be categorized according to whether (1) they are persistent or non-persistent and (2) whether they are correlated with student outcomes. First, factors that induce measurement error may be idiosyncratic events, specific to a given observation period or year, or may stem from persistent factors that contaminate teachers' observation scores across multiple years. Examples of idiosyncratic factors that may introduce noise to observation score include a teacher being ill or a fire alarm going off during an observation period, while persistent factors might include teacher characteristics, both observable (e.g. race/ethnicity, gender) and unobservable (e.g. likeability) in nature or biases held by specific raters that a teacher is observed by year after year.

Whether measurement errors are correlated or uncorrelated with student outcomes determines

the nature of bias these errors will impart to estimates of γ. The presence of uncorrelated measurement errors, as is well-known, leads to estimates of the effect of observation score changes on student outcomes that are attenuated relative to the effect of changes in the "true scores" on those same outcomes. For example, a fire alarm going off during the observation period is an example of a non-persistent, uncorrelated error as it may potentially affect a teacher's observation score but likely has no effect on any student outcome not measured on that same day. A stylized example of a persistent but uncorrelated source of measurement error may be if a particular observer systematically rates women higher than men but otherwise does not treat male and female teachers, or their students differently. In both cases, these factors inflate the variation of teachers observation scores but produce no corresponding variation in student outcomes, leading to attenuated estimates.

Sources of measurement error that also affect student outcomes introduce the possibility of directional bias into estimates of interest. Again, correlated errors can be both non-persistent and persistent in nature. For example, a school-wide flu might result in an increase in student absences while also temporarily depressing teacher observation scores if observations are conducted during periods in which teachers and/or raters are ill. As a result, this flu creates an endogenous negative relationship between student absences and observation scores. The flu is an example of a type of non-persistent (i.e., specific to a given year) correlated error. Plausible examples of persistent correlated measurement errors can stem from rater biases that manifest in both the observation scoring and student assignment processes. For example, if a principal is biased against a particular teacher, the principal may both artificially deflate this teacher's observation scores and assign low-performing students to this teacher. Should both the principal and this teacher remain employed in the same school, this is a type of measurement error in observation scores that is both persistent and correlated with student outcomes. Additionally, persistent, correlated errors can manifest if there is a systematic connection between administrators' direct effects on both student outcomes and observation scores (e.g., principals who systematically assign lower observation scores are more likely to be effective at raising test scores).

In my analysis, I explore methods for creating measures of teacher effectiveness from raw ob-

servation scores that omit or mitigate different types of measurement error. Specifically, I use a combination of leave-out estimation, shrinkage, and regression adjustment to produce three different "adjusted" observation score measures, (1) Leave-Year-Out (LYO), (2) Rater Fixed Effect Adjusted (RFE), and (3) Leave-Rater-Out (LRO) estimates. Each method is described in further detail below.

4.2.1 Producing Adjusted Observation Scores

For the purpose of accountability calculations, teachers are evaluated using an annual overall observation score, which is calculated by averaging across the observation item-specific scores assigned to them in that year. As described previously, these observation scores are likely contaminated by multiple sources of measurement error that may result in either attenuation or directional bias if used as regressors.

The three adjustment methods I employ, (1) Leave-Year-Out (LYO), (2) Rater Fixed Effect Adjusted (RFE), and (3) Leave-Rater-Out (LRO), account for different sources of errors. First, LYO adjustments account for "non-persistent" sources of measurement error that may affect both a given student's outcomes and her teacher's observation score during the year (t) in which they were assigned to one another by using observation scores from years *other* than year t. To address errors stemming from persistent sources that would be unaccounted for by the LYO process (i.e., persistent sorting to different types of raters and students), I calculate two additional adjusted observation score measures that, in addition to omitting endogenous year t observation scores, regression adjust scores on the basis of observable student characteristics and either remove rater fixed effects from scores (RFE) or omit scores from any year that were assigned by a teacher's year t rater(s). The LYO, RFE, and LRO procedures are described in more detail below.

Leave-Year-Out (LYO) Estimates. The first adjustment approach I consider is the "leave-year-out" (LYO) method, which constructs an adjusted observation score using data from all years *not* used to construct the student outcomes of interest (Chetty et al., 2014a). Therefore, for any student outcome in year t, a "leave-out" version of the observation score is constructed using data from

any year that is not year t. I subsequently refer to the years from where outcomes are drawn (year t in this example) as "in-sample" years and years where outcomes are not drawn as "out-of-sample" years (non t years) where "sample" is defined specifically as the years where the student data used to construct the outcome variable were assigned to a given teacher. Similarly, if the outcome is operationalized as a gain score that uses data from years t and $t-1$, the LYO measure used in that regression will omit in-sample observation scores from years t and $t-1$.

There are several options for how to combine out-of-sample data to form LYO estimates, ranging from simple averaging, similar to what is done in Jacob, Lefgren, and Sims (2010), to the "drift" adjustments done in Chetty et al. (2014a) that place different weights on out-of-sample information based on their proximity to the in-sample year(s). In my dissertation, I opt for a relatively simple leave-out procedure followed by Blazar and Kraft (2017) that uses maximum likelihood estimation to obtain LYO scores that are shrunken in accordance with how many out-of-sample scores are available for each teacher. Denoting $\overrightarrow{t}$ as a vector of all years other than year t (i.e., the "out-of-sample" years), I estimate the following null teacher random effects model separately for each year t:

$$Obs_{j\overrightarrow{t}} = \mu_j + e_{jt} \tag{4.6}$$

Where $Obs_{j\overrightarrow{t}}$ represents teachers' overall observation scores from all out-of-sample years. From this model, I calculate BLUPs of the teacher effect, i.e., $\hat{\mu}_j$, and use these predictions as my LYO observation score measure. The BLUP can be written as:

$$\hat{\mu}_{jt} = \overline{Obs}_{j\overrightarrow{t}} \frac{\sigma_j^2}{\sigma_j^2 + \sigma_e^2/\overrightarrow{T}} \tag{4.7}$$

Where $\overline{Obs}_{j\overrightarrow{t}}$ is an average of the Obs_{jt} in all years other than t, multiplied by a shrinkage factor $\frac{\sigma_j^2}{\sigma_j^2+\sigma_e^2/\overrightarrow{T}}$. σ_j^2 and σ_e^2 are estimates of between-teacher and within-teacher variance in the year-specific observations scores, respectively, and are estimated via maximum likelihood. Because this model is fit separately for each year, the exact estimates of σ_j^2 and σ_e^2 used in the shrinkage procedure also vary for each year though σ_j^2 generally accounts for roughly 70 percent of total

variation. $\overrightarrow{T}$ is the total number of years, not including t, a teacher has a non-missing value of Obs_{jt}. $\overrightarrow{T}$ is the only parameter in the shrinkage factor that varies across teachers and serves to decrease the amount by which $\overline{Obs}_{j\overrightarrow{t}}$ is shrunk as the number of years of observation scores a teacher has increases. This procedure is repeated for every year of data in the sample to produce $\hat{\mu}_{jt}$ (i.e. the shrunken LYO estimate) for every teacher-year combination.

The LYO approach rests on the assumption that (1) there is a time-persistent (or "stable") component of teacher effectiveness, making teacher effectiveness scores in non-t years informative of teacher effectiveness in year t and (2) the sources of measurement error that may confound the relationship between student outcomes and teacher effectiveness scores in year t do not manifest in teacher effectiveness scores from non-t years. Observation scores from out-of-sample years are not inherently free from measurement error. However, in omitting in-sample observation scores, the LYO process ensures that the measurement error contained in the LYO estimate does not stem from an endogenous factor that also affects student outcomes in that year (e.g., school-wide flu). Additionally, to the extent that non-persistent factors are genuinely non-persistent, LYO estimates, by averaging information across multiple years, have a better chance of "cancelling out" idiosyncrasies from one year with idiosyncrasies in the next, minimizing the amount of error variation stemming from idiosyncratic factors.

LYO estimation, as with any leave-out procedure, is a brute force method, representing a trade-off between removing a source of strong signal variation during the in-sample year in hopes of omitting error variation. Therefore, whether the LYO procedure forms reasonable estimates of teacher effectiveness during the in-sample years depends on the extent to which the components of teacher effectiveness that are captured by observation scores are persistent across years for teachers. A check for whether a LYO approach is feasible is simply to calculate within-teacher autocorrelations for observation scores across years. If observation scores fluctuate strongly across years, leave-out methods will likely be ineffective in producing precise estimates of effectiveness within the in-sample year. Conversely, LYO methods will be much more successful if there is evidence of persistence in the variation captured by observation scors across years. Figure 4.1 shows

pooled autocorrelations between observation scores 1-5 years apart from one another. Observation scores show an autoregressive pattern, with adjacent year observation scores correlated at .71 with the autocorrelation monotonically decreasing to .45 between observation scores five years apart.

The autoregression profile depicted in Figure 4.1 is similar to the profile of test score residuals used in Chetty et al. (2014a) to construct value-added estimates, a promising finding that supports the plausibility of using out-of-sample observation scores in a similar manner to construct LYO adjusted observation scores.[1] Furthermore, the autocorrelations in Figure 4.1 could potentially be used to inform how to weight teachers' observation scores across different years to create more precise leave-out estimates. Currently, the random effects model described in Equation 4.7 weights teachers' scores from each year identically. However, as indicated by the observation score autocorrelations, we would expect that observation scores from more proximate years would be more informative for predicting teacher effectiveness during the "in-sample" year than more distal ones. This type of differential weighting is implemented and referred to as "drift adjusting" by Chetty et al. (2014a) and could, in theory be implemented for observation scores. In Chetty et al. (2014a), autocovariances across years are used to calculate weights for each of the out-of-sample years, with respect to year t. These weights are calculated under the assumption that the covariance between scores in any two years is only a function of the amount of time between them, i.e., the covariance between teacher scores in 2014 (t) and 2013 ($t-1$) is the same as the covariance between 2013 (t) and 2012 ($t-1$). I estimate year-specific autocorrelations in Table 4.1 to see whether this assumption holds in the context of Tennessee observation scores.

While the general autoregressive pattern holds for all years (i.e. $t+1$ is more correlated with t than $t+2$), the strength of the autocorrelations appear to increase with each year. The correlation between residualized observation scores during the first two years of TEAM (2012 and 2013) is 0.66. Adjacent year correlation grows to 0.75 in the latest years available in the sample (2017 and 2018), with corresponding increases for $t,t+2$ and $t,t+3$ correlations as well. As teachers and

[1] Moderate-to-high autocorrelations are a necessary, but not sufficient, condition to demonstrate the validity of the LYO process as these correlations could be the result of both persistent signal and error variance across years. Other adjustment methods delve more deeply into accounting for persistent error variation in observation scores

observers continue to adapt to the TEAM observation system, the relationship between adjacent-year observation likely has not reached a "steady state", meaning that the stationarity assumptions used by Chetty et al. (2014a) do not hold in the context of my data. Therefore, direct application of the drift adjustment as used in CFR may not be appropriate given the evolving relationship between teacher observation scores across time within my data. The development of an appropriate method for drift adjustment when stationarity assumptions do not hold is a valuable topic for future research.

LYO-adjusted scores account for non-persistent sources of measurement error but remain susceptible to sources of error that systematically reoccur across multiple years. To address this, I produce two additional types of adjusted observation scores, a Rater Fixed Effect Adjusted (RFE) and a Leave-Rater-Out (LRO) score, that, like the LYO process, also omit out-of-sample scores but perform additional adjustments to account for persistent sources of measurement error. The RFE and LRO scores both use teacher value-added style regression adjustment to adjust scores on the basis of observable classroom characteristics to mitigate bias stemming from patterns of non-random student sorting. Where the RFE and LRO scores differ is in their approach for accounting for the presence of rater-driven errors. I describe the calculation of the RFE and LRO scores below in further detail.

Rater Fixed Effect Adjusted (RFE) Estimates. Observation scoring as done for teacher evaluation does not include any explicit adjustments for differences in the students and raters to which a teacher is assigned, two factors which prior research suggests could be substantial sources of measurement error (Grissom & Loeb, 2017; Steinberg & Garrett, 2016). To the extent that measurement error due to students or raters occurs idiosyncratically, these types of non-persistent errors would be accounted for using the LYO method. However, errors of this sort are likely to persist across years. Abundant research shows that specific types of teachers are often assigned specific types of students (Clotfelter, Ladd, & Vigdor, 2005; Kalogrides et al., 2013) and because raters are most likely school administrators, if teachers do not change schools, they are likely to be be observed by the same rater across years.

One method for removing some persistent sources of error from observation scores is to apply value-added style regression adjustment to raw observation scores prior to combining them to form a leave-out estimate.(Whitehurst et al., 2014) Specifically, I build from the LYO processes described above by first residualizing teachers' observation scores against observable student characteristics and rater fixed effects. Once residualized, the out-of-sample observation scores are combined using the same null random effects model used to make LYO scores (Equation 4.7). I refer to these regression-adjusted leave-year-out scores as Rater Fixed Effect Adjusted (RFE) scores, the name referring to the specific method employed to account for rater errors. To produce RFE scores for each teacher, I first begin by regressing teachers' item-level observation scores on a vector of classroom average student covariates (X_{jt}), rater fixed effects (ψ_r), and teacher fixed effects (α_j):

$$\begin{aligned} Obs_{jrt} &= X_{tt}\beta + \alpha_j + \psi_r + \varepsilon_{jrt} \\ Z_{jrt} &= Obs_{jrt} - X\hat{\beta} - \hat{\psi}_r \end{aligned} \tag{4.8}$$

The vector X_{jt} consists of student characteristics typically used in the estimation of teacher value-added, including student race/ethnicity, fender, FRPL, SPED, and ELL designation, and lagged achievement and attendance. Using the coefficients estimated in the "first stage" model, I remove the influence of both the student characteristics and rater fixed effects from teachers observation scores to calculate a residualized observation score Z_{jrt}. Teacher fixed effects, α_j, are not removed from observation scores, but rather, are included to protect against the estimated coefficients on the student characteristics and rater fixed effects from "overcontrolling" for correlations between observation scores and student/rater characteristics that are due to sorting along "true" teacher effectiveness (Ballou et al., 2004; McCaffrey et al., 2004. Table 4.2 shows estimates from the residualizing equations under multiple fixed effect specifications including the two-way teacher and observer fixed effect specification that is eventually used to residualize scores in analysis.

Echoing results from Steinberg and Garrett (2016), classroom characteristics are significantly associated with teachers' observation scores across all specifications, with the estimates generally

smallest, but still significant, under the two way fixed effect specifications. The direction of the coefficients indicates that teachers assigned higher-performing, better-resourced students also typically earn higher observation scores, even when accounting for persistent teacher and rater effects. Using the estimated coefficients from Column 4 as the estimated rater fixed effects, I calculate residualized item-level scores, averaging these scores within teacher-year to form annual observation score residuals, Z_{jt}. These annual residuals are then modeled according to Equation 4.7 in order to obtain shrunken RFE estimates.[2]

Leave-Rater-Out (LRO) Estimates. The RFE method is capable of accounting for fixed differences in rater scores, e.g. certain raters consistently assigning higher scores than other raters. These fixed rater errors could contribute to systematic bias in my primary analyses if raters who are systematically more likely to assign higher/lower scores were also more likely to systematically raise or lower student outcomes. While plausible scenarios could be described that conform to this pattern, such as tougher raters also being more effective at raising student test scores than easier raters, it is unlikely that rater biases would manifest in ways that are fixed across all teachers that he or she is rating. Rather, rater biases are more likely to occur within certain rater-teacher pairs, such as a principal displaying favoritism toward a specific teacher or raters holding subconscious racial or gender biases. If animus between a particular principal and teacher manifests in the form of the principal both (1) assigning lower observation scores and (2) assigning lower performing students to that teacher, bias in obervation scores specific to particular raters and teachers can result in biased estimates of the relationship between observations and student outcomes. Because these biases depend on the specific interaction between raters and teacher characteristics, observed or unobserved, adjustments for rater fixed effects will have limited ability for correction. Likewise, if raters hold biases across years, these errors will affect teachers' observation scores across multiple years, contaminating the LYO process.

[2]One important note regarding regression adjustment is that correlations between observation scores and student characteristics do not necessarily constitute a form of measurement error as these correlations could result from interactions between student characteristics and teacher quality (i.e., teachers are genuinely more effective in years when they are assigned better-prepared students). Therefore, removing the influence of student characteristics potentially constitutes a form of overcontrolling, though prior research suggests that the magnitude of heterogeneities in teacher effectiveness according to students' incoming characteristics is small (Lockwood & McCaffrey, 2009).

Rather than modeling and residualizing rater fixed effects, the leave-rater-out (LRO) procedure accounts for rater error by omitting any observation score, or portions of observation scores, in out-of-sample years assigned by raters who also rated that teacher during the omitted in-sample years. Similar to the RFE approach, the calculation of LRO scores begins with residualizing observation scores to partial out the influence of observable classroom characteristics; the residualizing regression model uses teacher fixed effects to obtain "clean" estimates, differing slightly from the RFE model which uses both teacher and rater fixed effects. The RFE and LRO methods differ in respect to how rater errors are treated, the former opting to remove estimated rater fixed effects while the latter leaves out scores from in-sample raters altogether. Residualized observation scores assigned from out-of-sample raters, averaged to teacher-year level, are then used to predict leave-rater-out (LRO) estimates using the same random effects procedure used to produce the LYO and RFE estimates (Equation 4.7).

To illustrate data availability when calculating LRO scores, consider the hypothetical observation score data for Mr. Johnson, presented in Table 4.3.

For the purposes of this example, the in-sample years are 2013-14 and 2014-15, where a change in student test scores is the outcome of interest. Under the LYO approach, all of Mr. Johnson's observation scores from years other than 2014 and 2015 would be used to construct the LYO estimate. The LRO approach similarly does not use observation scores from these in-sample years but will also omit observation scores from out-of-sample years that were assigned by Mr. Johnson's in-sample raters, Todd and Wolpin. Mr. Johnson's scores in 2013 and 2016 are where the LYO and LRO methods differ. As an out-of-sample year, Mr. Johnson's 2016 scores would be included in his LYO estimates. However, because Todd and Wolpin also served as Mr. Johnson's raters in 2016, these scores would be omitted when calculating LRO estimates. In 2013, Mr. Johnson received scores from two raters, Wolpin and Sanchez, the latter of whom did not rate Mr. Johnson during the in-sample years. The LRO estimates for Mr. Johnson would use only the 2013 ratings given by Sanchez, omitting the portion of his scores that year that were given by Wolpin.

Relative to omitting 2013 scores entirely, the decision to still use portions of observation scores

not scored by in-sample raters allows me to include more teachers (and teacher-years) than a more conservative approach of omitting this year altogether. The use of partial scores in certain years will entail that teachers' observation scores in that year may not represent scores from every item of the TEAM rubric. Within the TEAM system, teachers are never rated on all domains of the TEAM rubric in a single classroom visit. Scores for the "Professionalism" domain are scored at the end of the academic year and scores for the "Environment" and "Planning" domains are typically scored during separate visits; "Instruction" is typically scored during each visit. Because of the bunching of domain scores within visits, it is very unlikely that, for teachers scored by multiple raters, that each rater will have assigned a score for that teacher on all items of the TEAM rubric. As result, it is likely that LRO estimates will only omit certain domains in certain years, depending on whether that domain was scored by an in-sample or out-of-sample rater.

Lastly, while the LYO method depends on teachers having multiple years of observation score data, the LRO method adds an additional stipulation that teachers receive scores from different raters across years as well. For teachers who receive observation scores across multiple years from the same rater, an LYO estimate can be calculated but an LRO estimate cannot. The LRO method may be especially restrictive in rural districts where teacher and principal mobility is lower and there are fewer administrators available within a school to conduct observations. Fortunately, the patterns of teacher-rater assignment in my data allow for LRO estimation. On average, teachers included in my analytic sample receive observation scores from two distinct raters in any given year and seven distinct raters over the course of the panel. Overall, only a third of teachers are rated by the exact same raters across adjacent years, indicating that LRO estimates can be calculated for the majority of teachers who are eligible for LYO estimates.

In my analysis, I compare the robustness of my results to all three adjustment methods. In Table 4.4, I calculate pairwise correlations between teachers scores across the three adjustment methods (LYO, RFE, LRO), in addition to correlations with teachers' unadjusted observation scores (OBS), and leave-out value-added estimates (VA), averaged across all subjects in a given year. These correlations are useful for gauging how much the various adjustment methods result in meaningful

differences in teachers' observation scores in addition to how observation and value-added measures of teacher effectiveness differ relate within my sample.

Generally, the various observation score measures are strongly correlated with one another, with the lowest correlation being between the unadjusted and RFE observation scores ($r = 0.577$) and the highest being between the LYO and LRO scores ($r = 0.907$). All observation score measures are correlated with teacher VA estimates bewteen 0.335-0.386, a range consistent with prior findings from the literature (Chin & Goldhaber, 2016).

4.3 Identifying Teacher Effects Using Observation Scores

In this dissertation, I am interested in determining whether differences in teachers' observation scores credibly capture differences in how these teachers affect various student outcomes. This difference in expected student outcomes for a one standard deviation difference in teacher observation scores are what I refer to as an observation score "effect" throughout my analysis. As with many analyses, an experimental setting would offer the most internally valid way to estimate observation score effects. Balanced on expectation due to randomization, the differences in students outcomes between two classrooms, scaled by the difference in observation scores between the teachers of those classrooms, would provide an unbiased estimate of the effect of interest.

In lieu of any such randomization, I use other methods that attempt to isolate sources of exogenous variation in students' exposure to teachers of different observation score levels similar to what would be generated due to random student-teacher assignment. Specifically, I use two identification strategies for identifying observation score effects: (1) a naive OLS model that captures the cross-sectional relationship between teacher observation scores and student outcomes and (2) a teacher-switching quasi-experiment that leverages teacher mobility across departments as a source of potentially exogenous variation in observation scores Chetty et al. (2014b). The OLS model can be estimated using the following:

$$y_{i(t)} = \gamma Obs_{jt} + \varepsilon_{i(t)} \tag{4.9}$$

Where Obs_{jt} represents teacher j's observaton score (adjusted using either the LYO, RFE, or LRO methods) and $y_{i(t)}$ representing a target student outcome of interest. These models also include indicators for year and grade level (i.e. elementary, middle, high). The year subscript for the student outcome measure is wrapped in parentheses to denote that certain outcomes are repeatedly measured (e.g. test scores, absences) while others are observed as "snapshots" at a certain point in time for a given student (e.g., on-time HS graduation, post-secondary enrollment at age 19, wages at age 25). For models where student test scores are the outcome of interest, I stack observations by test score subject to increase precision and include a subject indicator to the model. For all OLS models, since the outcome is a level of a student outcome at a certain point in time, all teacher observation scores will be "leave-one-out" measures that omit either observation scores from year t in the case of the LYO and RFE methods or, for the LRO procedure, observation scores assigned by year t raters. The OLS model uses the same cross-classroom comparisons described in the experimental hypothetical to identify observation score effects and, under the identifying assumption that the structural error term, $\varepsilon_{i(t)}$, is well-balanced across different levels of Obs_{jt}, would provide an unbiased estimate of the effect of being assigned to a teacher with an observation score one SD above the average. Table 4.2, which shows correlations between observation scores and a host of immutable student characteristics suggests this assumption does not hold within the data. Furthermore, even if Equation 4.9 were augmented with student covariates, Rothstein (2009), among others, suggests the complex nature in which students sort make it unlikely that controlling for student observables is sufficient for guaranteeing conditionally exogenous variation in Obs_{jt}.

The second research design I use is the "teacher switching" quasi-experiment as implemented by Chetty et al. (2014a). Rather than using cross-sectional differences in student outcomes and teachers' observation scores, the teacher switching design uses adjacent cohorts passing through the same school as counterfactuals for one another, with variation in students' exposure to observation score levels driven by teacher mobility across schools and departments. The merits of the teacher switching design can be motivated through example. Consider James, a student entering the fourth grade at Anderson Elementary in the fall of 2017. James's exposure to teacher quality is

influenced by which of Anderson Elementary's fourth grade teachers he is ultimately assigned to. However, his exposure to teacher quality is also influenced by which *set of teachers* are assigned to teach fourth grade at Anderson Elementary in 2017. The intuition behind the teacher switching design is that while James and his family can readily negotiate among which of Anderson Elementary's set of fourth grade teachers he is assigned to, it is substantially more difficult for him to manipulate the set of teachers available to him in 2017.

Teacher switching designs work by restricting identifying variation to changes in the *average* quality of all teachers a student could have been potentially assigned to within a particular department. Departments are defined as school-grade combinations in the elementary grades, school-grade-subject combinations in the middle school grades, and school-subject combinations in the high school grades. Identifying variation in observation scores at the department level, as the name of the design implies, is driven by teacher mobility in the form of (1) switches, where teachers change departments across years, (2) exits, where teachers leave the profession across years, and (3) entries, where new teachers enter the profession in a given year. As teachers enter and leave particular departments across years, cohorts entering that department will be exposed to different levels of average teacher quality. Returning to our earlier example, if a fourth grade teacher at Anderson Elementary retires following 2017, the average quality of teachers available to James will differ from the average teacher quality available for Helen, a student entering fourth grade at Anderson Elementary in the fall of 2018. Under the assumption that Helen is unlikely to switch departments, by changing schools or being held back or promoted, in response to changes in teacher staffing, the change in teacher quality brought on by this teacher retirement serves as a plausibly exogenous source of variation in observation scores.

I implement the teacher switching design by starting with the student-level data represented in Equation 4.9 but then (1) aggregating these data to the department-year level and (2) taking first differences of the aggregated outcomes and teacher observation scores:

$$\Delta\overline{y}_{dt} = \alpha + \gamma\Delta\overline{Obs}_{dt} + \Delta\overline{\varepsilon}_{dt} \tag{4.10}$$

where $\bar{y}_{dt}$ and $\overline{Obs}_{dt}$ are the average student outcomes and teacher observation scores, respectively, of department d in year t. Analogous to the OLS model, when the outcome is student test scores, I create department-year-subject averages and stack these subject-specific cells together in the analysis. $\bar{\varepsilon}_{dt}$ represents the unobserved determinants of student outcomes, aggregated to the department level. The Δ symbols indicate that each variable is first differenced, where $\Delta\bar{y}_{dt} = \bar{y}_{dt} - \bar{y}_{dt-1}$. All specifications include year and school level (i.e., elementary, middle, high) indicators and test score models include subject indicators. Department-year cells with fewer than 10 students with non-missing values for the outcome of interest are dropped from the analysis and results are weighted are according to how many students are included in each cell. Importantly, because the outcome is a gain score and thus uses data from years t and $t-1$, I use out-of-sample "leave two out" observation scores that estimate teacher quality in year t from data in all years other than t and $t-1$; the LRO scores for the teacher switching design omit all scores assigned from raters who issued scores for that teacher in years t and $t-1$. Students assigned to teachers missing leave-out score used are omitted from the analysis prior to aggregating data to the department level.

4.3.1 Potential Threats to Validity in the Teacher Switching Design

Endogenous student mobility. The key identifying assumption for the teacher switching design is that changes in department average observation scores are uncorrelated with changes in student heterogeneity in that same department.[3] *Primia facie*, such endogenous cross-department sorting

[3] A separate concern regarding identification is the extent to which department average observation scores serve as suitable instruments for teacher-level observation scores. Findings from C. K. Jackson and Bruegmann (2009) and Sun, Loeb, and Grissom (2017) suggest that there are positive spillovers to teacher quality from the introduction of high quality peer teachers to a given school or department. The evidence of teacher peer effects raise concern that department average quality may be a function of more than merely the sum of individual teacher quality, potentially limiting whether results from the teacher switching design can be used to make inference regarding the impacts of individual teachers. There are two reasons why, within this study, these spillovers may post only minor threats to validity. First, evidence of teacher peer effects does not in and of itself invalidate the teacher switching design. Rather, these peer effects must change the department's effect on student outcomes in ways not reflected in changes to department observation scores. At face value, the TEAM observation rubric does contain a "Professionalism" domain which, in theory, directly captures some measure of colleague interaction through which potentially endogenous peer effects might flow. Secondly, results from the analytic models (Chapter 5) show that teacher switching estimates, across nearly all outcomes, are smaller than OLS estimates, suggesting that the magnitude of any effects of uncaptured positive spillovers are smaller than the effects of any bias reduction occurring by switching from the OLS to teacher switching model.

would be difficult given the complexities of switching schools. However, we can test this assertion by using students' immutable characteristics, X, (e.g. test scores, race/ethnicity, FRPL eligibility) as the outcome in the teacher switching design:

$$\Delta\overline{X}_{sgkt} = \alpha + \gamma\Delta\overline{Obs}_{sgkt} + \Delta\overline{\varepsilon}_{sgkt} \tag{4.11}$$

In Figure 4.2, I use the teacher switching design to estimate "effects" of changes to observation scores on six student immutable charcateristic (Proportions FRPL, SPED, Black, Female, Prior Test Scores, Prior Attendance). Prior to discussing the relationships between observation scores and characteristics, the plots in Figure 4.2 importantly reveal the large proportion of departments that show little to no variation in observation scores changes across years. Approximately two-thirds of departments saw year-over-year changes in LYO observations of less than .05 standard deviations, with the majority of identifying variation used in the analysis coming from the one-third of cells that saw more substantial changes in teacher quality due to teacher switching.

When we examine the point estimates for each plot, changes in observation scores produced statistically significant "effects" on all six student characteristics with these effects being in the direction hypothesized by patterns of assortative matching (e.g., more advantaged students are matched with higher performing teachers), with much of this "on-average" relationship being driven by cells with extremely large changes in observation scores. The magnitude of the effects on student demographic characteristics is substantively very small and arguably negligible. For example, a standard deviation increase in observation scores results in a .6 percentage point increase in the percent of students who are FRPL eligible. Relationships between changes in observation scores and prior achievement are more troubling and echo the results and concerns raised by Rothstein (2017) in his validation of the teacher switching design.

Whether relationships between lagged outcomes and leave-out measures are cause for concern is a topic of debate Bacher-Hicks et al., 2014; Chetty et al., 2014a). To avoid contamination with sources of measurement error in years used to construct the gain score outcomes (years t and $t-1$), "leave-two-out" scores omit data from these years and rely on data from "out-of-sample" years

(e.g., $t-3$, $t-2$, $t+1$, $t+2$) to construct measures of effectiveness. However, when the outcome of interest is a lagged outcome, gain scores will be calculated using data from years $t-1$ and $t-2$. The presence of $t-2$ data in both the outcome and the "leave-out" score potentially introduces the same source of measurement error to both the left and right hand sides of the equation, e.g. a school-wide flu in year $t-2$ affecting both a teacher's $t-2$ observation score and a student's $t-2$ attendance rate. Therefore, the use of lagged values as "outcomes" reintroduces sources of measurement error that would be otherwise omitted were the outcome constructed using "current year" values.

One difference between the context of this study and CFR's that may render the former more vulnerable to endogenous sorting under the teacher switching design is high-stakes nature within which observation scores are collected. Teacher VA, particularly of the variety estimated by CFR, was not used or publicized in any meaningful way in New York during the period of CFR's data. While principals and families may be able to detect, as suggested by Jacob and Lefgren (2008), elements of teacher quality reflected in teacher VA estimates, the specific scores themselves are largely unobservable to actors who would be involved in the teacher-student assignment process.

In contrast, Tennessee's observation scores are a highly visible and consequential element of the state's teacher evaluation system. Given that observation scores are both (1) directly observable by both administrators and teachers and (2) attached to meaningful human capital decisions, there is additional incentive to explicitly sort on observation scores that is not present in the context of the CFR teacher VA studies. Additionally, while it may be unlikely for students to follow particular teachers to different schools and/or departments within schools, the strategic movement of teachers to specific cohorts within a school may be much more feasible (Atteberry, Loeb, & Wyckoff, 2016). As additional checks for robustness, for all outcomes, I present additional specifications of the teacher switching design that (1) include changes in observable student characteristics and (2) incorporate school-year fixed effects which limit identification variation to differences across departments in a given school-year. The use of school-year fixed effects restricts identifying variation to comparisons of year-over-year changes across different departments within the same school,

accounting for endogenous factors that affect specific school-year combinations whereas the inclusion of student controls mitigates bias due to observable changes in student composition.

Missing observation score data. An additional concern brought up by Chetty et al. (2014a) and its subsequent replications (Bacher-Hicks et al., 2014; Rothstein, 2017) is how teachers with missing leave-out estimates are treated in the analysis. Because leave-out scores require that teachers have multiple years of evaluation data, teachers are not likely to be missing leave-out estimates at random. Teachers with missing leave-out scores are more likely to be less experienced overall, new to Tennessee public schools, or new to teaching in a tested grade and subject. If students are non-randomly assigned to teachers missing leave-out estimates, dropping classrooms led by these teachers potentially induces sampling bias (e.g., weaker students are systematically assigned to new teachers). A solution proposed by CFR is to impute zero for teachers missing the specific leave-out estimate used in the analysis. This decision corresponds with the logic that, absent other information, teachers should be assumed to have average effectiveness. Imputing zeroes for teachers with missing value-added scores caused validation coefficients to deviate from 1 in both Chetty et al. (2014a) and Rothstein (2017) but had minimal effect on the results of the quasi-experiment conducted by Bacher-Hicks et al. (2014). Like previous papers using the teacher switching design, my standard analysis omits classrooms taught by teachers for whom a leave-out estimate cannot be calculated. Therefore, to test the robustness of my results, I also estimate models that use the "zero-imputation" method to "fill in" sample average observation scores for teachers missing leave-out estimates. Because every observation score measure used in the dissertation is standardied, this entails imputing zeroes for teachers with missing leave-out scores. Zero-imputation is applied only to teachers in school-years in which other at least one other teacher was eligible for leave-out scores, i.e., teachers employed in districts that do not use the TEAM observation rubric are still omitted from the analytic sample.

4.3.2 Comparing and Combining Observation and Value-Added Scores

Lastly, I use the teacher switching design to estimate relationships between changes in teacher VA estimates and student outcomes of interest. In isolation, this teacher VA analysis serves as a replication of the prior teacher VA validations such as Chetty et al. (2014a), Chetty et al. (2014b), Bacher-Hicks et al. (2014), and Rothstein (2017). However, in the context of research questions posed in my dissertation, the ability to estimate the predictive power of both teacher value-added and observation scores within the same sample of students and teachers offers a way to compare and assess whether both measures differ in the information they capture about teachers' impacts on different student outcomes.

I estimate teacher VA effects on student outcomes (i.e. γ^{VA}) within a sample of cells for which both teacher VA and observation scores are available. I refer to this sample as the "matched" sample and the sample used for the observation score-only analysis, which is comprised of all cells with available observation scores, as the "full" sample. Since classroom rosters are identified using linkage files made for the purpose of VA estimation, differences between the "matched" and "full" samples are largely driven by (1) cells in years prior to the implementation of TEAM in 2011-12[4] and (2) cells where rosters were available but not test scores, such as English IV and US History.

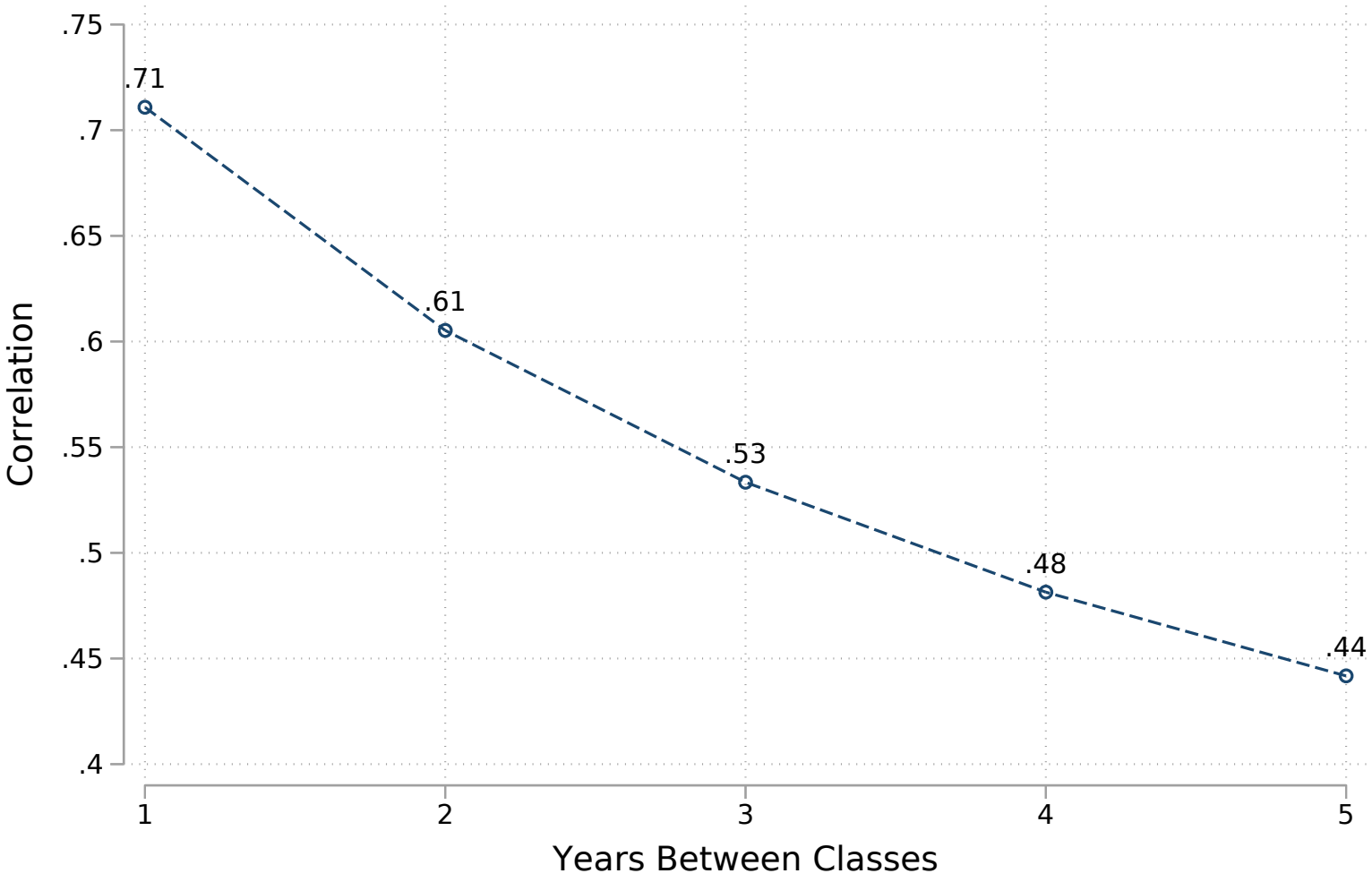
The simplest method for comparing the predictive power of observation scores and value-added estimates is to estimate the teacher switching design separately for each measure and simply compare the magnitude and significance of the estimated γ coefficients across the various student outcomes in the matched sample. This "vote-counting" method is valuable for determining whether each measure, used in isolation, is capable of detecting teacher impacts on the student outcome of interest. To explicitly compare whether observation scores and teacher VA hold distinct information on teachers impacts, I can the extend the teacher switching design by including both teacher observation and value-added estimates as explanatory variables to estimate "multiple measure" models:

[4]This restriction is later eased by the use of"back-projected" scores which I describe in further detail in the Results chapter.

$$\Delta\overline{Y}_{dst} = \alpha + \gamma_1 \Delta\overline{Obs}_{dst} + \gamma_2 \Delta\overline{VA}_{dst} + \Delta\overline{\varepsilon}_{dst} \tag{4.12}$$

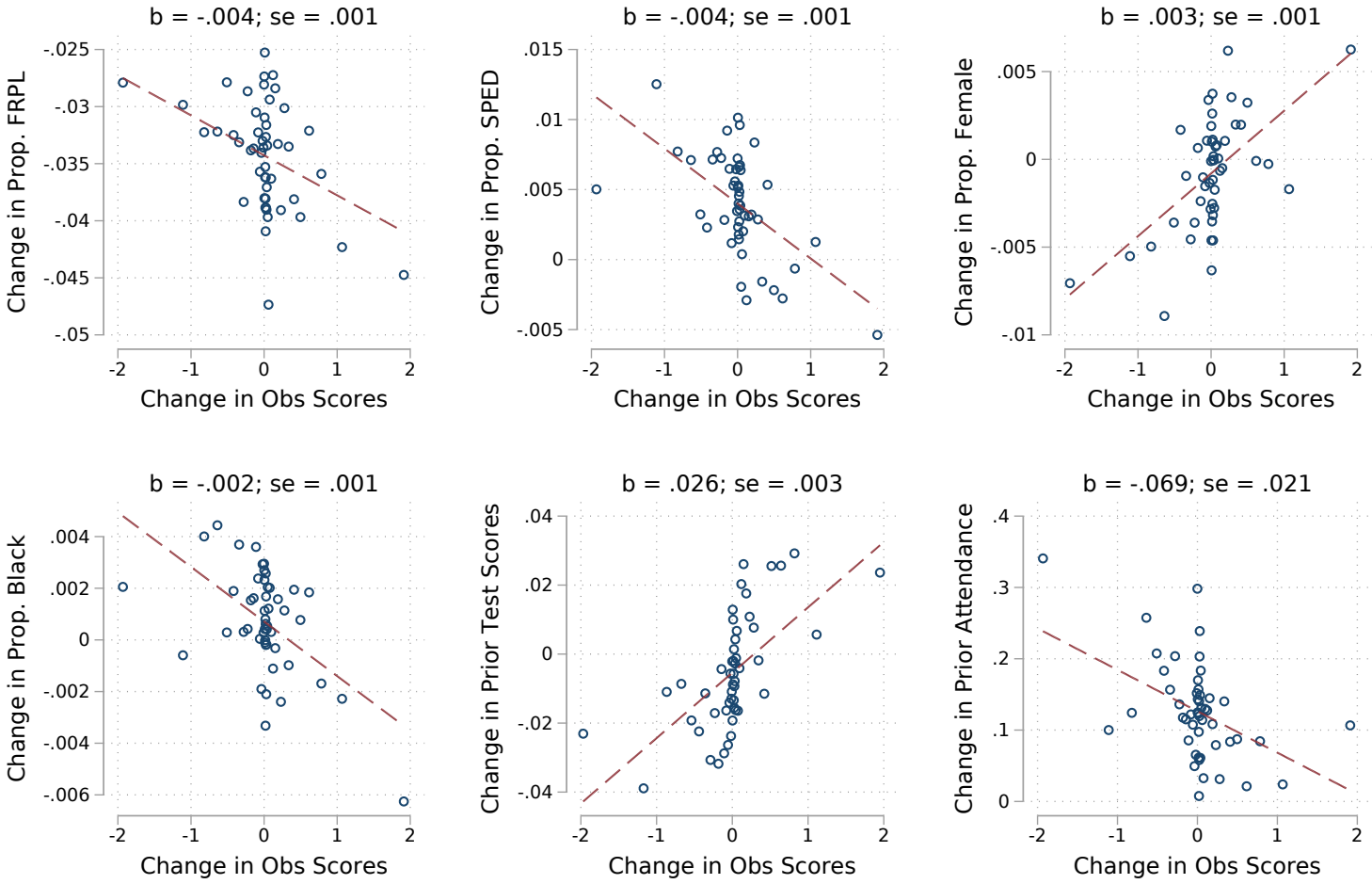
where γ_1 and γ_2 are the effects of changes in observation and value-added scores, respectively, on changes in outcome Y, aggregated to the department level. Because teacher VA is subject-specific, the data for all outcomes is aggregated to the department-subject-year level. A given teacher quality measure need not have a significant γ in the multiple measure design to be a valid predictor of outcome Y if it has a significant γ in the single measure design. This scenario is likely to occur if both measures share common information about teacher effects for a given outcome. Rather, the purpose of the multiple measure design is to identify whether each measure holds distinct, meaningful information regarding teachers' impacts on a particular student outcome.

Figure 4.1: Autocorrelations of Observation Scores



Note: Figure shows pairwise autocorrelations of teachers' annual observation scores

Figure 4.2: Changes in Student Characteristics and Teacher Observations



Note: Figure shows leave-year-out (LYO) observation score effects on department average student background characteristics. Estimates are obtained using the teacher switching (QE) design, meaning the unit of analyses are department-year cells. Point estimates and standard errors are presented above each plot. Both observation score and student background characteristics measures are operationalized as first differences (t - t-1) and results are weighted by the number of students in each cell. Models include dummy variables for subject, year, and grade level (i.e., elementary, middle, high) and use standard errors clustered at the school-cohort level.

Table 4.1: Year-Specific Autocorrelations of Observation Scores

	2012	2013	2014	2015	2016	2017	2018
2012	1						
2013	0.664	1					
2014	0.550	0.678	1				
2015	0.496	0.594	0.693	1			
2016	0.465	0.531	0.601	0.724	1		
2017	0.434	0.487	0.542	0.634	0.750	1	
2018	0.404	0.449	0.491	0.561	0.645	0.748	1

Note: Table shows pairwise year-specific autocorrelations of teachers' annual observation scores.

Table 4.2: Observation Scores and Classroom Characteristics (Item Level)

	(1)	(2)	(3)	(4)
Pr. FRPL Eligible	–0.132***	–0.303***	–0.0315***	–0.0469***
	(0.002)	(0.005)	(0.004)	(0.004)
Pr. SPED	0.0892***	0.0693***	0.0687***	0.0826***
	(0.002)	(0.002)	(0.005)	(0.005)
Pr. Female	0.158***	0.151***	0.0374***	0.0495***
	(0.005)	(0.005)	(0.006)	(0.006)
Pr. Black	–0.365***	–0.159***	–0.103***	–0.0460***
	(0.002)	(0.007)	(0.006)	(0.008)
Pr. Hispanic	–0.244***	–0.0972***	0.0729***	0.0631***
	(0.009)	(0.011)	(0.013)	(0.014)
Pr. Other	0.284***	0.0665***	0.0504**	0.0759***
	(0.014)	(0.016)	(0.019)	(0.020)
Pr. ELL	0.0814***	0.138***	–0.152***	–0.0706***
	(0.008)	(0.010)	(0.012)	(0.013)
Avg Prior Score	0.119***	0.0836***	0.0475***	0.0469***
	(0.001)	(0.001)	(0.002)	(0.002)
Avg Prior Att. Rate	1.012***	0.893***	0.0874**	0.155***
	(0.025)	(0.030)	(0.031)	(0.033)
N	3244565	3244565	3244565	3244565
R2	0.125	0.240	0.414	0.458
Model	OLS	SY FE	TCH FE	TCH + OBS FE

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors in parentheses.

Note: Table shows estimates from regressions of teacher-item-year observation scores on average classroom characteristics. In addition to an OLS model, three different fixed effect strategies are used: School-Year (SY), Teacher (TCH), and Teacher + Observer (TCH + OBS)

Table 4.3: Leave Rater(s) Out Data for Mr. Johnson

Years	Test Scores	Obs. Scores	Rater(s)
2012		Obs_{12}	Sanchez, Holmes
2013		Obs_{13}	~~Wolpin~~, Sanchez
2014	y_{14}	~~Obs_{14}~~	~~Todd~~, ~~Wolpin~~
2015	y_{15}	~~Obs_{15}~~	~~Todd~~, ~~Wolpin~~
2016		~~Obs_{16}~~	~~Todd~~, ~~Wolpin~~
2017		Obs_{17}	Murphy

Table 4.4: Correlations Across Observation and Value-Added Scores

	OBS	LYO	RFE	LRO	VA
OBS	1				
LYO	0.713	1			
RFE	0.577	0.782	1		
LRO	0.643	0.907	0.750	1	
VA	0.335	0.372	0.386	0.365	1

Note: Table shows pairwise correlations between teacher-year unadjusted observation scores (OBS), leave-year-out estimates (LYO), rater fixed effect adjusted (RFE) estimates, leave-rater-out (LRO) estimates, and teacher value-added (VA) estimates, averaged across subjects. LYO, RA, LRO, and VA scores are all "leave-one-out" measures that omit year t scores, i.e., the year represented by OBS. All scores are standardized within year.

Chapter 5

Results

This chapter provides estimates of the effects of students' exposure to different levels of observation scores (or observation score-based measures) on multiple student outcomes of interest. I divide student outcomes into two categories. First, I show estimated observation score effects on time-varying K-12 outcomes, including student test scores, absences, and suspensions. The robustness of these results are tested using the different observation score adjustment methods (LYO, RFE, LRO), zero imputation, and examining whether domain specific observation scores are differentially predictive of these outcomes. Next, I present effects of observation score changes on long-run student outcomes which are measured as "snapshots" at a particular age (e.g., high school graduation by age 18, post-secondary completion by age 23). Results for these outcomes are presented separately due to the use of "back-projected" observation scores to extend the number of available students with which to analyze long-run effects.

5.1 Time-Varying K-12 Outcomes

In Table 5.1, I show OLS (Columns 1-2) and teacher-switching (QE) (Columns 3-6) results for the effects of changes in LYO observation scores on student test score achievement. The OLS models use leave-one-out scores, standardized within year, that omit observation scores in year t when test scores in year t are the outcome and the QE models use leave-two-out scores, standardized within year, where observation scores in years t and $t-1$ are omitted when the outcome is the change in test scores between years t and $t-1$. The units of the independent variable in the teacher switching design is the change in standardized LYO scores, a measure not to be confused with the standardized change in LYO scores. For all models using test scores as outcomes, observations across different subjects (i.e., Math, ELA, Science, Social Studies) are "stacked" for additional

precision. Therefore, the sample sizes for the OLS and QE models refer to the number of student-year-subject and department-year-subject observations, respectively. All models use school-cohort clustered standard errors.

Column 1 in Table 5.1 shows the naive OLS regression, indicating that students taught by a teacher with an observation score one standard deviation above the average would be expected to score .21 standard deviations higher on standardized test scores than students assigned to a teacher with an average observation scores. Subsequent specifications that account for student heterogeneity in varying degree indicate that this naive estimate, as would be expected, suffers from positive bias due to the positive sorting of teachers and students. In Column 2, the OLS specification is extended by controlling for conventional student covariates (e.g., lagged test scores, student FRPL eligibility, SPED designation, race/ethnicity). The inclusion of this vector of student covariates halves the estimated observation score effect from .21 to .096; this result aligns, perhaps coincidentally, with Kane and Staiger (2008) who similarly find that the magnitude of estimated teacher effects halves when moving from "no control" specifications to specifications using lagged test scores.

The remaining columns (3-6) in Table 5.1 use the teacher switching design, which estimates the effects of department-average changes in observation scores on department-average student achievement. Differences across the teacher switching specifications do produce descriptively meaningful differences in point estimates, though 95 percent confidence intervals surrounding each estimate overlap for all pairwise comparisons across specifications except the standard QE (Column 2) and the QE with controls and school-year FE (Column 6); this pattern is broadly true across all models in this chapter. Using the "standard" QE specification, which does not include additional controls beyond subject and school level (i.e., elementary, middle, high) fixed effects, I estimate that, relative to a cohort of students passing through the same department in the prior year, an increase in the department average LYO observation scores by one standard deviation is expected to result in a 0.89 standard deviation increase in student achievement. This estimate from the standard teacher switching design is very similar, but slightly smaller, than the estimated observation

score effect obtained when using the student covariate adjusted OLS model (0.096). Extensions to the teacher switching design, including controls for changes in (aggregated) student-level covariates (Columns 4,6) and school-year fixed effects (Column 5,6) similarly result in decreases in the magnitude of the point estimates, though the most severe specification, which uses the teacher switching design and includes both student-level controls, still yields a significant estimated test score effect of .076.

Among student absences and student suspensions, I find similar patterns with regard to the relationship between LYO observation score effects and outcomes of interest. The magnitude of the naive OLS estimate for all outcomes is substantially larger than the estimated effects in all other specifications, providing further evidence that observed relationships between teacher quality and student outcomes under business-as-usual conditions are heavily confounded by numerous factors. Using the standard QE specification (Column 3) for each outcome, I estimate that a standard deviation increase in LYO observation scores results in a decrease of .15 student absences (Table 5.2) and a decrease of .019 student suspensions in a given year (Table 5.3). As with the student achievement results, the addition of school-year fixed effects and controls for year-over-year changes in observable student characteristics moves the estimated observation score effects on both absences and suspensions toward zero but these effects remain statistically significant at conventional levels.

One interesting consideration across K-12 outcomes is the extent to which point estimates obtained from OLS models that control for typically-available student covariates mirror the effects obtained from the more data-intensive teacher switching specification. Because the use of the department-level teacher switching design comes at severe cost to statistical power, a covariate-saturated student-level OLS model may be preferable on grounds of precision if there is no appreciable difference in bias reduction between the models. For the test score and student absence models, the saturated OLS model (Column 2) and the standard QE design (Column 3) produce relatively similar point estimates, with these effects differing by about 10 percent from one another (Achievement: .0961 vs. .0889; Absences: -0.133 vs. -0.145). However, point estimates from the two models measure differ when estimating suspensions effects, where the OLS and teacher

switching models estimate effects of -0.042 and -0.019, respectively.

The non-uniform manner in which model selection affects estimated effects strongly suggests that the unobservables that are accounted (and unaccounted) for by the various specifications differs across outcomes. While the use of an OLS model with commonly available student covariates suggest similar inference to the "teacher switching" estimates for student test scores and absences, these same covariates may be much more limited in their ability to reduce bias when estimating student suspension effects. In particular, it appears standard covariates may not properly account for sorting on behavioral and socioemotional skills that might affect students' disciplinary outcomes at greater magnitude than their academic or attendance outcomes. For this reason, I present results from the full range of OLS and teacher switching specifications for models across all outcomes. While model selection largely does not affect inferences regarding the significance of observation score effects, the magnitudes of the estimated effects often meaningfully differ across specifications.

5.1.1 Using Zero-Imputation to Account for Missing Teacher Data

A concern with models using leave-out estimates is that teachers with fewer years of experience are less likely to have the requisite number of annual observation scores to be eligible for the analysis. Therefore, if students are non-randomly sorted to teachers for whom leave-out estimates cannot be calculated (i.e., novice teachers), these students are more likely to be omitted from the analysis, inducing sampling bias. A robustness check to test the sensitivity of results to this sampling bias is to impute average observation scores for teachers who are missing leave-out estimates. In the case where teachers' scores are standardized, this means that teachers with missing leave-out scores are imputed to have scores of 0. This technique follows from Bayesian inference that, absent other information, teachers are assumed to be of average quality. Results for K-12 outcomes using the zero-imputed scores are presented in Table 5.4.

For achievement and absences, the use of zero-imputed scores does decrease the magnitude of the estimated effects. Comparing results from the "standard" teacher switching specification (see

), the estimated effect of a standard deviation increase in LYO scores on student achievement decreases from 0.089 to 0.074 when using zero-imputation and -0.145 to -0.108 for student absences. The decrease in estimated effects could be driven by both the omission of a non-random sample of students when using the non-imputed scores, as described previously, but could also result from increased measurement error due to assuming all teachers with missing scores are of average effectiveness. Given that more novice teachers are more likely to be missing leave out scores, and prior literature suggests that these teachers are likely to be less effective than their more experienced peers (Papay & Kraft, 2015), it may be more likely that teachers with missing scores would have lower-than-average observation scores. Importantly, despite decreases to the magnitude of the estimated coefficients, results using the zero-imputed scores remain statistically significant for all models and specifications. Effects on student suspensions are comparable whether zero imputation is used or not, with the imputed score effects actually being slightly larger than effects obtained when not using zero-imputation.

5.1.2 Comparing Methods for Accounting for Rater Error

Results shown in Tables 5.1, 5.2, and 5.3 use "Leave-Year-Out" (LYO) observation scores that omit a teacher's observation scores in years t and $t-1$ as a means of removing both random noise in single-year observation scores and covariation in observation scores and student outcomes due to common sources of measurement error. However, other persistent sources of measurement error, specifically the presence of the same raters across years, can potentially both (1) affect student outcomes in years t and $t-1$ and (2) teacher observation scores in years other than t and $t-1$. If these persistent errors are pervasive, the LYO approach may be inadequate for cleaning out sources of measurement error correlated with student outcomes in the "in-sample" years.

As I describe previously, I use two separate methods, Rater Fixed Effect Adjustment (RFE) and Leave-Rater-Out (LRO), to further adjust the LYO estimates to partially account for measurement error stemming from differences in student composition and raters across teachers. Both the RFE and LRO methods use value-added style regression adjustment to remove the "effects" of observ-

able student characteristics from observation scores but differ in how they approach adjustments for rater error. The RFE procedure models rater fixed effects and residualizes the estimated rater effects alongside observable student characteristics. The LRO method, by contrast, is a direct extension of the LYO process and, in addition to omitting all observation scores in years t and $t-1$, also omits scores, or portions of scores, in other years that were scored by raters assigned to that teacher in years t and $t-1$.

Full results for RFE and LRO estimates are presented in Tables 5.5 and 5.6, respectively. All three adjustment methods (LYO, RFE, LRO) adjust for measurement error in relatively crude fashion, removing, in all likelihood, both "signal" and "noise" from teacher observation scores. Therefore, the purpose of comparing results across the adjustment methods is less to select a "correct" method, but rather, to test the robustness of the estimated observation score effects against different assumptions of how measurement error might affect the relationships of interest. For all three K-12 outcomes, the effects across the three adjustment methods remain substantively similar and all remain statistically significant. Relative to estimates obtained using the LYO scores, the magnitude of effects obtained using the RFE scores are generally larger across all student outcomes whereas the coefficients are slightly smaller when using the LRO scores. The larger effects estimated using the RFE scores may reflect the residualization process "cleaning out" measurement error in the observation scores, producing a less-attenuated estimate than what is obtained when using LYO scores. Conversely, the smaller effects obtained from using LRO scores could result from (1) the LRO process omitting correlated rater biases that inflated the LYO observation score effects or (2) the LRO estimates suffering from a smaller signal-to-noise-ratio than LYO scores on account of incorporating fewer raw observation score ratings.

Lastly, a separate reason why estimates between adjustments procedures may differ is that the RFE and LRO scores are estimated only for teachers in districts reporting item-level observation scores, i.e., TEAM districts. Because the differences in estimates across the LYO, RFE, and LRO adjustments are relatively subtle and LYO estimates are available for a slightly larger share of teachers across the state, I use LYO observation scores as my default observation score measure

for all subsequent analyses.

In addition to comparing how the various observation score adjustments affect the estimates of observation score effects, I also estimate the effects of changes to the unadjusted single year observation scores that are used for accountability calculations in Tennessee on K-12 outcomes. I standardize the unadjusted scores in order to facilitate comparisons with effects using the adjusted scores. The comparison between the effects of unadjusted and adjusted scores can be helpful in several ways. First, differences across the unadjusted and adjusted estimates may be informative of what type of variation is being removed in the various adjustment procedures. The primary motivation in making observation score adjustments in this analysis is to mitigate the risk that the observation score "effects" I estimate are being driven by the measurement error, rather than true score, components of the observation scores. However, in performing these adjustments, it is likely that the resulting adjusted scores are likely more reliable than what could be obtained in typical observation score practice. Therefore, analyses using adjusted observation scores may give overly-optimistic estimates of impacts that could be obtained when using the less-reliable unadjusted observation scores. Secondly, because the adjusted scores are estimated using information across multiple years, changes in observation scores for a given department are primarily driven by changes in teacher assignment across years.[1] This construction largely omits variation in observation scores coming from changes within teachers, across years. The use of single-year scores allows these within-teacher changes to be incorporated into the analysis, providing an estimate of the effect of changes to observation scores that includes changes due to both teacher staffing and professional development.

Results from models using teachers' single-year, unadjusted observation scores are presented in Table 5.7 with Figure 5.2 comparing point estimates across all adjustments and K-12 outcomes. Naive estimates OLS estimates when using the unadjusted scores (Column 1) are roughly similar to estimates from equivalent models when using teachers' adjusted scores. However, when com-

[1] There is some degree of within-teacher variation in leave-out observation scores across years since the observation scores which are left out vary across years. However, this variation is very small as the ICC of leave-out scores within teacher, across years for any adjustment method if above 0.95

paring the estimates from the teacher switching design, unadjusted observation score effects are noticeably smaller for both student test scores and absences, with effects on student suspensions roughly the same across the unadjusted and adjusted observation scores. What is very clear from this pattern of results is that it is unlikely that the majority of the measurement error removed through the various adjustment strategies is of the variety that is correlated with student outcomes. Rather, much of the error removed through the adjustment process appears to be random "noise" that would otherwise attenuate the relationship between scores and student outcomes. [2] Thus, the various adjustments serve to produce a "cleaner" estimate of a teacher's effectiveness, as captured by observation scores, than would otherwise be obtained from using a single year of unadjusted observation scores. This finding suggests that while adjusted observation scores provided a less-attenuated estimate of how the underlying "signal" contained inside observation scores affects student outcomes, this estimate will overstate the precision with which decisions based on the actual observation scores used in TEAM can be used to affect student outcomes but may provide policymakers with a higher bound on what gains to the predictive power of observation scores could be if efforts to improve their reliability were to take place.

5.1.3 K-12 Observation Score Effects by Grade Level

An important moderator of the effect of observation score differences is grade level (i.e., elementary, middle, high). We might expect to see differences in effects by grade level for several reasons. First, the size of teacher effects on each student outcome may differ by grade level (e.g., differences in student absences are better explained by teacher quality in high school than in elementary school). If the size of the underlying teacher effect is larger for teachers at a given grade level than others, all else held equal, estimates of the observation score effect for that level will also be larger. Second, the observation process may be differentially reliable for teachers at dif-

[2] Given the direction of the coefficients, it is also technically possible that the unadjusted scores are contaminated with systematic but compensatory measurement error. However, this is unlikely given that regressions of observation scores on classroom characteristics, such as those calculated using my data (See Table 4.2) and experimental data from Steinberg & Garrett, 2016, suggest that the nature of this error is likely to be assortative, where being assigned more advantaged students is likely to raise, rather than lower, a teacher's observation score.

ferent grade levels. This could be due to systematic differences in either rater reliability at each grade level (e.g., middle school principals more reliable raters than elementary school principals) or differences in ability on certain rubric items being easier to discern for certain grade levels (e.g., differences in "Teacher Content Knowledge" being more readily apparent for high school teachers teaching more specialized content). Because Tennessee's observation rubrics are not differentiated by grade, this is unlikely to be a major factor but could subtly affect grade level differences in observation score effects. Lastly, and most obviously, under the assumption that teacher effects have a non-zero rate of decay (Jacob et al., 2010), proximity between when a student was assigned to a given teacher and when a particular outcome was measured will affect the magnitude of the estimated effects.

I examine how the K-12 effects of LYO observation score changes differ across students and teachers in elementary (Grades 3-5), middle (Grades 6-8) and high (Grades 9-12) school. Figure 5.3 below shows point estimates and 95 percent confidence intervals from the "standard" (i.e., no school-year FE, no covariates) specification of the teacher switching design that is modified to include an interaction between grade level and observation scores.

A clear pattern emerges across K-12 outcomes. The magnitude of observation score changes increases across grade levels, with the elementary and high school effects being significantly different from one another across all outcomes; the significance of differences between elementary and middle and middle and high vary by outcome. Importantly, the elementary school effects for both absences and suspensions are not significantly different from zero, indicating that the significance of the main effects for these outcomes is driven primarily by teachers and students in middle and high school cells. Because test scores, absences, and suspensions are contemporaneously measured with when a student was assigned to a given teacher, proximity to student outcomes is equivalent across grade levels. Therefore, differences in effects by grade level are driven by either differential (1) reliability of observation scores and/or (2) magnitude of underlying teacher effects at each grade level. Of the two, the latter is more likely to be the primary factor. The nature of absences and suspensions lend themselves to teachers of older students having more

prominent effects, as the attendance decisions of younger students are often predominantly driven by parental decisions and the occurrence of serious disciplinary infractions requiring suspensions being much rarer for younger students. When it comes to test score effects, the muted teacher effects in elementary grades are likely driven by the complexity, or lack thereof, of the material being tested. Family members can more readily help younger students learn simpler material, dampening the importance of in-school resources. However, as lessons become more complex and out-of-school resources are less available, the importance of teacher quality on the mastery of these topics becomes increasingly important, growing the size of teacher effects on these students. This result is echoed in prior literature on teacher test score effects that generally find that the variance of teacher effects is larger among middle school teachers relative to elementary school teachers (Bacher-Hicks et al., 2014; Chetty et al., 2014a).[3]

5.1.4 Domain-Specific Effects

Thus far, the observation score measures in my analyses are constructed using, when possible, information from all observation score items made available under a given adjustment procedure. However, it may be possible that certain domains of the TEAM observation rubric are more sensitive to certain outcomes than others. For example, we may expect observation ratings from the "Instruction" domain to be more indicative of teachers' ability to raise test scores while "Environment" ratings may be more predictive of behavioral outcomes such as suspensions. Correlations between teachers' scores across each of the four TEAM domains (see Table 5.8), ranging from 0.650 (Planning and Professionalism) to 0.848 (Instruction and Environment), show moderate-to-strong relationships between teachers' domain-specific scores.

I investigate domain-specific observation score effects by first calculating leave-year-out (LYO) estimates using teachers' annual average scores for each domain with comparisons across outcomes plotted in Figure 5.4 with full results presented in Tables 5.9, 5.10, and 5.11. First, in Table

[3]C. K. Jackson (2014) finds contrary results, concluding that test score results are smaller at the high school level than in younger grades.

5.9, I show domain-specific LYO effects on student achievement. I find that changes in observation scores across all four domains significantly capture changes in student achievement. While the differences in magnitude are small, LYO estimates constructed using teachers' "Instruction" scores yield the largest effects (0.094) on achievement across all specifications, closely followed by LYO estimates constructed using "Environment" scores (0.084). Effects using "Planning" (0.080) and "Professionalism" (0.073) lag slightly behind with regards to capturing teacher effects on all K-12 outcomes.

With regard to the number of student absences (Table 5.10) and suspensions (Table 5.11), "Instruction" and "Environment" observation scores again yield the largest effect estimates, with "Environment" effects being slightly larger on absences and "Instruction" effects being slightly larger for suspensions. The differences between "Instruction" and "Environment", as currently estimated, are too subtle to make distinctions regarding the types of information on teacher quality captured by each domain. However, the domain-specific results do suggest that scores from both domains generally exceed the predictive power of teachers' scores from the "Planning" and "Professionalism" domains. Researchers interested in using observation scores as a measure of teachers' impact on student outcomes may consider using only teachers' scores from the "Instruction" and "Environment" domains in order to potentially construct a cleaner signal of these impacts.

5.2 Long-Run Student Outcomes

5.2.1 Using "Back-Projected" scores

Tennessee is an ideal setting in which to study observation scores given that TEAM is one of the longest-standing multiple measure evaluation systems in the country. However, despite being mature in comparison to other statewide teacher evaluation systems, TEAM, and the data collected as part of the system, has still existed for less than a decade. While I have access to student information and classroom rosters since 2006-07, the later adoption of the TEAM observation score system limits my analysis to the period within which TEAM is active, 2011-12 to 2017-18, losing

five years of available student data. This relatively short panel severely limits the number of cohorts that are available for analysis of long-term outcomes that are measured in early adulthood, prohibiting the analysis of particular outcomes (e.g., annual wages at age 30) altogether. Compounding this issue is the fact that, with exception of English IV, no departments included in my analysis are comprised predominantly of high school seniors. By comparison, the analysis conducted by Chetty et al. (2014b) leverages 20 years worth of student-teacher data and 15 years of IRS-provided outcome data, far exceeding the reach available within my data.

One way to extend the number of observations that can be included in the analysis is to apply the principle of "leave-out" estimation to "back-project" LYO observation scores for teachers observed in the data during the pre-TEAM period (2006-07 to 2010-11). Following the same random effects approach as is used for all other leave-out scores (Equation 4.6), since all years during the TEAM period are "out-of-sample" years, the back-projected observation LYO score for any teacher will simply be the shrunken average of all of a teacher's available observation scores.

Obtaining "back-projected" scores using a "leave-out" approach is, in theory, sensible, but carries a number of substantive and technical challenges. Teachers' observation scores show moderate to strong autocorrelations (see Figure 4.1), suggesting that information during the TEAM period will perform reasonably for producing back-projected scores, particularly for years closest to 2011-12. However, because Tennessee's statewide evaluation system did not exist prior to 2012, back-projected scores are best thought of as estimates of teacher effectiveness made using available information rather than explicit projections of a teacher's observation score during that period.[4]

A simple test to examine the feasibility of using back-projected scores is to conduct the teacher switching analysis for the annually repeating K-12 outcomes (test scores, absences, suspensions) using student outcome data from only the pre-TEAM period, exclusively using back-projected scores as the observation score measure. If the estimated effects using the back-projected scores are comparable to those estimated in the standard analyses, this would provide evidence that these scores function similarly to the standard LYO scores, granting confidence that they can be used to

[4]The changing magnitude of the year-specific autocorrelations (see Table 4.1) between observation scores gives further cause for concern regarding the validity of back-projected scores.

augment the number cells available for my analysis of long-run outcomes. K-12 outcome results using only back-projected scores are presented in Table 5.12 with a visual comparison between LYO and back-projected scores plotted in Figure 5.5.

Estimates obtained using the back-projected scores are roughly similar to those estimated from the main analyses (Tables 5.1, 5.2, 5.3) though the nature in which the back-projected estimates differ varies across outcome. Effects on test scores (Panel A) appear to be the most "well-behaved" of the three K-12 outcomes, with estimates using back-projected scores (0.090) being virtually identical to LYO estimates from full sample 5.1. That teachers' observation scores in the future produce reasonable projections of how these teachers impacted student test score performance in the past speaks to the ability of observation scores to capture the persistent components of teacher effectiveness as they relate to teachers' effects on achievement. However, teachers' back-projected effects on student absences (Panel B) are more strongly attenuated than test score effects, with back-projected estimates (-0.07) being roughly half the size of estimates obtained from the TEAM-era sample (-0.15). Absence effect estimates using the standard quasi-experiment specification (Panel B, Columns 4-6) are only significant when using school-year fixed effects. The back-projected estimates of teachers' effects on student suspensions move in opposite direction from absences, with the back-projected estimates for suspensions being slightly larger than estimates from the main analyses.

5.2.2 Teacher Effects on Long-Run Outcomes

For analyses of students' long-run outcomes, a group of student outcomes which includes high school graduation, post-secondary attendance and completion, and early career wages, I use cells from the full range of student data, spanning from 2007-08 to 2017-18; outcomes from the first year of available data, 2006-07, are not used as prior test scores are used as a covariate in certain specifications. During the TEAM period (2011-12 to 2017-18), I use teachers' LYO observation scores as my measure of teacher effectiveness and their back-projected scores during the pre-TEAM period (2007-08 to 2010-11). For all long-run outcomes, I select a student's level of attainment at a spe-

cific age for that outcome in order to provide a more equal comparison of attainment across student cohorts. I use attainment at the following time points as my primary method of operationalizing student long-run outcomes:

- Obtained HS degree by age 18 (completion of 12th grade year)
- Enroll in a PS Institution by age 19
- Earned PS award/degree by age 23
- Annual wages at age 25

First, I examine the main effects of changes to students' exposure to teacher quality, as measured by LYO and back-projected observation scores, on their likelihood of graduating high school "on-time", defined as receiving a high school diploma by at least the end of their twelfth grade or age 18 year (Table 5.13). All but one specification yields a positive, significant effect of observation score changes on high school graduation though the magnitude of these coefficients is very small, ranging from a 1.2 percentage point increase per standard deviation increase in LYO observation scores in the naive model (Column 1) to .1 percentage points in the most saturated teacher switching models. One plausible technical reason for this smaller effect is imprecision in the graduation indicator as described in the Data section. As operationalized for analysis, it is likely that dropouts and students otherwise failing to attain on-time graduation (i.e., 0s for the graduation indicator) are conflated with students who have (1) moved out-of-state, (2) attend an in-state private school, (3) shift to homeschooling.

Next, in Table 5.14, I present results for the effects of observation score changes on four different types of post-secondary enrollment at the age of 19: enrollment at (1) any post-secondary institution, (2) a Bachelor's degree granting institution, (3) a non-Bachelor's degree granting institution, and (4) a non-THEC reporting institution where students' intended degree cannot be ascertained.

Across all types of post-secondary institutions (Panel A), a standard deviation increase in observation scores is estimated to result in roughly a 1.5 percentage point increase in the probability

of enrollment at age 19. From the institution-specific results, I find that much of the overall post-secondary enrollment effect is carried by enrollment in Bachelor's degree granting institutions (1.3 percentage point increase), with the effects on enrollments at Non-Bachelor's granting institutions (.8 percentage points) and Non-Reporting institutions (.4 percentage points) much smaller. The larger effect for enrollments at BA institutions would seem to suggest that increases to teacher quality affect both the intensive and extensive margins of post-secondary enrollment: exposure to higher teacher quality results in both increased participation in post-secondary education in addition to an increase in the quality of post-secondary education, e.g., students who would have otherwise enrolled in a 2-year program now enrolling in a 4-year program as a result of improvements to teacher quality. Observation score effects on the probability of completing a post-secondary degree mirror patterns found in post-secondary enrollment, with positive, significant effect on the attainment of any type of post-secondary award that is driven largely by effects on Bachelor's Degree attainment (See Table 5.15). Specifically, a standard deviation increase in LYO scores, using the primary teacher switching specification (Column 3), is estimated to result in a 1.7 percentage point increase in the probability of earning any post-secondary degree or credential, with a 2.1 percentage point increase specifically for Bachelor's degree attainment and no significant effect on Associates Degree or Post-Secondary Certificate completion. As with the wage outcomes, which are presented subsequently, it is important to note that data on post-secondary completion is limited to students who graduate from public, in-state institutions, and therefore, students attending private or out-of-state institutions are not represented in the post-secondary completion results.

In Table 5.16, I present estimates for the effects of changes to LYO scores on annual wages at ages 23, 24, 25. Recall that the wage data available for my data set comprise only in-state earnings reported by companies participating in Tennessee's unemployment insurance program, with out-of-state income and income stemming from self-employment being the largest omitted categories. Given that I find that changes to observation scores are linked to students' propensity to enroll in a post-secondary institution, with stronger effects for enrollment at Bachelor's granting institutions, I would expect that wage effects stemming from changes in teacher effectiveness to

manifest in later, rather than more proximate earnings. Indeed, the pattern of wage effects by age appears to correspond to this hypothesis. Using results from the standard teacher-switching design (Column 3) as a point of comparison, I find that a standard deviation increase in LYO scores results in increases of $111.8, $325.7, and $756.3 at ages 23, 24, and 25, respectively, with the age 23 effect not significant at conventional levels. A necessary note of caution when interpreting results from Table 5.16 are the small sample sizes for the teacher switching effects and subsequently large standard errors. For example, the 95 percent confidence interval for the standard QE specification (Column 3) Age 25 estimate spans from roughly $350 per standard deviation increase in LYO observation scores to $1150.

5.2.3 Long-Run Observation Score Effects by Grade Level

When effects on long-run outcomes are estimated separately by grade level (See Figure 5.6), like with K-12 outcomes, I find that changes to observation scores have the most pronounced, and precisely estimated, effects at the high school level. In particular, the grade level analysis indicates that the significant overall effect of observation scores on on-time high school graduation (Table 5.13) is driven exclusively by changes in teacher quality at the high school level. Post-secondary enrollment effects are strongest at the high school level, with a standard deviation increase in LYO scores estimates to result in a 2.6 percentage point increase in the probability of enrolling in any post-secondary institution at 19, with changes at the middle school level also statistically significant albeit at a much smaller magnitude (0.7 percentage points). Observation scores at all three grade levels appear to result in significant effects on the probability of earning any post-secondary credential at age 23, though the estimates for observation score changes at both the elementary and middle school levels are bounded with large confidence intervals relative to the more precisely estimated high school effect. Age 25 wage effects are identified almost exclusively from high school cells, and therefore, elementary and middle school-specific effects are not estimable with the current data.

The larger estimated high school teacher effects on long-run outcomes likely comes from both

high school teachers having larger effects on these long-run outcomes in addition to elementary and middle school teacher effects having experienced substantial decay by the time these outcomes are measured for students. The ability to graduate high school on time inherently depends on students successfully having earned credits in the courses taught by high school teachers, making differences in quality among teachers in these departments especially important. Additionally, while students' academic achievement in the elementary and middle grades may promote future learning and may help students establish placements in more advanced academic tracks in high school, students' high school GPAs are directly considered in college application processes in ways that their elementary and middle school grades are not, similarly granting outsized importance to high school teacher quality on these outcomes.

Even if there were no fundamental differences in how teachers across grade level affect the skills required to attain the various long-run outcomes, the estimated effects of high school teachers would still likely be the largest due to their proximity to the points in times at which student attainment for long-run outcomes are measured. It is possible that the impact of a 5th grade teacher on college-going skills on 5th grade students is similar in size to the impact that a 11th grade teacher has on these same skills on 11th grade students. But, if these contemporaneous impacts fade over time, the importance of teachers who are more proximate to the point in time in which an outcome is measured will be larger than those teachers who are more distal. The patterns in effects across grade level presented in Figure 5.6 are consistent with the notion that these effects decay over time. However, it should be noted that the specific rate at which these effects decay, particularly for non-test score outcomes, is uncertain. There is an extensive literature on the decay of teacher effects on test scores, with estimates of the average rate of decay after one year of 50 to 80 percent (Chetty et al., 2014b; Jacob et al., 2010; Kane & Staiger, 2008; Rothstein, 2010). Equivalent decay estimates are not available for long-run outcomes as they are not contemporaneously measured in the same fashion as test scores, but numerous studies examining the effects of various educational interventions on students' longitudinal outcomes (e.g., Chetty, Friedman, Hilger, et al., 2011; Chetty et al., 2014b; Heckman, Moon, Pinto, Savelyev, & Yavitz, 2010) find similar patterns of significant

effects on adulthood outcomes despite rapid decay on test score effects, suggesting that the ability for teachers and other educational inputs to impact student attainment of non-test outcomes may be more persistent than effects on academic achievement.

5.3 Comparing and Combining Observation and Value-Added Effects

While I find that changes in observation scores are statistically significant predictors of several student outcomes, it is reasonable to question whether the magnitude of these effects warrant their use in a teacher evaluation system. One method for benchmarking the utility of the information contained in observation scores is to compare estimated effects of observation score changes with analogously-estimated effects using teacher value-added (VA). As the "outcomes-based" teacher evaluation measure, teacher VA is often assumed to provide a strong tie to student outcomes, at potential cost to their clarity, buy-in from teachers and administrators, and clear application to formative evaluation, properties typically enjoyed by the observation score process (Goldring et al., 2015). Therefore, even if observation score effects are slightly smaller to than teacher VA effects on the same outcomes, stakeholders may find that these additional benefits to observation scores may be substantial enough to consider these measures on equal footing to teacher VA.

I first compare the effects of observation scores and VA estimates by estimating the "teacher switching" model separately for each measure of teacher effectiveness within a sample of department-year cells for which both observation scores and teacher VA estimates are available; I hereafter refer to this sample as the "matched" sample. For all outcomes, I use back-projected LYO observation scores in order to include cells during the pre-TEAM period (2007-08 to 2010-11). While the majority of departments included in my data set are eligible for this comparison analysis, I am unable to estimate VA scores for teachers in select HS departments for which I have access to student-teacher rosters but not test score data. The most important omission of this type is English IV cells, which are particularly important in the analysis of long-run student outcomes as English IV rosters constitute the majority of grade 12 observations for students included in my data set. Therefore, the "single measure" observation score estimates I obtain in the comparison analysis

will not necessarily be the same as those estimated in the full observation score models presented previously, particularly for analysis of long-run outcomes.

Figure 5.7 depicts LYO and teacher VA effects on K-12 outcomes with more detailed results presented in Tables 5.17, 5.20, and 5.18. Using the matched sample of department-year cells that have both LYO observation scores and teacher VA, I present estimates from the single (Columns 2,3 and Columns 5,6) and multiple measure models (Columns 1 and 4) for student test scores in Table 5.17. For all comparison analyses, I use teacher VA estimates that have been standardized to have mean zero and unit variance in order to facilitate comparison with the observation score effects. Therefore, the coefficient on teacher VA for student test scores does not lend itself to the familiar "equal to one" test of forecast bias used in Chetty et al. (2014a). These estimates are obtained using both the standard teacher switching design (Columns 1-3) and the teacher switching design that incorporates school-year fixed effects and changes in aggregated student characteristics as covariates (Columns 4-6). Using the standard QE specification, I estimate that one SD increases in observation scores and teacher VA result in .089 and .129 SD increases in student test scores, respectively; notably, the observation score effect on student test scores estimated within the matched sample closely mirrors the effect estimated in the full sample (see Table 5.1). Given that test score VA estimates are derived from student test scores, it is expected that teacher VA would exceed the predictive power of observation scores as it pertains to achievement. Therefore, that the magnitude of the observation score effect on test scores is roughly 70 percent the size of the VA effect should be interpreted as a promising finding with regard to the predictive validity of observation scores.

Using the matched sample of department-year cells that have both LYO observation scores and teacher VA, I present estimates from the single (Columns 2,3 and Columns 5,6) and multiple measure models (Columns 1 and 4) for student test scores in Table 5.17. For all comparison analyses, I use teacher VA estimates that have been standardized to have mean zero and unit variance in order to facilitate comparison with the observation score effects. Therefore, the coefficient on teacher VA for student test scores does not lend itself to the familiar "equal to one" test of forecast bias used in Chetty et al. (2014a). These estimates are obtained using both the standard teacher switching

design (Columns 1-3) and the teacher switching design that incorporates school-year fixed effects and changes in aggregated student characteristics as covariates (Columns 4-6). Using the standard QE specification, I estimate that one SD increases in observation scores and teacher VA result in .10 and .13 SD increases in student test scores, respectively; notably, the observation score effect on student test scores estimated within the matched sample closely mirrors the effect estimated in the full sample (see Table 5.1). Given that test score VA estimates are derived from student test scores, it is expected that teacher VA would exceed the predictive power of observation scores as it pertains to achievement. Therefore, that the magnitude of the observation score effect on test scores is roughly 75 percent the size of the VA effect should be interpreted as a promising finding with regard to the predictive validity of observation scores.

The multiple measure models for student test scores (Columns 1 and 4 of Table 5.17) show that, when both observation scores and teacher VA are used as regressors, both measures remain statistically significant predictors of student test scores albeit with smaller magnitudes. Importantly, this finding indicates that both observation scores and teacher VA detect distinct portions of teachers' effects on student achievement. Similar to the full sample results, the inclusion of school-year fixed effects and covariates attenuate the effects in both the single measure and full measure models, but all effects on student test score achievement remain positive and statistically significant.

Single and multiple measure estimates for student absences, Table 5.20 show that changes in observation scores and teacher VA are both (1) separately predictive and (2) have significant, orthogonal effects on absences. The magnitude of the coefficients for each teacher quality measure decreases only slightly when switching from the single measure to multiple measure models, suggesting that both measures largely capture distinct portions of teachers' ability to affect the attendance of their students.

Likewise, results for the "single measure" student suspension models (Table 5.18) indicate that both LYO observation scores and teacher VA are significant predictors of student suspensions, with the magnitude of the observation score effect twice as large as the teacher VA effects in both teacher

switching specifications. When regressing suspensions on both observation and VA scores in the multiple measure models (Columns 1 & 4), estimates for observation scores remain significant, albeit slightly smaller than the "single measure" models, while estimates for teacher VA fall out of significance, suggesting that the predictive power of teacher VA on suspensions largely overlaps with information on teacher quality already captured by observation scores.

For the K-12 outcomes, the magnitudes of the single measure observation score effects in "matched" sample are very similar to the effects from the 'full" sample observation score models. However, there are more prominent differences in the observation score effects between the matched and full samples when analyzing long-run outcomes, likely driven by the lack of Grade 12 observations in the matched sample. Observation and VA effects for high school graduation, post-secondary enrollment, and post-secondary completion are presented in Figure 5.8.

With on-time HS graduation (Table 5.19, neither the estimated effects using observation scores or teacher VA are significant, with point estimates routinely being below .1 percentage points. Previously discussed issues regarding data quality for high school graduation (see "Data" section) and evidence of significant effects only among high school teachers within the full sample (see Figure 5.6 may also contribute to the null findings obtained when estimating main effects within the "matched" sample.

With regard to Table 5.21, which shows single and multiple measure effects for post-secondary enrollment, the standard teacher switching specification (Columns 1-3) finds significant effects for both LYO observation scores (1.1 percentage point increase) and teacher VA (0.8 percentage point increase) in the single measure model, with observation scores retaining a significant, albeit smaller, effect on enrollment when included alongside teacher VA. Using the saturated teacher switching specification (Columns 4-6), I find that observation scores have a significant "single measure" effect but no significant effect in the "multiple measure" model. I do not estimate that teacher VA has a significant effect on post-secondary enrollment when using the saturated teacher switching specification.

The comparison analysis of observation and teacher VA effects on post-secondary completion

finds that, within the matched sample, observation scores have a significant effect on completion when used as the sole measure of teacher effectiveness (1.2 percentage points) in the standard teacher switching design. Neither observation scores nor teacher VA are significant in any other specification.

Lastly, the comparison analysis on annual wages at age 25 reveals a curious pattern of results which strongly suggest that the estimated wage effects for both observation scores and teacher VA are highly imprecise. In the teacher switching specifications shown in Table 5.23, changes in teacher VA are estimated to have large, significant effects on students' age 25 wages, though the range of point estimates is quite wide ($948 per standard deviation increase in teacher VA to $1730). Estimates of observation score effects, in comparison are highly imprecise, with standard errors nearly twice as large as the point estimates across all specifications in the matched sample. At face value, it would appear that changes in teachers' VA scores are more sensitive to teachers' ability to increase wages than equivalently sized changes in observation scores.

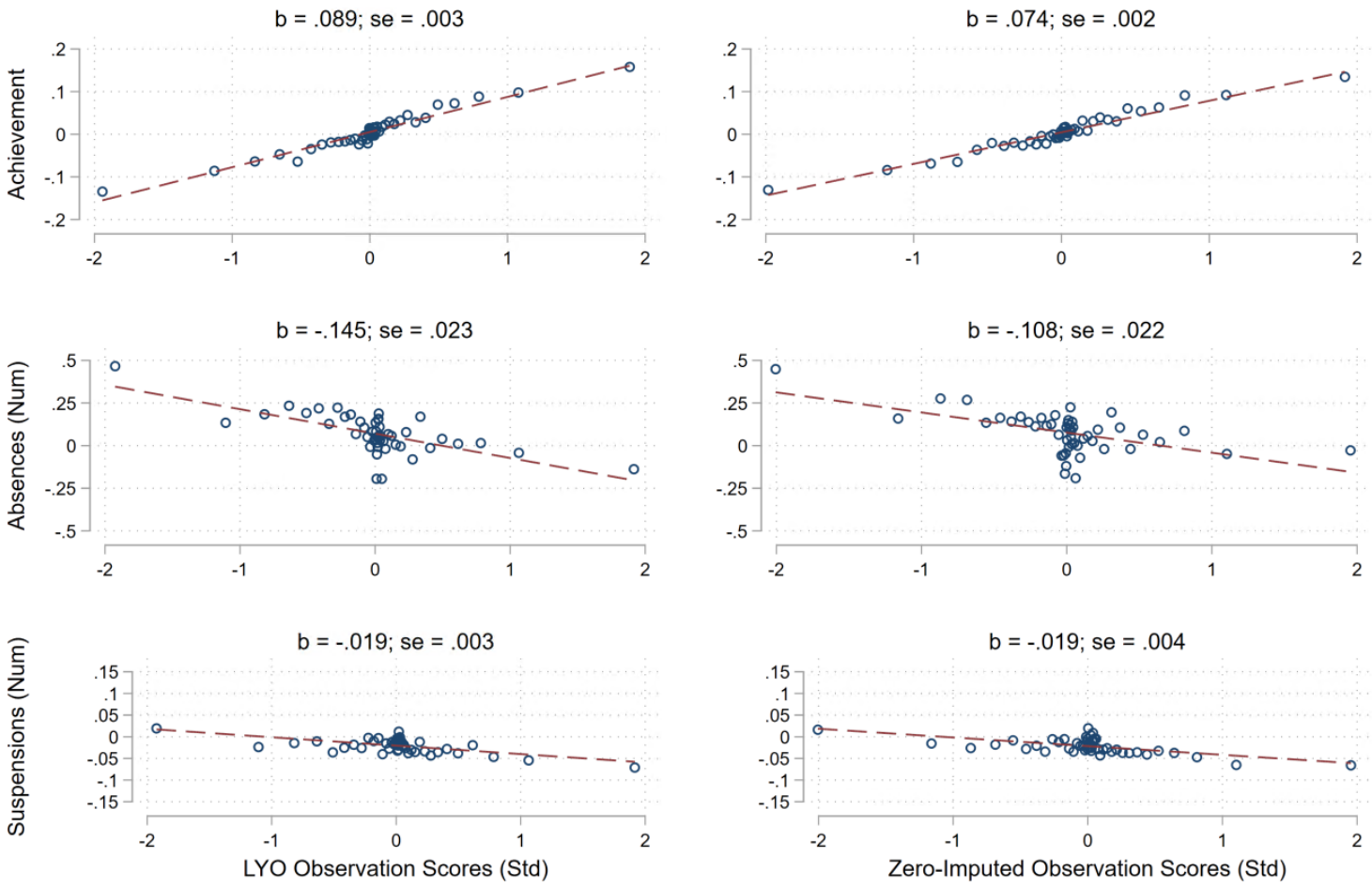
However, there are several reasons to be skeptical of this interpretation. First, post-secondary attendance and completion is an important mechanism through which exposure to K-12 teacher quality affects students' adult wages. In the "full" sample results for observation scores, significant, positive effects of observation score changes on post-secondary attendance, completion, and wages lend support to this hypothesis on how K-12 teacher quality "flows" through to affect adult wages. If teacher VA affected adult wages through similar channels, we would expect the larger VA effect on wages to be accompanied by larger effects on both post-secondary enrollment and completion. However, teacher VA shows relatively muted effects on post-secondary outcomes, with slightly smaller, but significant effects on post-secondary enrollments (Table 5.21 and no significant effects on post-secondary awards (Table 5.22. This does not preclude students' exposure to high VA teachers from having effects on wages through channels beyond post-secondary participation; Arcidiacono, Bayer, and Hizmo (2010), among others, find that student test performance, interpreted as a proxy for underlying ability, is predictive of wages even conditional on college completion. That said, the lack of pronounced teacher VA effects on intermediate post-secondary

outcomes calls into question the validity of the more distal wage effects I estimate.

Another issue complicating the interpretation of the estimated teacher VA effects on wages is the imprecision of the estimated effects across specifications. Previously, I present only estimates from the teacher switching design for various multiple measure analyses as I consider this design my preferred specification for most outcomes. However, as shown in tables for "full sample" analysis of observation scores, there is variation in the point estimates across the full suite of specifications. This is particularly true for more distal outcomes, such as wages, where I have far fewer cells when using the teacher switching design. From the full sample analysis of observation score effects of age 25 wages (Table 5.16, Panel C), I estimate point estimates ranging from $367.0 to $1020.8, a nearly threefold difference between the largest and smallest estimates. This range of estimates is even more severe when I conduct a "full sample" analysis on age 25 wages of changes to teacher VA (see Table 5.24), where the largest point estimate ($1397.6) obtained using the teacher switching design is nearly ten times the size of the smallest ($142.6) estimated when using OLS.

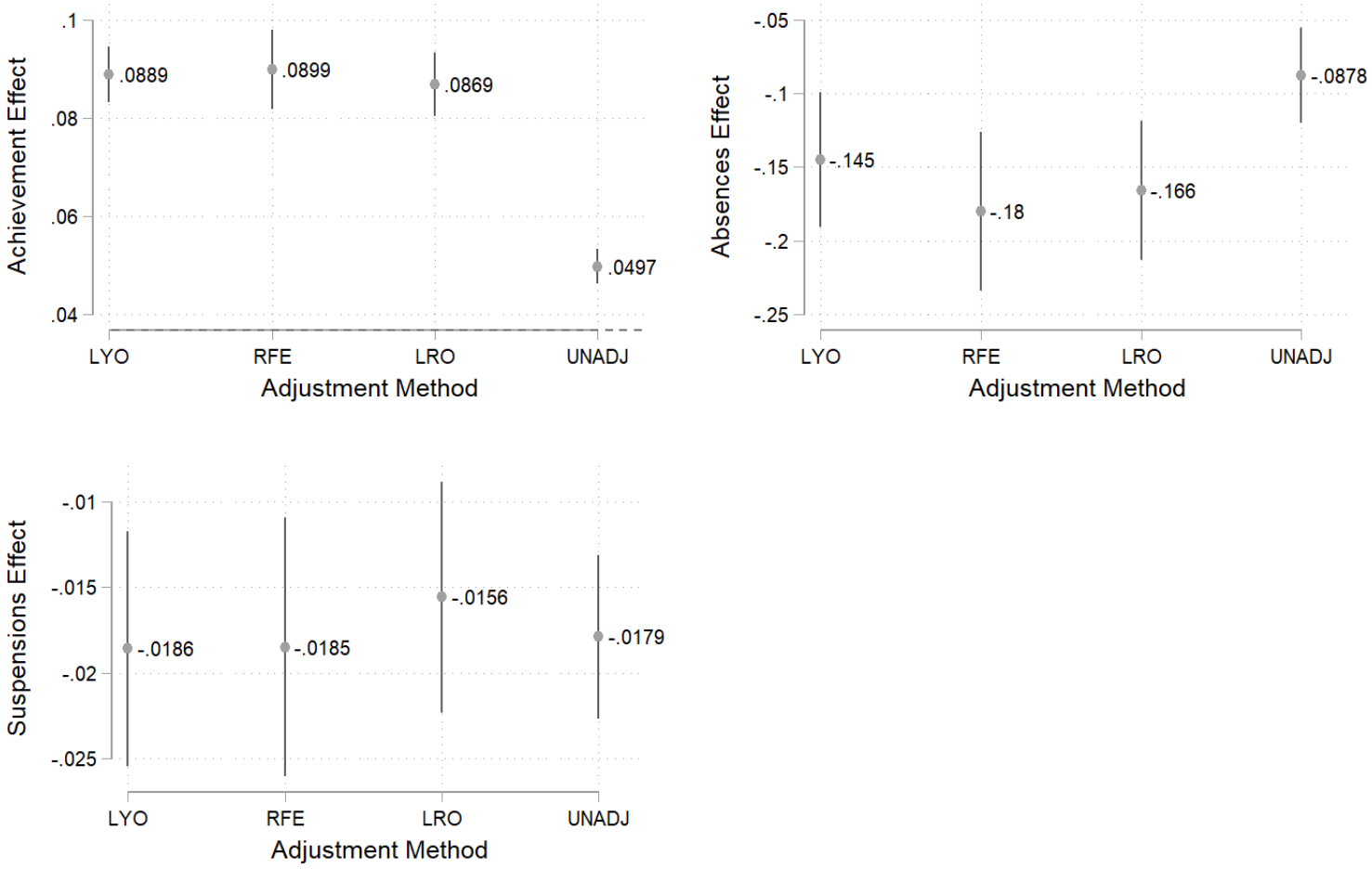
In their analysis of the effects of teacher VA on wages, Chetty et al. (2014b) note similar issues with the heavily inflated estimates coming from the teacher switching specification when applied to the smaller sample of department-year cells with observable wage outcomes, referring to their estimates as "very imprecise and fragile" (p.25). Given the evidence at hand, this appears to be a fitting description of the results I obtain when analyzing wage effects as well.

Figure 5.1: LYO and Zero-Imputed K-12 Effects



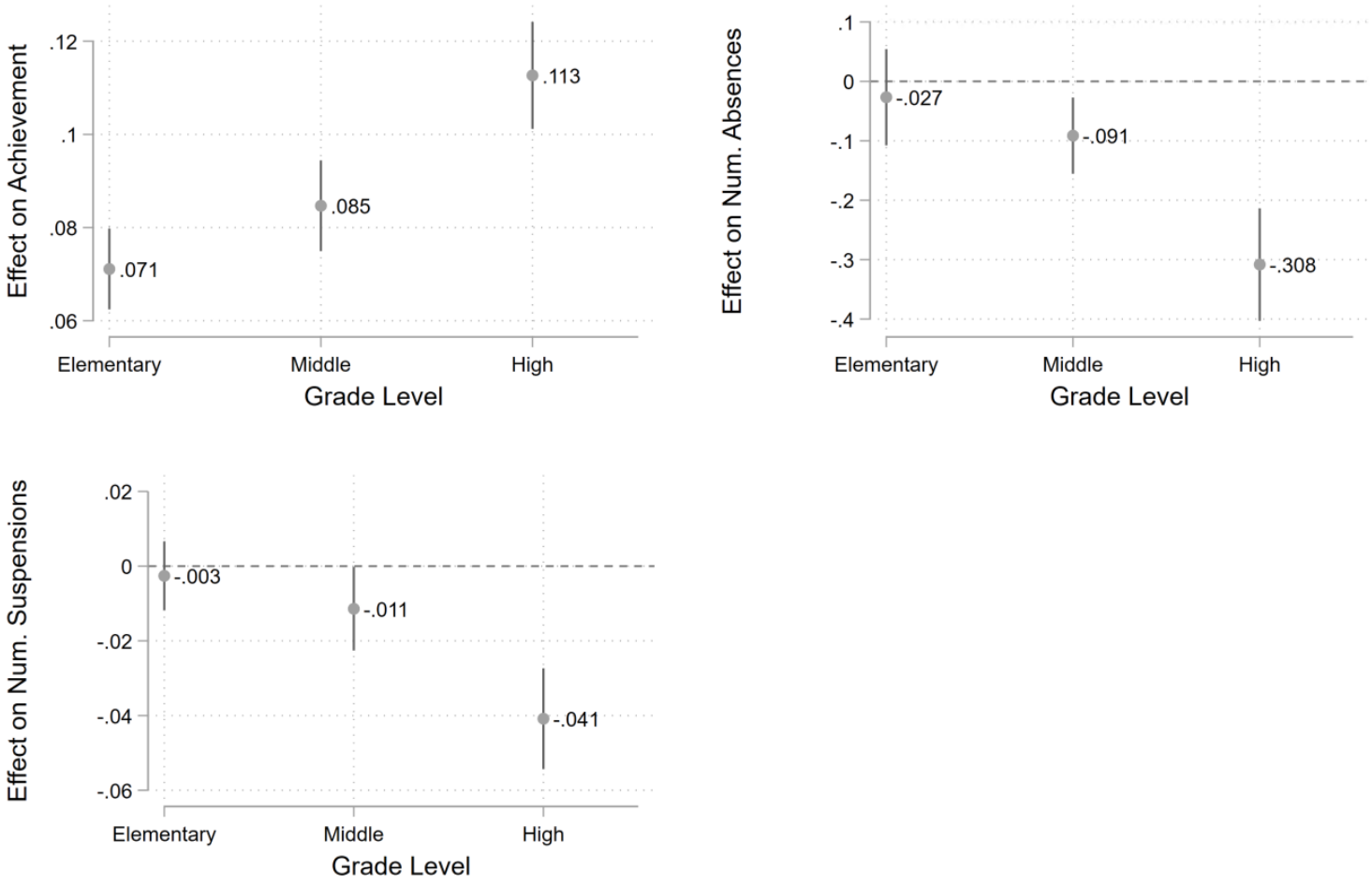
Note: Figure shows Leave Year Out (LYO) and zero-imputed observation score effects on standardized student achievement, the number of student absences, and the number of student suspensions. Point estimates and standard errors are shown above each plot, with the red dotted line showing the implied regression line and blue dots representing clustered bins of the data. Estimates are obtained using the "standard" teacher switching specification that does not include controls beyond separate dummy variables for year and grade level (i.e., elementary, middle, high) and use standard errors clustered at the school-cohort level.

Figure 5.2: K-12 Observation Score Effects by Adjustment Method



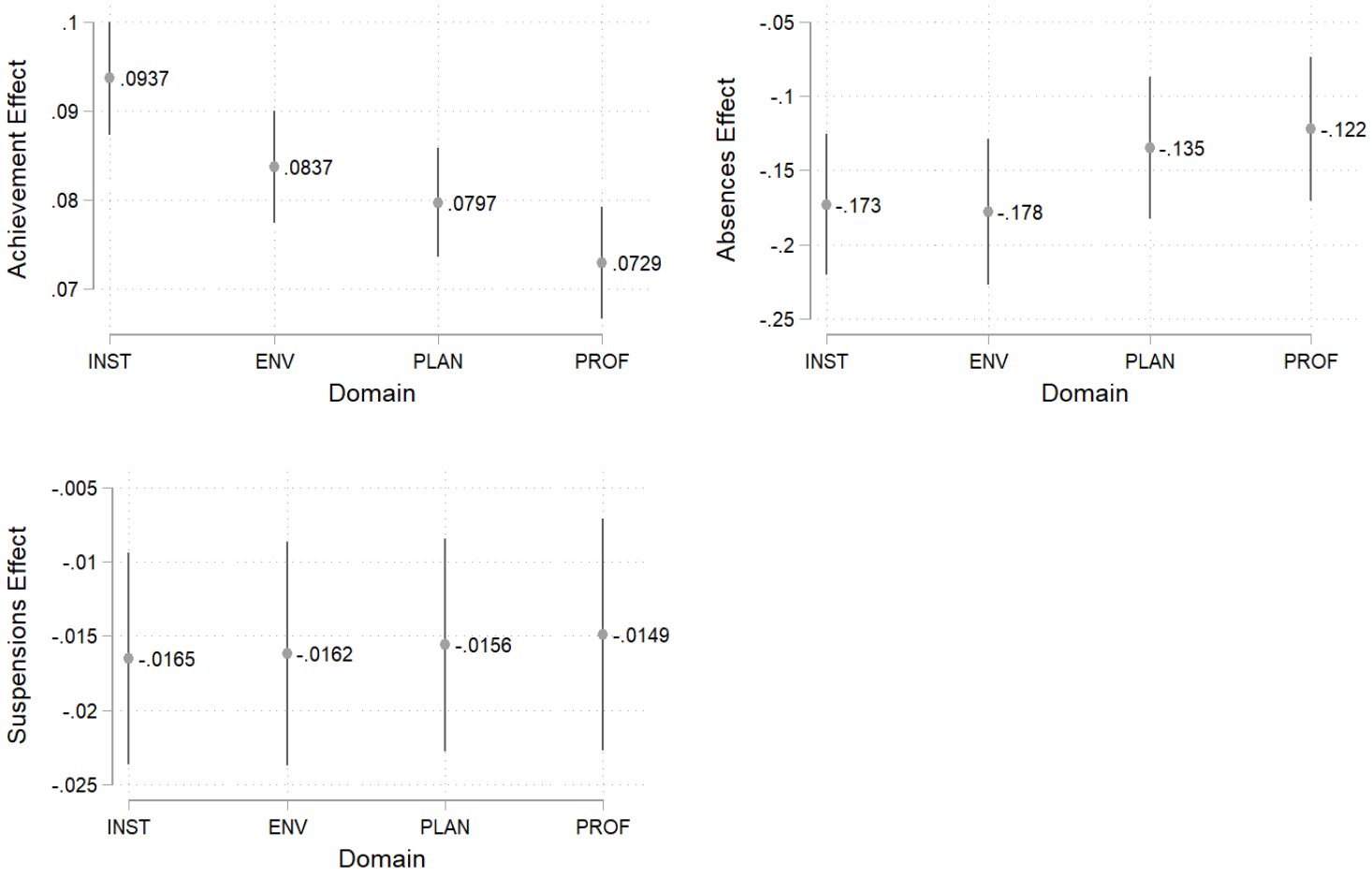
Note: Figure shows Leave Year Out (LYO), Rater Fixed Effect Adjusted (RFE), Leave Rater Out (LRO), and unadjusted observation score effects and 95 percent confidence intervals on standardized student achievement, the number of student absences, and the number of student suspensions. Estimates are obtained using the "standard" teacher switching specification that does not include controls beyond separate dummy variables for year and grade level (i.e., elementary, middle, high) and use standard errors clustered at the school-cohort level

Figure 5.3: LYO Effects on K-12 Outcomes, by Grade Level



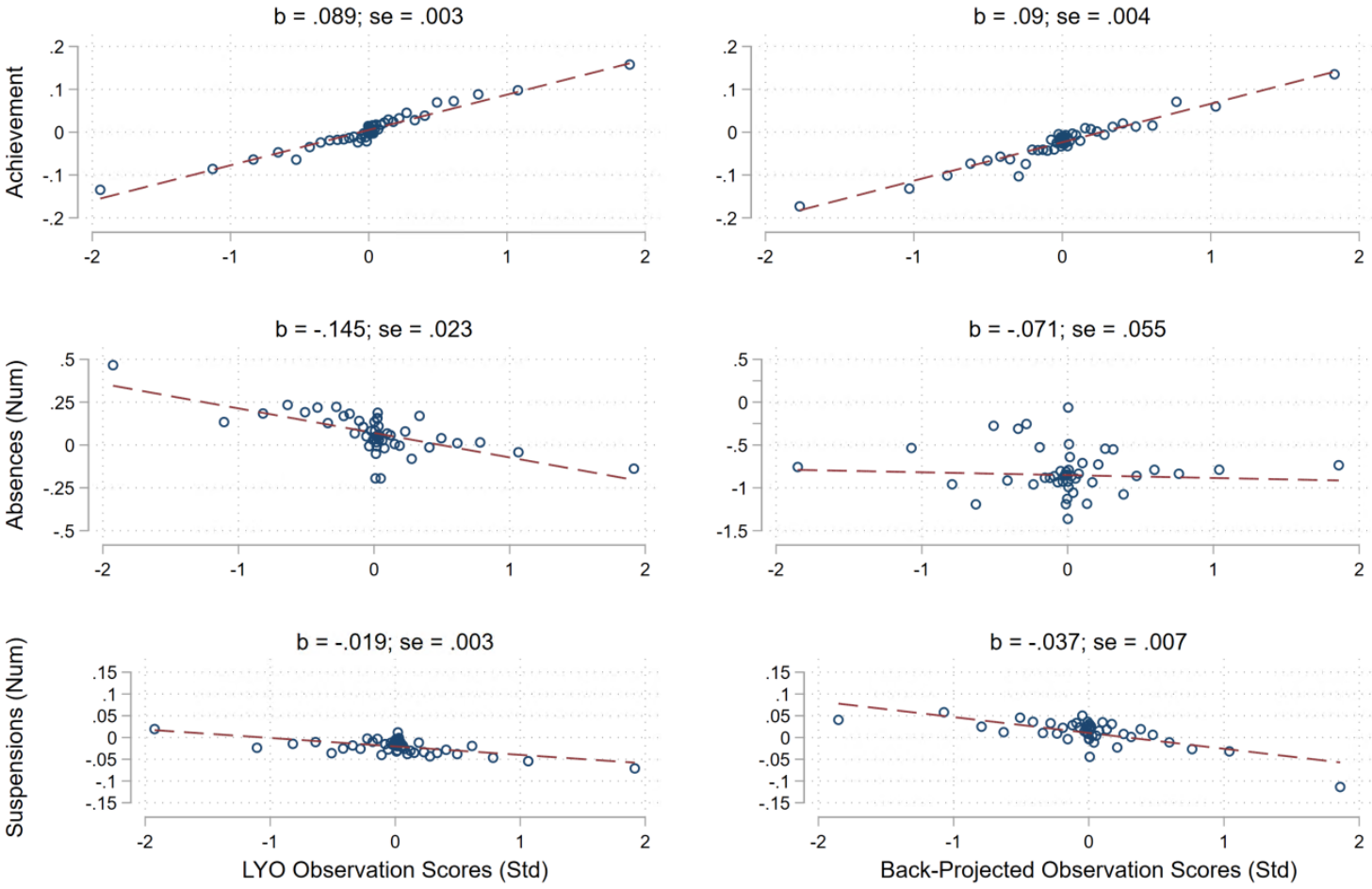
Note: Figure shows effects of changes to LYO estimates on K-12 student outcomes using the "standard" teacher switching specification, by grade level. Vertical lines depict 95 percent confidence intervals.

Figure 5.4: K-12 Observation Score Effects by Domain



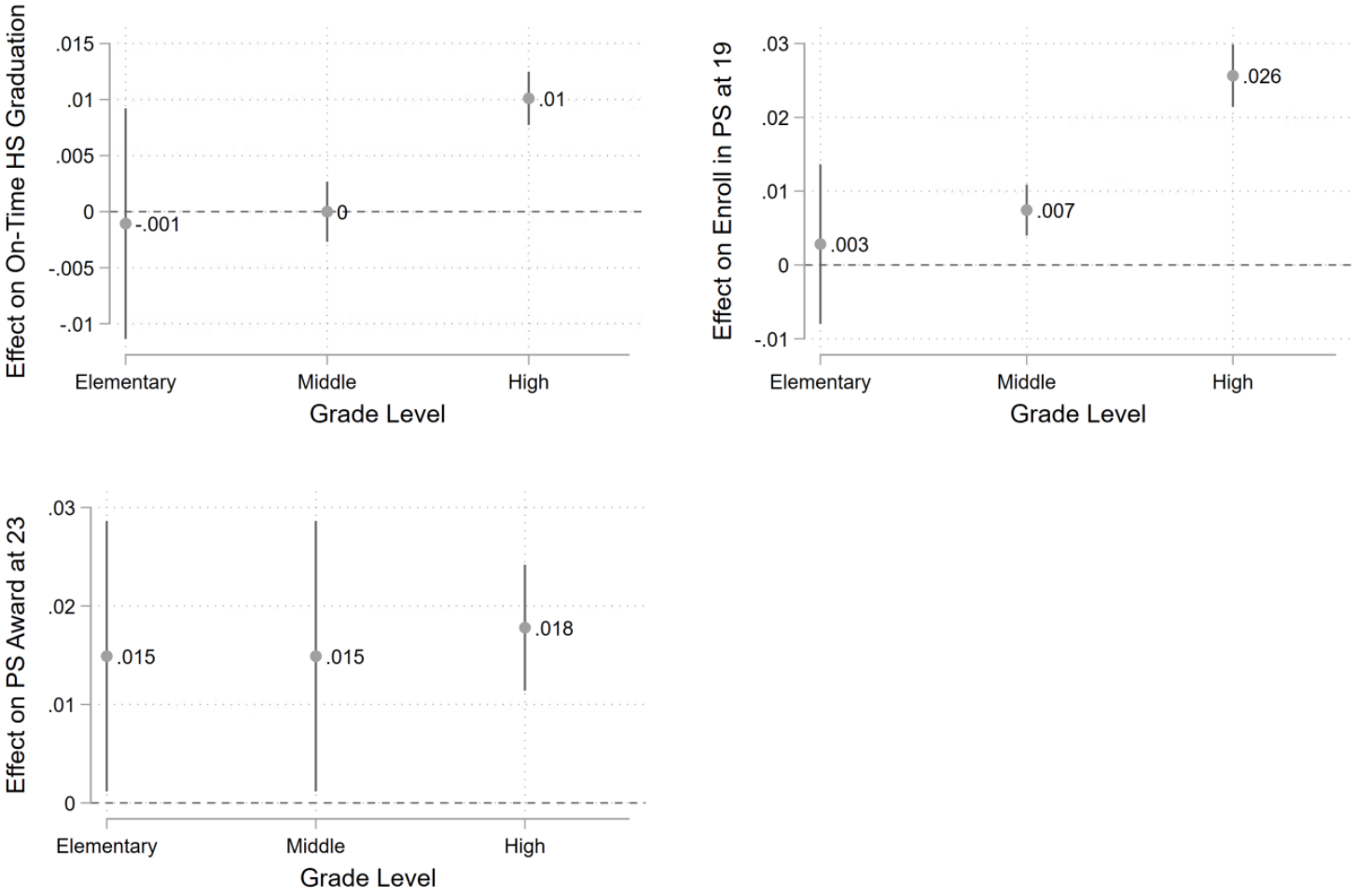
Note: Figure shows Instruction (INST), Environment (ENV), Planning (PLAN), and Professionalism (PROF) effects and 95 percent confidence intervals on standardized student achievement, the number of student absences, and the number of student suspensions. Estimates are obtained using the "standard" teacher switching specification that does not include controls beyond separate dummy variables for year and grade level (i.e., elementary, middle, high) and use standard errors clustered at the school-cohort level

Figure 5.5: LYO and Back-Projected K-12 Effects



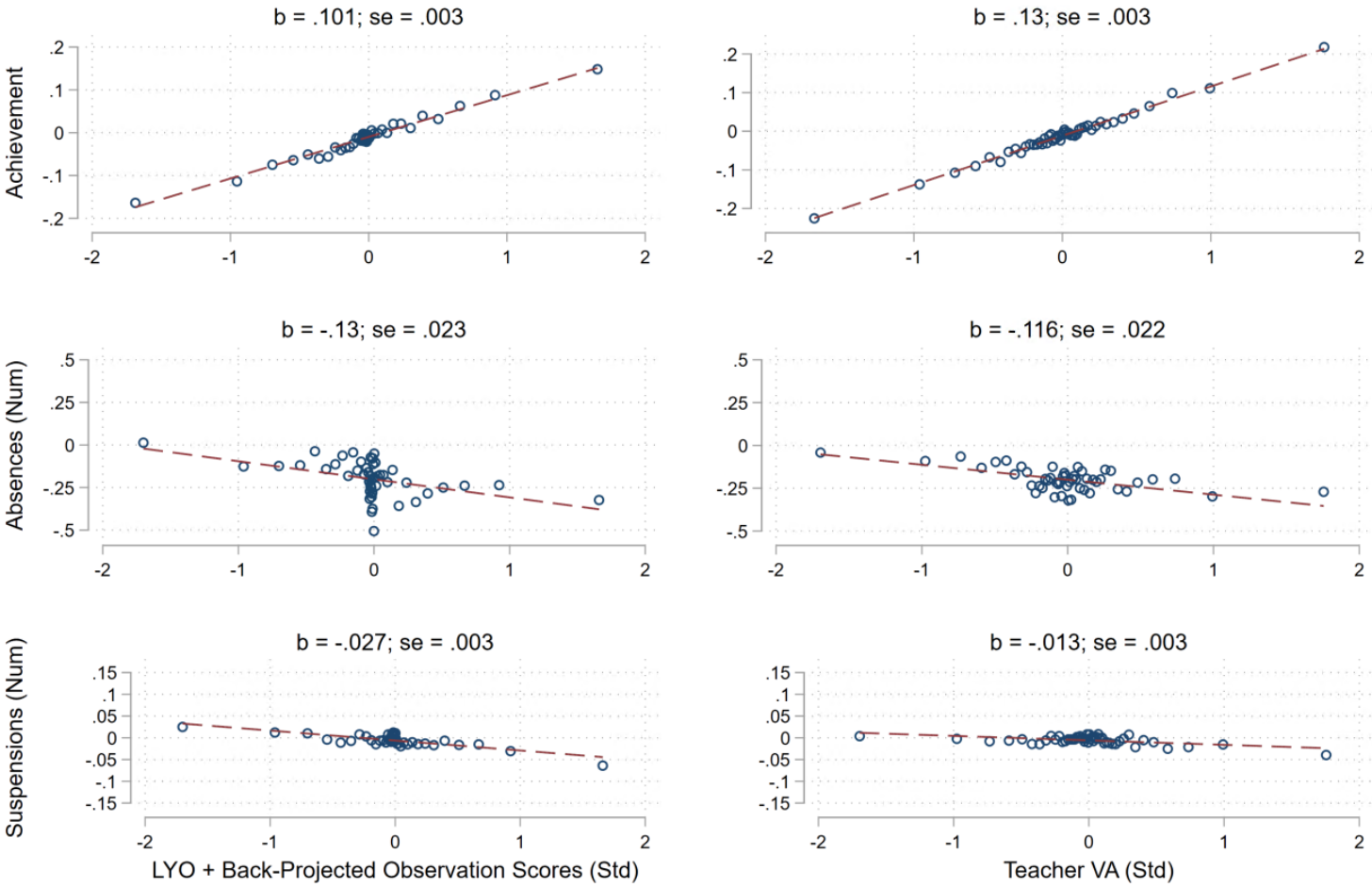
Note: Figure shows Leave Year Out (LYO) and Back-Projected observation score effects on standardized student achievement, the number of student absences, and the number of student suspensions. Point estimates and standard errors are shown above each plot, with the red dotted line showing the implied regression line and blue dots representing clustered bins of the data. The LYO sample uses student and teacher data from 2011-12 to 2017-18 whereas the Back-Projected data uses teacher data from 2011-12 to 2017-18 to estimate teacher effectiveness for student data from 2007-08 to 2010-11. Estimates are obtained using the "standard" teacher switching specification that does not include controls beyond separate dummy variables for year and grade level (i.e., elementary, middle, high) and use standard errors clustered at the school-cohort level.

Figure 5.6: LYO Effects on Long-Run Outcomes, by Grade Level



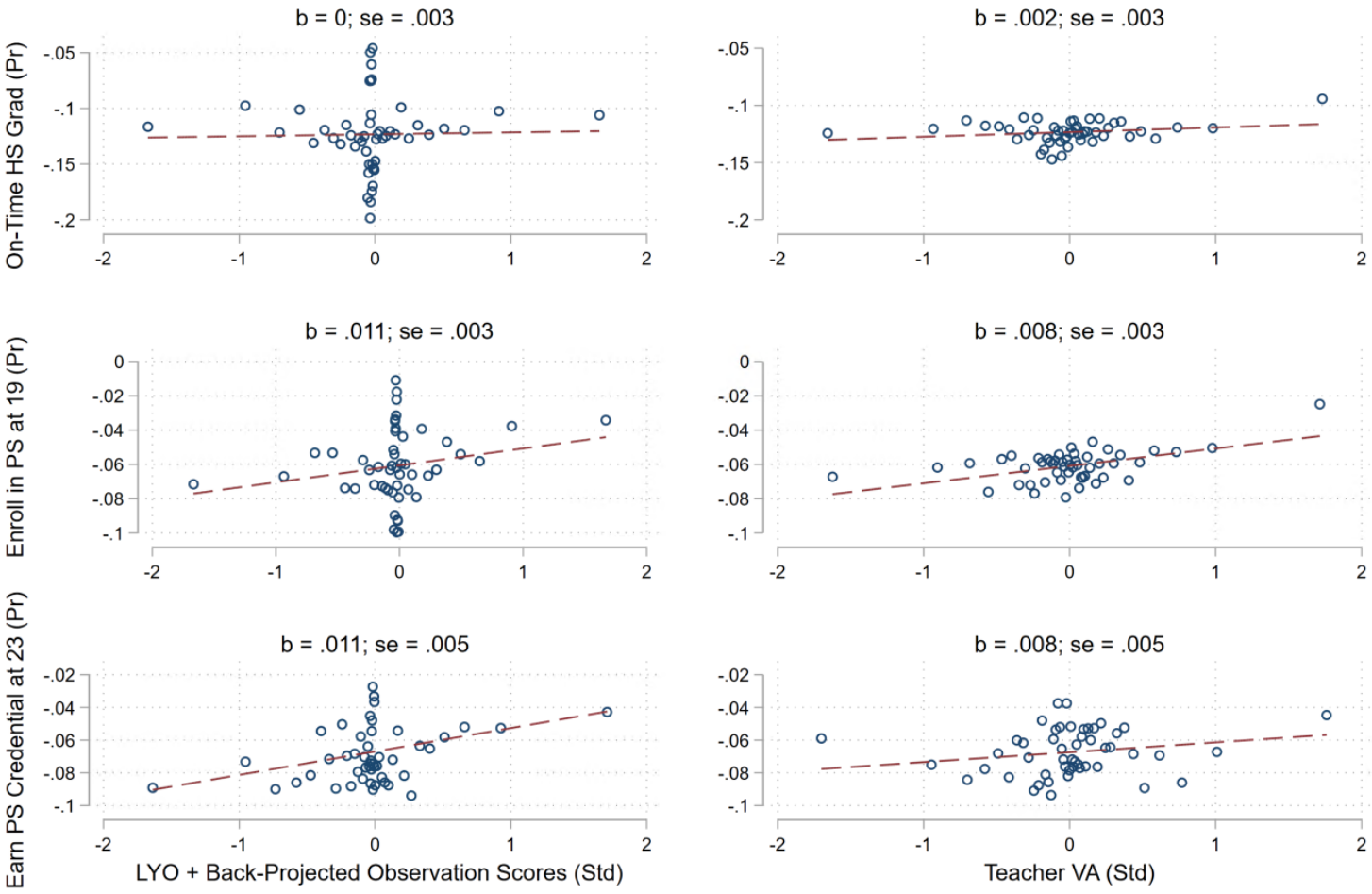
Note: Figure shows effects of changes to LYO estimates on on-time high school graduation, any post-secondary enrollment at age 19, and any post-secondary award or degree at age 23 using the "standard" teacher switching specification, by grade level. Vertical lines depict 95 percent confidence intervals.

Figure 5.7: Observation and VA Effects on K-12 Outcomes, Matched Sample



Note: Figure shows Leave Year Out (LYO) and Teacher VA effects on standardized student achievement, the number of student absences, and the number of student suspensions. Point estimates and standard errors are shown above each plot, with the red dotted line showing the implied regression line and blue dots representing clustered bins of the data. Estimates are obtained using the "standard" teacher switching specification that does not include controls beyond separate dummy variables for year and grade level (i.e., elementary, middle, high) and use standard errors clustered at the school-cohort level.

Figure 5.8: Observation and VA Effects on Long-Run Outcomes, Matched Sample



Note: Figure shows Leave Year Out (LYO) and Teacher VA effects on on-time high school graduation, any post-secondary enrollment at age 19, and any post-secondary credential at age 23. Point estimates and standard errors are shown above each plot, with the red dotted line showing the implied regression line and blue dots representing clustered bins of the data. Estimates are obtained using the "standard" teacher switching specification that does not include controls beyond separate dummy variables for year and grade level (i.e., elementary, middle, high) and use standard errors clustered at the school-cohort level.

Table 5.1: LYO Observation Score Effects on Achievement

	(1)	(2)	(3)	(4)	(5)	(6)
Change in Obs Scores (LYO)	0.208***	0.0961***	0.0889***	0.0811***	0.0861***	0.0763***
	(0.00262)	(0.000974)	(0.00293)	(0.00263)	(0.00290)	(0.00253)
N	6,468,387	6,468,387	60,213	43,482	60,213	43,482
R2	0.044	0.570	0.035	0.298	0.272	0.521
Design	OLS	OLS	QE	QE	QE	QE
Controls		X		X		X
School-Year FE					X	X

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors in parentheses.

Note: Table shows leave-year-out (LYO) observation score effects on standardized student achievement. Observations are "stacked" across subjects for additional precision, meaning the unit of analyses are student-subject-year observations in the OLS model and department-subject-year observations in the QE model. "Controls" refer to a vector of observable student characteristics and include lagged outcomes, student race/ethnicity, gender, FRPL status, ELL status, and SPED status. For the QE specification, all student outcome, teacher quality, and control variables are operationalized as first differences (t - t-1) and results are weighted by the number of students in each cell. All models include separate dummy variables for subject, year, and grade level (i.e., elementary, middle, high) and use standard errors clustered at the school-cohort level.

Table 5.2: LYO Observation Score Effects on Absences

	(1)	(2)	(3)	(4)	(5)	(6)
Change in Obs Scores (LYO)	–0.395***	–0.133***	–0.145***	–0.129***	–0.123***	–0.0586***
	(0.0160)	(0.00962)	(0.0233)	(0.0230)	(0.0184)	(0.0155)
N	9,053,002	9,053,002	48,828	48,828	48,828	48,828
R2	0.017	0.379	0.031	0.046	0.631	0.719
Design	OLS	OLS	QE	QE	QE	QE
Controls		X		X		X
School-Year FE					X	X

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors in parentheses.

Note: Table shows leave-year-out (LYO) observation score effects on the number of student absences. The unit of analyses are student-year observations in the OLS model and department-year observations in the QE model. "Controls" refer to a vector of observable student characteristics and include lagged outcomes, student race/ethnicity, gender, FRPL status, ELL status, and SPED status. For the QE specification, all student outcome, teacher quality, and control variables are operationalized as first differences (t - t-1) and results are weighted by the number of students in each cell. All models include separate dummy variables for year and grade level (i.e., elementary, middle, high) and use standard errors clustered at the school-cohort level.

Table 5.3: LYO Observation Score Effects on Suspensions

	(1)	(2)	(3)	(4)	(5)	(6)
Change in Obs Scores (LYO)	–0.0772***	–0.0418***	–0.0186***	–0.0149***	–0.0207***	–0.0151***
	(0.00262)	(0.00203)	(0.00350)	(0.00342)	(0.00279)	(0.00263)
N	9,053,261	9,053,261	48,835	48,833	48,835	48,833
R2	0.022	0.087	0.010	0.042	0.559	0.589
Design	OLS	OLS	QE	QE	QE	QE
Controls		X		X		X
School-Year FE					X	X

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors in parentheses.

Note: Table shows leave-year-out (LYO) observation score effects on the number of student absences. The unit of analyses are student-year observations in the OLS model and department-year observations in the QE model. "Controls" refer to a vector of observable student characteristics and include lagged outcomes, student race/ethnicity, gender, FRPL status, ELL status, and SPED status. For the QE specification, all student outcome, teacher quality, and control variables are operationalized as first differences (t - t-1) and results are weighted by the number of students in each cell. All models include separate dummy variables for year and grade level (i.e., elementary, middle, high) and use standard errors clustered at the school-cohort level.

Table 5.4: Zero-Imputed LYO Effects on K-12 Outcomes

Panel A: Student Test Scores

	(1)	(2)	(3)	(4)	(5)	(6)
Change in Obs Scores (Imputed)	0.204***	0.0932***	0.0744***	0.0722***	0.0706***	0.0675***
	(0.00252)	(0.000914)	(0.00246)	(0.00239)	(0.00234)	(0.00226)
N	7,522,581	7,522,581	66,940	48,332	66,940	48,332
R2	0.042	0.570	0.028	0.238	0.275	0.475
Design	OLS	OLS	QE	QE	QE	QE
Controls		X		X		X
School-Year FE					X	X

Panel B: Student Absences

	(1)	(2)	(3)	(4)	(5)	(6)
Change in Obs Scores (Imputed)	–0.392***	–0.133***	–0.108***	–0.101***	–0.0937***	–0.0536***
	(0.0151)	(0.00898)	(0.0217)	(0.0217)	(0.0157)	(0.0137)
N	10,457,551	10,457,551	53,461	53,461	53,461	53,461
R2	0.018	0.377	0.028	0.039	0.648	0.728
Design	OLS	OLS	QE	QE	QE	QE
Controls		X		X		X
School-Year FE					X	X

Panel C: Student Suspensions

	(1)	(2)	(3)	(4)	(5)	(6)
Change in Obs Scores (Imputed)	–0.0797***	–0.0432***	–0.0187***	–0.0168***	–0.0165***	–0.0133***
	(0.00264)	(0.00203)	(0.00350)	(0.00344)	(0.00253)	(0.00242)
N	10,457,863	10,457,863	53,463	53,461	53,463	53,461
R2	0.022	0.088	0.010	0.036	0.593	0.615
Design	OLS	OLS	QE	QE	QE	QE
Controls		X		X		X
School-Year FE					X	X

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors in parentheses.

Note: Table shows zero-imputed Leave Year Out (LYO) observation score effects on standardized student achievement (A), the number of student absences (B), and the number of student suspensions (C). The unit of analyses are student-year observations in the OLS model and department-year observations in the QE model. "Controls" refer to a vector of observable student characteristics and include lagged outcomes, student race/ethnicity, gender, FRPL status, ELL status, and SPED status. For the QE specification, all student outcome, teacher quality, and control variables are operationalized as first differences (t - t-1) and results are weighted by the number of students in each cell. All models include separate dummy variables for year and grade level (i.e., elementary, middle, high) and use standard errors clustered at the school-cohort level.

Table 5.5: Rater Fixed Effect Adjusted (RFE) Effects on K-12 Outcomes

Panel A: Student Test Scores

	(1)	(2)	(3)	(4)	(5)	(6)
Change in Obs Scores (RFE)	0.167***	0.0841***	0.0899***	0.0755***	0.0902***	0.0736***
	(0.00269)	(0.00110)	(0.00416)	(0.00332)	(0.00406)	(0.00314)
N	4,547,349	4,547,349	39,311	30,267	39,311	30,267
R2	0.030	0.561	0.034	0.377	0.279	0.577
Design	OLS	OLS	QE	QE	QE	QE
Controls		X		X		X
School-Year FE					X	X

Panel B: Student Absences

	(1)	(2)	(3)	(4)	(5)	(6)
Change in Obs Scores (RFE)	–0.298***	–0.114***	–0.180***	–0.149***	–0.166***	–0.0947***
	(0.0170)	(0.0104)	(0.0276)	(0.0268)	(0.0235)	(0.0188)
N	6,045,958	6,045,958	34,174	34,174	34,174	34,174
R2	0.015	0.387	0.042	0.060	0.634	0.727
Design	OLS	OLS	QE	QE	QE	QE
Controls		X		X		X
School-Year FE					X	X

Panel C: Student Suspensions

	(1)	(2)	(3)	(4)	(5)	(6)
Change in Obs Scores (RFE)	–0.0581***	–0.0338***	–0.0185***	–0.0129***	–0.0215***	–0.0138***
	(0.00274)	(0.00216)	(0.00384)	(0.00377)	(0.00310)	(0.00296)
N	6,046,128	6,046,128	34,176	34,175	34,176	34,175
R2	0.015	0.077	0.011	0.044	0.536	0.570
Design	OLS	OLS	QE	QE	QE	QE
Controls		X		X		X
School-Year FE					X	X

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors in parentheses.

Note: Table shows Rater Fixed Effect adjusted (RFE) observation score effects on standardized student achievement (A), the number of student absences (B), and the number of student suspensions (C). The unit of analyses are student-year observations in the OLS model and department-year observations in the QE model. "Controls" refer to a vector of observable student characteristics and include lagged outcomes, student race/ethnicity, gender, FRPL status, ELL status, and SPED status. For the QE specification, all student outcome, teacher quality, and control variables are operationalized as first differences (t - t-1) and results are weighted by the number of students in each cell. All models include separate dummy variables for year and grade level (i.e., elementary, middle, high) and use standard errors clustered at the school-cohort level.

Table 5.6: Leave-Rater-Out (LRO) Effects on K-12 Outcomes

Panel A: Student Test Scores

	(1)	(2)	(3)	(4)	(5)	(6)
Change in Obs Scores (LRO)	0.187***	0.0880***	0.0869***	0.0734***	0.0848***	0.0705***
	(0.00257)	(0.000996)	(0.00328)	(0.00283)	(0.00320)	(0.00277)
N	5326054	5326054	48835	35128	48835	35128
R2	0.0405	0.565	0.0358	0.330	0.274	0.542
Design	OLS	OLS	QE	QE	QE	QE
Controls		X		X		X
School-Year FE					X	X

Panel B: Student Absences

	(1)	(2)	(3)	(4)	(5)	(6)
Change in Obs Scores (LRO)	–0.340***	–0.119***	–0.166***	–0.144***	–0.142***	–0.0773***
	(0.0154)	(0.00929)	(0.0242)	(0.0239)	(0.0201)	(0.0167)
N	7360163	7360163	39660	39660	39660	39660
R2	0.0171	0.386	0.0424	0.0568	0.633	0.724
Design	OLS	OLS	QE	QE	QE	QE
Controls		X		X		X
School-Year FE					X	X

Panel C: Student Suspensions

	(1)	(2)	(3)	(4)	(5)	(6)
Change in Obs Scores (LRO)	–0.0676***	–0.0370***	–0.0156***	–0.0114***	–0.0178***	–0.0119***
	(0.00272)	(0.00204)	(0.00343)	(0.00337)	(0.00283)	(0.00267)
N	7360359	7360359	39661	39660	39661	39660
R2	0.0190	0.0785	0.0108	0.0415	0.543	0.574
Design	OLS	OLS	QE	QE	QE	QE
Controls		X		X		X
School-Year FE					X	X

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors in parentheses.

Note: Table shows Leave-Rater-Out (LRO) observation score effects on standardized student achievement (A), the number of student absences (B), and the number of student suspensions (C). The unit of analyses are student-year observations in the OLS model and department-year observations in the QE model. "Controls" refer to a vector of observable student characteristics and include lagged outcomes, student race/ethnicity, gender, FRPL status, ELL status, and SPED status. For the QE specification, all student outcome, teacher quality, and control variables are operationalized as first differences (t - t-1) and results are weighted by the number of students in each cell. All models include separate dummy variables for year and grade level (i.e., elementary, middle, high) and use standard errors clustered at the school-cohort level.

Table 5.7: Unadjusted Observation Score Effects on K-12 Outcomes

Panel A: Student Test Scores

	(1)	(2)	(3)	(4)	(5)	(6)
Change in Obs Scores (Unadjusted)	0.192***	0.0862***	0.0497***	0.0426***	0.0619***	0.0525***
	(0.00256)	(0.000942)	(0.00179)	(0.00181)	(0.00187)	(0.00184)
N	7,176,303	7,176,303	63,822	45,825	63,822	45,825
R2	0.040	0.569	0.027	0.240	0.271	0.479
Design	OLS	OLS	QE	QE	QE	QE
Controls		X		X		X
School-Year FE					X	X

Panel B: Student Absences

	(1)	(2)	(3)	(4)	(5)	(6)
Change in Obs Scores (Unadjusted)	–0.372***	–0.130***	–0.0878***	–0.0808***	–0.0991***	–0.0595***
	(0.0147)	(0.00893)	(0.0165)	(0.0165)	(0.0125)	(0.0108)
N	10004927	10004927	50,918	50,918	50,918	50,918
R2	0.018	0.381	0.030	0.043	0.636	0.720
Design	OLS	OLS	QE	QE	QE	QE
Controls		X		X		X
School-Year FE					X	X

Panel C: Student Suspensions

	(1)	(2)	(3)	(4)	(5)	(6)
Change in Obs Scores (Unadjusted)	–0.0812***	–0.0467***	–0.0179***	–0.0156***	–0.0193***	–0.0153***
	(0.00280)	(0.00217)	(0.00244)	(0.00240)	(0.00188)	(0.00181)
N	10005197	10005197	50,924	50,922	50,924	50,922
R2	0.023	0.090	0.012	0.040	0.581	0.605
Design	OLS	OLS	QE	QE	QE	QE
Controls		X		X		X
School-Year FE					X	X

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors in parentheses.

Note: Table shows unadjusted observation score effects on student test z-scores (Panel A), the number of student absences (Panel B), and the number of student suspensions (Panel C). Observation scores are operationalized as annual averages from year t and do not incorporate shrinkage, regression adjustment, or data beyond year t. The unit of analyses are student-year observations in the OLS model and department-year observations in the QE model. "Controls" refer to a vector of observable student characteristics and include lagged outcomes, student race/ethnicity, gender, FRPL status, ELL status, and SPED status. For the QE specification, all student outcome, teacher quality, and control variables are operationalized as first differences (t - t-1) and results are weighted by the number of students in each cell. All models include separate dummy variables for year and grade level (i.e., elementary, middle, high) and use standard errors clustered at the school-cohort level.

Table 5.8: Correlations Across Domain Scores

	INST	ENV	PLAN	PROF
INST	1			
ENV	0.848	1		
PLAN	0.843	0.733	1	
PROF	0.712	0.681	0.659	1

Note: Table shows pairwise correlations between teachers' annual scores on the Instruction (INST), Environment (ENV), Planning (PLAN), and Professionalism (PROF) domains of the TEAM observation rubric. Teachers' domain-specific observation scores are standardized within year to have mean zero and unit variance.

Table 5.9: Domain-Specific Effects on Achievement

Panel A: Instruction

	(1)	(2)	(3)	(4)	(5)	(6)
Change in Obs Scores (INST)	0.201***	0.0959***	0.0937***	0.0855***	0.0920***	0.0830***
	(0.00262)	(0.00101)	(0.00323)	(0.00296)	(0.00312)	(0.00279)
N	5,685,693	5,685,693	52,552	37,886	52,552	37,886
R2	0.043	0.565	0.038	0.298	0.273	0.521
Design	OLS	OLS	QE	QE	QE	QE
Controls		X		X		X
School-Year FE					X	X

Panel B: Environment

	(1)	(2)	(3)	(4)	(5)	(6)
Change in Obs Scores (ENV)	0.199***	0.0923***	0.0837***	0.0767***	0.0819***	0.0743***
	(0.00267)	(0.00102)	(0.00320)	(0.00303)	(0.00308)	(0.00284)
N	5,668,279	5,668,279	52,460	37,820	52,460	37,820
R2	0.040	0.564	0.031	0.295	0.268	0.517
Design	OLS	OLS	QE	QE	QE	QE
Controls		X		X		X
School-Year FE					X	X

Panel C: Planning

	(1)	(2)	(3)	(4)	(5)	(6)
Change in Obs Scores (PLAN)	0.168***	0.0779***	0.0797***	0.0720***	0.0777***	0.0682***
	(0.00262)	(0.000995)	(0.00311)	(0.00284)	(0.00303)	(0.00272)
N	5,676,702	5,676,702	52,484	37,830	52,484	37,830
R2	0.033	0.563	0.029	0.290	0.266	0.514
Design	OLS	OLS	QE	QE	QE	QE
Controls		X		X		X
School-Year FE					X	X

Panel D: Professionalism

	(1)	(2)	(3)	(4)	(5)	(6)
Change in Obs Scores (PROF)	0.155***	0.0726***	0.0729***	0.0658***	0.0701***	0.0629***
	(0.00268)	(0.00103)	(0.00321)	(0.00294)	(0.00315)	(0.00281)
N	5,631,102	5,631,102	52,211	37,661	52,211	37,661
R2	0.027	0.562	0.025	0.291	0.262	0.514
Design	OLS	OLS	QE	QE	QE	QE
Controls		X		X		X
School-Year FE					X	X

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors in parentheses.

Note: Table shows leave-year-out (LYO) domain-specific observation score effects on student achievement. The unit of analyses are student-year observations in the OLS model and department-year observations in the QE model. "Controls" refer to a vector of observable student characteristics and include lagged outcomes, student race/ethnicity, gender, FRPL status, ELL status, and SPED status. For the QE specification, all student outcome, teacher quality, and control variables are operationalized as first differences (t - t-1) and results are weighted by the number of students in each cell. All models include separate dummy variables for year and grade level (i.e., elementary, middle, high) and use standard errors clustered at the school-cohort level.

Table 5.10: Domain-Specific Effects on Absences

Panel A: Instruction

	(1)	(2)	(3)	(4)	(5)	(6)
Change in Obs Scores (INST)	−0.358***	−0.128***	−0.173***	−0.160***	−0.134***	−0.0729***
	(0.0158)	(0.00940)	(0.0244)	(0.0242)	(0.0197)	(0.0167)
N	7,920,863	7,920,863	42,054	42,054	42,054	42,054
R2	0.017	0.386	0.042	0.054	0.637	0.724
Design	OLS	OLS	QE	QE	QE	QE
Controls		X		X		X
School-Year FE					X	X

Panel B: Environment

	(1)	(2)	(3)	(4)	(5)	(6)
Change in Obs Scores (ENV)	−0.397***	−0.154***	−0.178***	−0.166***	−0.146***	−0.0832***
	(0.0166)	(0.0101)	(0.0251)	(0.0248)	(0.0200)	(0.0171)
N	7,900,930	7,900,930	41,995	41,995	41,995	41,995
R2	0.017	0.386	0.042	0.054	0.636	0.724
Design	OLS	OLS	QE	QE	QE	QE
Controls		X		X		X
School-Year FE					X	X

Panel C: Planning

	(1)	(2)	(3)	(4)	(5)	(6)
Change in Obs Scores (PLAN)	−0.306***	−0.122***	−0.135***	−0.121***	−0.0992***	−0.0464**
	(0.0156)	(0.00932)	(0.0244)	(0.0243)	(0.0193)	(0.0161)
N	7,909,844	7,909,844	42,000	42,000	42,000	42,000
R2	0.017	0.386	0.042	0.054	0.636	0.724
Design	OLS	OLS	QE	QE	QE	QE
Controls		X		X		X
School-Year FE					X	X

Panel D: Professionalism

	(1)	(2)	(3)	(4)	(5)	(6)
Change in Obs Scores (PROF)	−0.250***	−0.0800***	−0.122***	−0.108***	−0.115***	−0.0690***
	(0.0165)	(0.0100)	(0.0248)	(0.0245)	(0.0203)	(0.0174)
N	7,845,292	7,845,292	41,760	41,760	41,760	41,760
R2	0.016	0.386	0.042	0.054	0.636	0.724
Design	OLS	OLS	QE	QE	QE	QE
Controls		X		X		X
School-Year FE					X	X

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors in parentheses.

Note: Table shows leave-year-out (LYO) domain-specific observation score effects on the number of student absences. The unit of analyses are student-year observations in the OLS model and department-year observations in the QE model. "Controls" refer to a vector of observable student characteristics and include lagged outcomes, student race/ethnicity, gender, FRPL status, ELL status, and SPED status. For the QE specification, all student outcome, teacher quality, and control variables are operationalized as first differences (t - t-1) and results are weighted by the number of students in each cell. All models include separate dummy variables for year and grade level (i.e., elementary, middle, high) and use standard errors clustered at the school-cohort level.

Table 5.11: Domain-Specific Effects on Suspensions

Panel A: Instruction

	(1)	(2)	(3)	(4)	(5)	(6)
Change in Obs Scores (INST)	−0.0720***	−0.0405***	−0.0165***	−0.0130***	−0.0190***	−0.0132***
	(0.00273)	(0.00202)	(0.00365)	(0.00356)	(0.00300)	(0.00283)
N	7,921,066	7,921,066	42,057	42,056	42,057	42,056
R2	0.020	0.078	0.011	0.039	0.556	0.585
Design	OLS	OLS	QE	QE	QE	QE
Controls		X		X		X
School-Year FE					X	X

Panel B: Environment

	(1)	(2)	(3)	(4)	(5)	(6)
Change in Obs Scores (ENV)	−0.0810***	−0.0459***	−0.0162***	−0.0135***	−0.0206***	−0.0159***
	(0.00317)	(0.00236)	(0.00385)	(0.00380)	(0.00301)	(0.00287)
N	7,901,133	7,901,133	41,998	41,997	41,998	41,997
R2	0.021	0.079	0.011	0.039	0.556	0.584
Design	OLS	OLS	QE	QE	QE	QE
Controls		X		X		X
School-Year FE					X	X

Panel C: Planning

	(1)	(2)	(3)	(4)	(5)	(6)
Change in Obs Scores (PLAN)	−0.0635***	−0.0342***	−0.0156***	−0.0125***	−0.0150***	−0.0102***
	(0.00258)	(0.00191)	(0.00367)	(0.00358)	(0.00294)	(0.00279)
N	7,910,046	7,910,046	42,003	42,002	42,003	42,002
R2	0.019	0.078	0.011	0.040	0.558	0.587
Design	OLS	OLS	QE	QE	QE	QE
Controls		X		X		X
School-Year FE					X	X

Panel D: Professionalism

	(1)	(2)	(3)	(4)	(5)	(6)
Change in Obs Scores (PROF)	−0.0508***	−0.0265***	−0.0149***	−0.0119**	−0.0166***	−0.0124***
	(0.00252)	(0.00199)	(0.00398)	(0.00391)	(0.00319)	(0.00307)
N	7,845,493	7,845,493	41,763	41,762	41,763	41,762
R2	0.018	0.077	0.011	0.040	0.550	0.580
Design	OLS	OLS	QE	QE	QE	QE
Controls		X		X		X
School-Year FE					X	X

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors in parentheses.

Note: Table shows leave-year-out (LYO) domain-specific observation score effects on the number of student suspensions. The unit of analyses are student-year observations in the OLS model and department-year observations in the QE model. "Controls" refer to a vector of observable student characteristics and include lagged outcomes, student race/ethnicity, gender, FRPL status, ELL status, and SPED status. For the QE specification, all student outcome, teacher quality, and control variables are operationalized as first differences (t - t-1) and results are weighted by the number of students in each cell. All models include separate dummy variables for year and grade level (i.e., elementary, middle, high) and use standard errors clustered at the school-cohort level.

Table 5.12: Back-Projected Observation Score Effects on K-12 Outcomes

Panel A: Student Test Scores

	(1)	(2)	(3)	(4)	(5)	(6)
Change in Obs Scores	0.178***	0.0825***	0.0898***	0.0774***	0.0870***	0.0728***
	(0.00324)	(0.00111)	(0.00401)	(0.00332)	(0.00397)	(0.00329)
N	3,812,868	3,812,868	45,164	35,576	45,164	35,576
R2	0.039	0.596	0.034	0.276	0.255	0.487
Design	OLS	OLS	QE	QE	QE	QE
Controls		X		X		X
School-Year FE					X	X

Panel B: Student Absences

	(1)	(2)	(3)	(4)	(5)	(6)
Change in Obs Scores	–0.328***	–0.145***	–0.0711	–0.0713	–0.0912**	–0.0642**
	(0.0220)	(0.0157)	(0.0547)	(0.0531)	(0.0288)	(0.0247)
N	4,080,792	4,080,792	21,943	21,887	21,943	21,887
R2	0.023	0.299	0.086	0.156	0.842	0.870
Design	OLS	OLS	QE	QE	QE	QE
Controls		X		X		X
School-Year FE					X	X

Panel C: Student Suspensions

	(1)	(2)	(3)	(4)	(5)	(6)
Change in Obs Scores	–0.0926***	–0.0532***	–0.0366***	–0.0323***	–0.0254***	–0.0196***
	(0.00404)	(0.00324)	(0.00711)	(0.00699)	(0.00609)	(0.00581)
N	4,080,902	4,080,902	21,967	21,893	21,967	21,893
R2	0.026	0.091	0.021	0.053	0.604	0.632
Design	OLS	OLS	QE	QE	QE	QE
Controls		X		X		X
School-Year FE					X	X

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors in parentheses.

Note: Table shows back-projected observation score effects on student test z-scores (Panel A), the number of student absences (Panel B), and the number of student suspensions (Panel C). Student outcome data and rosters are drawn from 2007-08 to 2010-11 and matched to back-projected observation scores calculated using observation score data from 2011-12 to 2017-18. The unit of analyses are student-year observations in the OLS model and department-year observations in the QE model. "Controls" refer to a vector of observable student characteristics and include lagged outcomes, student race/ethnicity, gender, FRPL status, ELL status, and SPED status. For the QE specification, all student outcome, teacher quality, and control variables are operationalized as first differences (t - t-1) and results are weighted by the number of students in each cell. All models include separate dummy variables for year and grade level (i.e., elementary, middle, high) and use standard errors clustered at the school-cohort level.

Table 5.13: LYO Observation Score Effects on On-Time HS Graduation

	(1)	(2)	(3)	(4)	(5)	(6)
Change in Obs Scores	0.0124***	0.00426***	0.00464***	0.00247**	0.00549***	0.00142
	(0.000611)	(0.000553)	(0.000942)	(0.000917)	(0.000983)	(0.000937)
N	5,162,671	5,162,671	29,778	29,714	29,778	29,714
R2	0.034	0.085	0.018	0.074	0.355	0.446
Design	OLS	OLS	QE	QE	QE	QE
Controls		X		X		X
School-Year FE					X	X

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors in parentheses.

Note: Table shows leave-year-out (LYO) observation score effects on the probability that a student graduates high school on-time. All results include data for teachers for whom LYO scores have been "back-projected". The unit of analyses are student-year observations in the OLS model and department-year observations in the QE model. "Controls" refer to a vector of observable student characteristics and include lagged test scores, absences, and suspensions, student race/ethnicity, gender, FRPL status, ELL status, and SPED status. For the QE specification, all student outcome, teacher quality, and control variables are operationalized as first differences (t - t-1) and results are weighted by the number of students in each cell. All models include separate dummy variables for year and grade level (i.e., elementary, middle, high) and use standard errors clustered at the school-cohort level.

Table 5.14: LYO Observation Score Effects on Post-Secondary Enrollment at Age 19

Panel A: Enroll at any PS Institution at Age 19

	(1)	(2)	(3)	(4)	(5)	(6)
Change in Obs Scores	0.0367***	0.0198***	0.0156***	0.0114***	0.0164***	0.00933***
	(0.000968)	(0.000743)	(0.00138)	(0.00121)	(0.00156)	(0.00122)
N	5,326,577	5,326,577	29,632	29,569	29,632	29,569
R2	0.006	0.112	0.020	0.138	0.307	0.486
Design	OLS	OLS	QE	QE	QE	QE
Controls		X		X		X
School-Year FE					X	X

Panel B: Enroll at Non-BA Granting Institution at Age 19

	(1)	(2)	(3)	(4)	(5)	(6)
Change in Obs Scores	0.0125***	–0.00194	0.00783***	0.00574***	0.00808***	0.00451***
	(0.00134)	(0.00122)	(0.00125)	(0.00121)	(0.00128)	(0.00119)
N	5,326,577	5,326,577	29,632	29,569	29,632	29,569
R2	0.005	0.053	0.014	0.049	0.321	0.375
Design	OLS	OLS	QE	QE	QE	QE
Controls		X		X		X
School-Year FE					X	X

Panel C: Enroll at BA Granting Institution at Age 19

	(1)	(2)	(3)	(4)	(5)	(6)
Change in Obs Scores	0.0310***	0.0208***	0.0125***	0.00904***	0.0131***	0.00737***
	(0.00111)	(0.000891)	(0.00128)	(0.00117)	(0.00142)	(0.00118)
N	5,326,577	5,326,577	29,632	29,569	29,632	29,569
R2	0.006	0.072	0.014	0.110	0.281	0.421
Design	OLS	OLS	QE	QE	QE	QE
Controls		X		X		X
School-Year FE					X	X

Panel D: Enroll at Non-THEC Reporting Institution at Age 19

	(1)	(2)	(3)	(4)	(5)	(6)
Change in Obs Scores	0.01000***	0.00823***	0.00402***	0.00323***	0.00373***	0.00242***
	(0.000592)	(0.000480)	(0.000575)	(0.000555)	(0.000612)	(0.000572)
N	5,326,577	5,326,577	29,632	29,569	29,632	29,569
R2	0.002	0.019	0.005	0.034	0.286	0.324
Design	OLS	OLS	QE	QE	QE	QE
Controls		X		X		X
School-Year FE					X	X

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors in parentheses.

Note: Table shows leave-year-out (LYO) observation score effects on the probability a student enrolls in any (Panel A), a Non-BA Granting (Panel B), a BA-Granting (Panel C) or Non-THEC Reporting post-secondary institution at age 19 (Panel D). All results include "back-projected" scores. "Controls" refer to a vector of observable student characteristics and include lagged outcomes, student race/ethnicity, gender, FRPL, ELL, and SPED status. For the QE specification, all outcome, teacher quality, and control variables are operationalized as first differences and results are weighted by the number of students in each cell. All models include separate dummy variables for year and grade level (i.e., elementary, middle, high) and use standard errors clustered at the school-cohort level.

Table 5.15: LYO Observation Score Effects on Post-Secondary Completion at Age 23

Panel A: Any PS Award at Age 23 (Reporting Institutions)

	(1)	(2)	(3)	(4)	(5)	(6)
Change in Obs Scores	0.0348***	0.0188***	0.0173***	0.0102***	0.0188***	0.00935***
	(0.00152)	(0.00116)	(0.00295)	(0.00255)	(0.00369)	(0.00284)
N	1,189,795	1,189,795	6,320	6,311	6,320	6,311
R2	0.006	0.086	0.031	0.179	0.338	0.538
Design	OLS	OLS	QE	QE	QE	QE
Controls		X		X		X
School-Year FE					X	X

Panel B: Associates/Certificate at Age 23 (Reporting Institutions)

	(1)	(2)	(3)	(4)	(5)	(6)
Change in Obs Scores	0.00365***	–0.00262**	–0.00191	–0.00314*	–0.000985	–0.00293
	(0.00101)	(0.000919)	(0.00151)	(0.00148)	(0.00181)	(0.00178)
N	1,189,795	1,189,795	6,320	6,311	6,320	6,311
R2	0.000	0.019	0.002	0.019	0.383	0.414
Design	OLS	OLS	QE	QE	QE	QE
Controls		X		X		X
School-Year FE					X	X

Panel C: Bachelor's Degree at Age 23 (Reporting Institutions)

	(1)	(2)	(3)	(4)	(5)	(6)
Change in Obs Scores	0.0329***	0.0217***	0.0207***	0.0144***	0.0212***	0.0129***
	(0.00150)	(0.00115)	(0.00267)	(0.00228)	(0.00343)	(0.00267)
N	1,189,795	1,189,795	6,320	6,311	6,320	6,311
R2	0.009	0.084	0.046	0.197	0.351	0.539
Design	OLS	OLS	QE	QE	QE	QE
Controls		X		X		X
School-Year FE					X	X

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors in parentheses.

Note: Table shows leave-year-out (LYO) observation score effects on the probability a student earns any post-secondary credential (Panel A), an Associates/Certificate (Panel B), or a Bachelor's Degree (Panel C) at age 19. All results include data for teachers for whom LYO scores have been "back-projected". Observation scores are operationalized as annual averages from year t and do not incorporate shrinkage, regression adjustment, or data beyond year t. The unit of analyses are student-year observations in the OLS model and department-year observations in the QE model. "Controls" refer to a vector of observable student characteristics and include lagged test scores, absences, and suspensions, student race/ethnicity, gender, FRPL status, ELL status, and SPED status. For the QE specification, all student outcome, teacher quality, and control variables are operationalized as first differences (t - t-1) and results are weighted by the number of students in each cell. All models include separate dummy variables for year and grade level (i.e., elementary, middle, high) and use standard errors clustered at the school-cohort level.

Table 5.16: LYO Observation Score Effects on In-State Wages

Panel A: In-State Wages at Age 23 (Reporting Industries)

	(1)	(2)	(3)	(4)	(5)	(6)
Change in Obs Scores	241.3***	−54.43	111.8	107.9	101.5	51.68
	(39.18)	(30.47)	(75.71)	(73.30)	(90.83)	(89.17)
N	649,246	649,246	5,518	5,503	5,518	5,503
R2	0.002	0.066	0.012	0.066	0.392	0.433
Design	OLS	OLS	QE	QE	QE	QE
Controls		X		X		X
School-Year FE					X	X

Panel B: In-State Wages at Age 24 (Reporting Industries)

	(1)	(2)	(3)	(4)	(5)	(6)
Change in Obs Scores	575.8***	157.9***	325.7**	288.8**	483.5***	367.1*
	(56.74)	(42.42)	(109.5)	(105.1)	(143.4)	(144.4)
N	388,932	388,932	3,159	3,159	3,159	3,159
R2	0.002	0.071	0.012	0.071	0.347	0.420
Design	OLS	OLS	QE	QE	QE	QE
Controls		X		X		X
School-Year FE					X	X

Panel C:In-State Wages at Age 25 (Reporting Industries)

	(1)	(2)	(3)	(4)	(5)	(6)
Change in Obs Scores	871.3***	367.0***	756.3***	620.4**	1020.8**	673.0*
	(85.40)	(62.05)	(198.3)	(200.4)	(318.1)	(308.0)
N	211,917	211,917	1,722	1,721	1,722	1,721
R2	0.004	0.080	0.012	0.068	0.476	0.538
Design	OLS	OLS	QE	QE	QE	QE
Controls		X		X		X
School-Year FE					X	X

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors in parentheses.

Note: Table shows leave-year-out (LYO) observation score effects on students' annual in-state wages at age 23 (Panel A), age 24 (Panel B), and age 25 (Panel C). All results include data for teachers for whom LYO scores have been "back-projected". The unit of analyses are student-year observations in the OLS model and department-year observations in the QE model. "Controls" refer to a vector of observable student characteristics and include lagged test scores, absences, and suspensions, student race/ethnicity, gender, FRPL status, ELL status, and SPED status. For the QE specification, all student outcome, teacher quality, and control variables are operationalized as first differences (t - t-1) and results are weighted by the number of students in each cell. All models include separate dummy variables for year and grade level (i.e., elementary, middle, high) and use standard errors clustered at the school-cohort level.

Table 5.17: Single and Multiple Measure Effects on Achievement (Matched Sample)

	(1)	(2)	(3)	(4)	(5)	(6)
Change in Obs Score	0.0573***	0.101***		0.0395***	0.0784***	
	(0.00321)	(0.00307)		(0.00228)	(0.00228)	
Change in VA	0.110***		0.130***	0.0970***		0.111***
	(0.00300)		(0.00276)	(0.00232)		(0.00216)
N	79,390	79,390	79,390	79,390	79,390	79,390
R2	0.079	0.043	0.070	0.580	0.558	0.577
Design	QE	QE	QE	QE	QE	QE
Controls				X	X	X
School-Year FE				X	X	X

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors in parentheses.

Note: Table shows leave-year-out (LYO) observation score effects and teacher VA effects on student achievement using two specifications of the teacher switching quasi-experimental (QE) model: A standard specification (Columns 1-3) and a saturated specification that includes School-Year fixed effects and controls. "Controls" refer to a vector of observable student characteristics and include lagged test scores, absences, and suspensions, student race/ethnicity, gender, FRPL status, ELL status, and SPED status. All student outcome, teacher quality, and control variables are operationalized as first differences (t - t-1) and results are weighted by the number of students in each cell. Both observation scores and teacher VA are standardized to be mean zero with unit variance. All models include separate dummy variables for year and grade level (i.e., elementary, middle, high) and use standard errors clustered at the school-cohort level.

Table 5.18: Single and Multiple Measure Effects on Suspensions (Matched Sample)

	(1)	(2)	(3)	(4)	(5)	(6)
Change in Obs Score	–0.0253***	–0.0269***		–0.0139***	–0.0156***	
	(0.00369)	(0.00341)		(0.00265)	(0.00248)	
Change in VA	–0.00406		–0.0128***	–0.00437		–0.00923***
	(0.00327)		(0.00302)	(0.00236)		(0.00221)
N	102,652	102,652	102,652	102,652	102,652	102,652
R2	0.016	0.016	0.015	0.595	0.595	0.595
Design	QE	QE	QE	QE	QE	QE
Controls				X	X	X
School-Year FE				X	X	X

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors in parentheses.

Note: Table shows leave-year-out (LYO) observation score effects and teacher VA effects on the number of student suspensions using two specifications of the teacher switching quasi-experimental (QE) model: A standard specification (Columns 1-3) and a saturated specification that includes School-Year fixed effects and controls. "Controls" refer to a vector of observable student characteristics and include lagged test scores, absences, and suspensions, student race/ethnicity, gender, FRPL status, ELL status, and SPED status. All student outcome, teacher quality, and control variables are operationalized as first differences (t - t-1) and results are weighted by the number of students in each cell. Both observation scores and teacher VA are standardized to be mean zero with unit variance. All models include separate dummy variables for year and grade level (i.e., elementary, middle, high) and use standard errors clustered at the school-cohort level.

Table 5.19: Single and Multiple Measure Effects on On-Time HS Graduation (Matched Sample)

	(1)	(2)	(3)	(4)	(5)	(6)
Change in Obs Score	–0.00119	–0.000178		–0.00210	–0.00206	
	(0.00362)	(0.00334)		(0.00341)	(0.00320)	
Change in VA	0.00248		0.00204	0.0000939		–0.000686
	(0.00363)		(0.00335)	(0.00325)		(0.00304)
N	56,870	56,870	56,870	56,870	56,870	56,870
R2	0.047	0.047	0.047	0.389	0.389	0.389
Design	QE	QE	QE	QE	QE	QE
Controls				X	X	X
School-Year FE				X	X	X

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors in parentheses.

Note: Table shows leave-year-out (LYO) observation score effects and teacher VA effects on the number of student absences using two specifications of the teacher switching quasi-experimental (QE) model: A standard specification (Columns 1-3) and a saturated specification that includes School-Year fixed effects and controls. "Controls" refer to a vector of observable student characteristics and include lagged test scores, absences, and suspensions, student race/ethnicity, gender, FRPL status, ELL status, and SPED status. All student outcome, teacher quality, and control variables are operationalized as first differences (t - t-1) and results are weighted by the number of students in each cell. Both observation scores and teacher VA are standardized to be mean zero with unit variance. All models include separate dummy variables for year and grade level (i.e., elementary, middle, high) and use standard errors clustered at the school-cohort level.

Table 5.20: Single and Multiple Measure Effects on Absences (Matched Sample)

	(1)	(2)	(3)	(4)	(5)	(6)
Change in Obs Score	–0.0977***	–0.130***		–0.0355**	–0.0513***	
	(0.0251)	(0.0233)		(0.0131)	(0.0122)	
Change in VA	–0.0826***		–0.116***	–0.0403**		–0.0527***
	(0.0238)		(0.0221)	(0.0123)		(0.0115)
N	102,652	102,652	102,652	102,652	102,652	102,652
R2	0.075	0.075	0.075	0.812	0.812	0.812
Design	QE	QE	QE	QE	QE	QE
Controls				X	X	X
School-Year FE				X	X	X

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors in parentheses.

Note: Table shows leave-year-out (LYO) observation score effects and teacher VA effects on the number of student absences using two specifications of the teacher switching quasi-experimental (QE) model: A standard specification (Columns 1-3) and a saturated specification that includes School-Year fixed effects and controls. "Controls" refer to a vector of observable student characteristics and include lagged test scores, absences, and suspensions, student race/ethnicity, gender, FRPL status, ELL status, and SPED status. All student outcome, teacher quality, and control variables are operationalized as first differences (t - t-1) and results are weighted by the number of students in each cell. Both observation scores and teacher VA are standardized to be mean zero with unit variance. All models include separate dummy variables for year and grade level (i.e., elementary, middle, high) and use standard errors clustered at the school-cohort level.

Table 5.21: Single and Multiple Measure Effects on PS Enrollment at 19 (Matched Sample)

	(1)	(2)	(3)	(4)	(5)	(6)
Change in Obs Score	0.00834*	0.0105***		0.00469	0.00560*	
	(0.00330)	(0.00304)		(0.00287)	(0.00266)	
Change in VA	0.00539		0.00846**	0.00223		0.00399
	(0.00331)		(0.00305)	(0.00283)		(0.00262)
N	42,129	42,129	42,129	42,129	42,129	42,129
R2	0.119	0.119	0.119	0.509	0.509	0.509
Design	QE	QE	QE	QE	QE	QE
Controls				X	X	X
School-Year FE				X	X	X

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors in parentheses.

Note: Table shows leave-year-out (LYO) observation score effects and teacher VA effects on the number of student absences using two specifications of the teacher switching quasi-experimental (QE) model: A standard specification (Columns 1-3) and a saturated specification that includes School-Year fixed effects and controls. "Controls" refer to a vector of observable student characteristics and include lagged test scores, absences, and suspensions, student race/ethnicity, gender, FRPL status, ELL status, and SPED status. All student outcome, teacher quality, and control variables are operationalized as first differences (t - t-1) and results are weighted by the number of students in each cell. Both observation scores and teacher VA are standardized to be mean zero with unit variance. All models include separate dummy variables for year and grade level (i.e., elementary, middle, high) and use standard errors clustered at the school-cohort level.

Table 5.22: Single and Multiple Measure Effects on PS Award at 23 (Matched Sample)

	(1)	(2)	(3)	(4)	(5)	(6)
Change in Obs Score	0.00931	0.0115*		0.00935	0.00780	
	(0.00537)	(0.00476)		(0.00504)	(0.00445)	
Change in VA	0.00484		0.00846	–0.00345		0.000252
	(0.00553)		(0.00492)	(0.00524)		(0.00465)
N	8,544	8,544	8,544	8,544	8,544	8,544
R2	0.111	0.111	0.111	0.510	0.510	0.509
Design	QE	QE	QE	QE	QE	QE
Controls				X	X	X
School-Year FE				X	X	X

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors in parentheses.

Note: Table shows leave-year-out (LYO) observation score effects and teacher VA effects on the number of student absences using two specifications of the teacher switching quasi-experimental (QE) model: A standard specification (Columns 1-3) and a saturated specification that includes School-Year fixed effects and controls. "Controls" refer to a vector of observable student characteristics and include lagged test scores, absences, and suspensions, student race/ethnicity, gender, FRPL status, ELL status, and SPED status. All student outcome, teacher quality, and control variables are operationalized as first differences (t - t-1) and results are weighted by the number of students in each cell. Both observation scores and teacher VA are standardized to be mean zero with unit variance. All models include separate dummy variables for year and grade level (i.e., elementary, middle, high) and use standard errors clustered at the school-cohort level.

Table 5.23: Single and Multiple Measure Effects on Annual Wages at 25 (Matched Sample)

	(1)	(2)	(3)	(4)	(5)	(6)
Change in Obs Score	−304.7	222.7		−552.2	323.6	
	(562.1)	(513.9)		(800.8)	(770.0)	
Change in VA	1061.0*		948.0*	1730.0*		1505.9*
	(495.7)		(449.1)	(775.3)		(725.3)
N	2,175	2,175	2,175	2,175	2,175	2,175
R2	0.015	0.013	0.015	0.392	0.388	0.391
Design	QE	QE	QE	QE	QE	QE
Controls				X	X	X
School-Year FE				X	X	X

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors in parentheses.

Note: Table shows leave-year-out (LYO) observation score effects and teacher VA effects on the number of student absences using two specifications of the teacher switching quasi-experimental (QE) model: A standard specification (Columns 1-3) and a saturated specification that includes School-Year fixed effects and controls. "Controls" refer to a vector of observable student characteristics and include lagged test scores, absences, and suspensions, student race/ethnicity, gender, FRPL status, ELL status, and SPED status. All student outcome, teacher quality, and control variables are operationalized as first differences (t - t-1) and results are weighted by the number of students in each cell. Both observation scores and teacher VA are standardized to be mean zero with unit variance. All models include separate dummy variables for year and grade level (i.e., elementary, middle, high) and use standard errors clustered at the school-cohort level.

Table 5.24: Teacher VA Effects on In-State Wages at Age 25

	(1)	(2)	(3)	(4)	(5)	(6)
Change in VA (Std)	142.6	206.3***	665.7	652.3	1397.6*	1236.0
	(81.95)	(57.01)	(462.0)	(463.1)	(650.6)	(649.9)
N	147,855	147,855	2,400	2,400	2,400	2,400
R2	0.004	0.077	0.012	0.019	0.375	0.384
Design	OLS	OLS	QE	QE	QE	QE
Controls		X		X		X
School-Year FE					X	X

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors in parentheses.

Note: Table shows leave-year-out (LYO) observation score effects and teacher VA effects on annual wages at age 25 using two specifications of the teacher switching quasi-experimental (QE) model: A standard specification (Columns 1-3) and a saturated specification that includes School-Year fixed effects and controls. "Controls" refer to a vector of observable student characteristics and include lagged test scores, absences, and suspensions, student race/ethnicity, gender, FRPL status, ELL status, and SPED status. All student outcome, teacher quality, and control variables are operationalized as first differences (t - t-1) and results are weighted by the number of students in each cell. Both observation scores and teacher VA are standardized to be mean zero with unit variance. All models include separate dummy variables for year and grade level (i.e., elementary, middle, high) and use standard errors clustered at the school-cohort level.

Chapter 6

Conclusion

In my dissertation, I evaluate whether teacher observation scores collected as part of Tennessee's statewide teacher evaluation system validly identify differences in how teachers impact student success across K-12, post-secondary, and labor market outcomes. When applying techniques similar to those used to rigorously evaluate teacher value-added, I find considerable evidence that this is the case. Specifically, by using a teacher switching design that compares the outcomes and exposure to teacher quality, as measured by observation scores, of students passing through the same departments in adjacent years, I find that increases in observation scores appear to credibly identify improvements in both proximate K-12 outcomes, such as student achievement, attendance, and discipline, as well as more distal long-run outcomes, such as the rates of on-time high school graduation, post-secondary enrollment and completion, and wages in early adulthood. These findings appear to be largely robust across a range of modeling specifications and measurement adjustments. Comparisons with the analogously estimated effects when using teacher value-added estimates as the measure of teacher effectiveness indicate that while the magnitude of observation score "effects" is slightly smaller with regard to student achievement, observation score effects as large, if not larger, on the majority of non-test outcomes analyzed in this dissertation.

These findings lend themselves to a number of recommendations regarding how the classroom observation process could be strengthened in the future. First, the increase in predictive power between the various "adjusted" observation score measures, which use multiple years of teacher observation scores, and the single-year "unadjusted" observation score strongly suggests that school leaders should consider trends and multi-year averages of observation scores when making human capital decisions. While school leaders cannot obtain scores from the future to calculate "leave-out" estimates of the kind estimated in this analysis, they can use rolling three-year averages of

observation scores, as is used to produce teachers' overall TVAAS scores, to produce an observation score measure less susceptible to variation due to idiosyncratic factors that may distort a teacher's observation score in a single year.

In general, results from my dissertation conform with findings from Jacob (2011), Jacob and Lefgren (2008), Grissom and Loeb (2017) that indicate that, on average, administrators aided by a defined rubric are able to identify teacher quality with respect to multiple student outcomes. However, there is likely wide heterogeneity in this statement. Though both are prone to measurement error, the tools used to estimate teacher value-added, standardized test scores and a value-added model, are at least consistently applied for all students and teachers within a given setting. While most teachers in Tennessee are evaluated using the same TEAM rubric, the quality of the other components needed to produce an observation score, a human rater, varies widely across settings. Disinterested or overburdened administrators may not invest sufficient time and effort into conducting classroom observations, producing observation scores that contain little valuable information. Conversely, observation scores can potentially be highly informative if rated by administrators invested and adequately trained in the observation process. The strong observation score effects identified in this analysis are a promising signal that, on average, the effort invested by Tennessee school administrators in the observation process has resulted in the collection of meaningful information about teacher quality. This finding should encourage Tennessee, in addition to other states using observation scores, to invest in the training of observation raters or to provide support for additional administrator roles that are dedicated to teacher evaluation. An interim step that could be implemented to mitigate rater-driven error would be to encourage schools where multiple raters are available to assign raters to teachers on a rotating basis, which would help reduce the influence of specific teacher-rater biases on teachers' overall observation scores. Future research will investigate methods for identifying rater quality, using the extent to which raters differentiate scores or raters' own evaluation ratings moderate the extent to which their assessments of teacher practice accurately reflect teachers' underlying effectiveness.

Next, my analysis of TEAM domain-specific effects on K-12 outocomes finds only subtle dif-

ferences between what each of the domains capture about how teachers affect student success, particularly for non-test outcomes. In on-going "TEAM 2.0" conversations to reduce the length of the observation rubric, results from this dissertation suggest that it is unlikely that removing select items or even entire domains from the rubric would meaningfully alter the "type" of teacher effectiveness captured by the observation process. One consideration is that an overall reduction in items would, of course, lower the reliability of the resulting overall observation scores but this decrease in reliability would be overshadowed by the gains in precision afforded by using observation scores across multiple years, as previously suggested.

Comparison analysis of the predictive power of observation scores and value-added suggest that the magnitude of impacts to student outcomes as a result of differences along these two measures are comparable across many outcomes. However, while the magnitudes of the overall "effects" stemming from changes in both measures are similarly sized, this does not entail that observation scores and value-added capture the same dimensions of teacher quality. Instead, a number of results in my dissertation, specifically the "multiple measure" results for student achievement and absences, support the notion that both measures capture distinct dimensions of teachers' ability to raise student outcomes. Previously, Tennessee has used different mechanisms to encourage principals to better align their classroom observation scores with teacher VA (Poon & Schwartz, 2015). Evidence that each measure is capable of identifying different pathways to affecting student outcomes suggests that these so-called "alignment" interventions may be short-sighted and could hamper the ability of observation scores to capture complimentary, rather than overlapping, information on teacher quality.

Lastly, arguably the most profound results in this dissertation are the significant impacts of the changes of students' exposure to teacher observation score levels on their rates of post-secondary attendance, post-secondary completion, and early career wages, with observation score effects on increased enrollment and completion of Bachelor's degree programs being particularly prominent. Effects on the long-run outcomes are only possible through the use of "back-projected" scores that use teachers' observation scores during the TEAM period to project their effectiveness

during years prior to the implementation of TEAM. While auxiliary analyses suggest that these "back-projected" scores are reasonable projections of teachers' past effectiveness, it will be worth investigating in the future whether the effects identified without the use of "back-projected" scores and additional precision from increased observations align with those presented in this dissertation. The benefit of additional years of data will be most strongly felt for analyses of the effects of teacher quality on wages, where the results I currently find, while encouraging, are imprecise. In light of recent policymaker interest in identifying ways that K-12 teachers can support student success "across the pipeline", the results from this dissertation provide promising evidence that teachers' classroom observation scores can be tools that are well-aligned with this goal.

REFERENCES

Arcidiacono, P., Bayer, P., & Hizmo, A. (2010). Beyond signaling and human capital: Education and the revelation of ability. *American Economic Journal: Applied Economics*, *2*(4), 76-104.

Atteberry, A., Loeb, S., & Wyckoff, J. (2016). Teacher churning reassignment rates and implications for student achievement. *Educational Evaluation and Policy Analysis*, 3-30.

Bacher-Hicks, A., Chin, M. J., Kane, T. J., & Staiger, D. O. (2017). *An evaluation of bias in three measures of teacher quality: Value-added, classroom observations, and student surveys* (Working Paper No. 23478). National Bureau of Economic Research.

Bacher-Hicks, A., Kane, T. J., & Staiger, D. O. (2014). *Validating teacher effect estimates using changes in teacher assignments in Los Angeles* (NBER Working Paper No. 20657). National Bureau of Economic Research.

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.

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics*, *29*(1), 37–65.

Blazar, D., & Kraft, M. A. (2017). Teacher and teaching effects on students' attitudes and behaviors. *Educational Evaluation and Policy Analysis*, *39*(1), 146–170.

Borko, H. (2004, November). Professional Development and Teacher Learning: Mapping the Terrain. *Educational Researcher*, *33*(8), 3-15.

Boyce, A. C. (1912). Qualities of merit in secondary school teachers. *Journal of Educational Psychology*, *3*(3), 144.

Brighouse, H., Ladd, H. F., Loeb, S., & Swift, A. (2016). Educational goods and values: A framework for decision makers. *Theory and Research in Education*, *14*(1), 3-25.

Callahan, R. E. (1962). *Education and the cult of efficiency: A study of the social forces that have shaped the administration of the public schools*. Chicago, IL: University of Chicago Press.

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., & Yagan, D. (2011). How does your kindergarten classroom affect your earnings? Evidence from Project STAR. *The Quarterly Journal of Economics*, *126*(4), 1593-1660.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2011). *The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood* (Tech. Rep.). National Bureau of Economic Research.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014a). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review*, *104*(9), 2593-2632.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014b). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *American Economic Review*, *104*(9), 2633-2679.

Chin, M., & Goldhaber, D. D. (2016). *Exploring explanations for the "weak" relationship between value added and observation-based measures of teacher performance* (Working Paper). Cambridge, MA: Center for Education Policy Research.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. (2005). Who teaches whom? Race and the distribution of novice teachers. *Economics of Education Review*, *24*(4), 377-392.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2006). Teacher-student matching and the assessment of teacher effectiveness. *Journal of Human Resources*, *41*(4), 778–820.

Cogan, M. (1973). *Clinical supervision*. Boston, MA: Houghton Mifflin Company.

Coleman, J. S. (1966). *Equality of educational opportunity* (Tech. Rep.). Washington D.C.: National Center on Educational Statistics.

Danielson, C. (1996). *Enhancing professional practice: A framework for teaching*. Alexandria, VA: Association for Supervision and Curriculum Development.

Dee, T. S., Jacob, B., & Schwartz, N. L. (2013). The effects of NCLB on school resources and practices. *Educational Evaluation and Policy Analysis*, *35*(2), 252-279.

Dieterle, S., Guarino, C. M., Reckase, M. D., & Wooldridge, J. M. (2015). How do principals assign students to teachers? Finding evidence in administrative data and the implications for value added. *Journal of Policy Analysis and Management*, *34*(1), 32-58.

Garrett, R., & Steinberg, M. P. (2015). Examining teacher effectiveness using classroom observation scores: Evidence from the randomization of teachers to students. *Educational Evaluation and Policy Analysis*, *37*(2), 224–242.

Gershenson, S. (2016). Linking teacher quality, student attendance, and student achievement. *Education Finance and Policy*, *11*(2), 125-149.

Gershenson, S., Hart, C., Lindsay, C., & Papageorge, N. W. (2017). The long-run impacts of same-race teachers.

Goe, L. (2010). *Evaluating teaching with multiple measures* (Tech. Rep.). Washington, DC: American Federation of Teachers.

Goldhammer, R. (1969). *Clinical supervision*. New York: Holt, Rinehart & Winston.

Goldring, E., Grissom, J. A., Rubin, M., Neumerski, C. M., Cannata, M., Drake, T., & Schuermann, P. (2015). Make room value added: Principals' human capital decisions and the emergence of teacher observation data. *Educational Researcher*, *44*(2), 96-104.

Green, E. (2014). *Building a better teacher: How teaching works (and how to teacher it to everyone)*. New York: W.W. Norton & Company.

Grissom, J. A., & Loeb, S. (2017). Assessing principals' assessments: Subjective evaluations of teacher effectiveness in low- and high-stakes environments. *Education Finance and Policy*, *12*(3), 369-395.

Grissom, J. A., & Youngs, P. (Eds.). (2015). *Improving teacher evaluation systems: Making the most of multiple measures*. New York: Teachers College Press.

Grossman, P., Cohen, J., Ronfeldt, M., & Brown, L. (2014). The test matters: The relationship between classroom observation scores and teacher value added on multiple types of assessment. *Educational Researcher*, *43*(6), 293–303.

Grossman, P., Loeb, S., Cohen, J., & Wyckoff, J. (2013). Measure for measure: The relationship between measures of instructional practice in middle school english language arts and teachers' value-added scores. *American Journal of Education*, *119*(3), 445-470.

Hallgren, K., James-Burdumy, S., & Perez-Johnson, I. (2014). *State requirements for teacher evaluation policies promoted by Race to the Top* (Tech. Rep.). National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences.

Hamilton, L. S., Stecher, B. M., & Yuan, K. (2008). *Standards-Based Reform in the United States: History, Research, and Future Directions* (Tech. Rep.). Santa Monica, CA: Center on Education Policy, RAND.

Hamre, B. K., Pianta, R. C., Mashburn, A. J., & Downer, J. T. (2007). *Building a science of classrooms: Application of the CLASS framework in over 4,000 US early childhood and elementary classrooms* (Tech. Rep.). New York: Foundation for Child Development.

Hansen, M., & Goldhaber, D. D. (2015). *Response to AERA statement on value-added measures: Where are the cautionary statements on alternative measures?*

Hanushek, E. A. (1970). *The value of teachers in teaching*. Santa Monica, CA: RAND.

Hanushek, E. A. (1971). Teacher characteristics and gains in student achievement: Estimation using micro data. *The American Economic Review*, *61*(2), 280-288.

Hanushek, E. A. (1981). Education policy research - An industry perspective. *Economics of Education Review*, *1*(2), 193-223.

Hanushek, E. A. (1992). The trade-off between child quantity and quality. *Journal of Political Economy*, *100*(1), 84-117.

Heckman, J. J., Moon, S. H., Pinto, R., Savelyev, P. A., & Yavitz, A. (2010). The rate of return to the HighScope Perry Preschool Program. *Journal of Public Economics*, *94*(1-2), 114-128.

Hill, H. C., Blunk, M. L., Charalambous, C. Y., Lewis, J. M., Phelps, G. C., Sleep, L., & Ball, D. L. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and Instruction*, *26*(4), 430-511.

Hill, H. C., Charalambous, C. Y., Blazar, D., McGinn, D., Kraft, M. A., Beisiegel, M., ... Lynch, K. (2012). Validating Arguments for Observational Instruments: Attending to Multiple Sources of Variation. *Educational Assessment*, *17*(2-3), 88-106.

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher*, *41*(2), 56-64.

Ho, A. D., & Kane, T. J. (2013). The reliability of classroom observations by school personnel. *Bill & Melinda Gates Foundation*.

Hock, H., & Isenberg, E. (2012). Methods for accounting for co-teaching in value-added models. working paper. *Mathematica Policy Research, Inc.*.

Howell, W. G. (2015). Results of President Obama's Race to the Top. *Education Next*, *15*(4).

Jackson, C. (2018). What Do Test Scores Miss? The Importance of Teacher Effects on Non–Test Score Outcomes. *Journal of Political Economy*, *126*(51), 2072-2107.

Jackson, C., & Cowan, J. (2018). *Assessing the evidence on teacher evaluation reforms* (CALDER Research Brief No. 13-1218-1). Washington, D.C.: National Center for Analysis of Longitudinal Data in Education Research.

Jackson, C. K. (2014). Teacher quality at the high school level: The importance of accounting for tracks. *Journal of Labor Economics*, *32*(4), 645–684.

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.

Jacob, B. A. (2011). Do principals fire the worst teachers? *Educational Evaluation and Policy Analysis*, *33*(4), 403-434.

Jacob, B. A., & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics*, *26*(1), 101-136.

Jacob, B. A., Lefgren, L., & Sims, D. P. (2010). The persistence of teacher-induced learning. *Journal of Human resources*, *45*(4), 915–943.

Jennings, J. L., & DiPrete, T. A. (2010). Teacher effects on social and behavioral skills in early elementary school. *Sociology of Education*, *83*(2), 135-159.

Kalogrides, D., Loeb, S., & Béteille, T. (2013). Systematic sorting: Teacher characteristics and class assignments. *Sociology of Education*, *86*(2), 103-123.

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). Have we identified effective teachers? Validating measures of effective teaching using random assignment. *Bill & Melinda Gates Foundation*.

Kane, T. J., & Staiger, D. O. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation* (Tech. Rep.). National Bureau of Economic Research.

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains* (Tech. Rep.). Bill and Melinda Gates Foundation.

Kane, T. J., Taylor, E. S., Tyler, J. H., & Wooten, A. L. (2011). Identifying effective classroom practices using student achievement data. *Journal of Human Resources*, *46*(3), 587–613.

Koedel, C. (2008). Teacher quality and dropout outcomes in a large, urban school district. *Journal of Urban Economics*, *64*(3), 560-572.

Koedel, C., Mihaly, K., & Rockoff, J. E. (2015). Value-added modeling: A review. *Economics of Education Review*, *47*, 180-195.

Kraft, M. A., & Gilmour, A. F. (2017). Revisiting the widget effect: Teacher evaluation reforms and the distribution of teacher effectiveness. *Educational Researcher*.

Ladd, H. F. (2017). No Child Left Behind: A deeply flawed federal policy. *Journal of Policy Analysis and Management*, *36*(2), 461-469.

Levin, H. M. (1970). *A new model of school effectiveness.* Palo Alto, CA: Stanford Center for Research and Development in Teaching.

Lockwood, J. R., & McCaffrey, D. F. (2009). Exploring student-teacher interactions in longitudinal achievement data. *Education Finance and Policy*, *4*(4), 439-467.

Mansfield, R. K. (2015). Teacher quality and student inequality. *Journal of Labor Economics*, *33*(3), 751-788.

Martínez, J. F., Schweig, J., & Goldschmidt, P. (2016). Approaches for combining multiple measures of teacher performance: Reliability, validity, and implications for evaluation policy. *Educational Evaluation and Policy Analysis*, *38*(4), 738-756.

McCaffrey, D. F., Lockwood, J., Koretz, D. M., & Hamilton, L. S. (2004). *Evaluating value-added models for teacher accountability*. Santa Monica, CA: RAND.

McGreal, T. L. (1983). *Successful teacher evaluation*. Alexandria, Va: Association for Supervision and Curriculum Development.

Milanowski, A. T. (2011). Validity research on teacher evaluation systems based on the framework for teaching. *Online Submission*.

Minnici, A., & Hill, D. D. (2007). *Educational architechs: Do state education agencies have the tools necessary to implement NLCB?* (Tech. Rep.). Washington, D.C.: Center on Education Policy.

Morgan, J. G. (2004). *The Education Improvement Act*. Tennessee State Comptroller.

Murnane, R. J. (1975). *The impact of school resources on the learning of inner city children*. Cambridge, MA: Balinger Publishing Company.

NIET. (2010). *TAP evaluation and compensation guide* (Tech. Rep.). National Institute for Excellence in Teaching.

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

Papay, J. P., & Kraft, M. A. (2015). Productivity returns to experience in the teacher labor market: Methodological challenges and new evidence on long-term career improvement. *Journal of Public Economics*, *130*, 105-119.

Pennington, A. (2014). *ESEA waivers and teacher-evaluation plans: State oversight of district-designed teacher-evaluation systems* (Tech. Rep.). Washington, D.C.: Center for American Progress.

Polikoff, M. S., McEachin, A. J., Wrabel, S. L., & Duque, M. (2014). The waive of the future? School accountability in the waiver era. *Educational Researcher*, *43*(1), 45-54.

Poon, A., & Schwartz, N. (2015). Improving feedback in teacher evaluations: An evaluation of Tennessee's TEAM coach initiative. In *Annual Meeting of the Association for Education Finance and Policy*. Washington, D.C..

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, *73*(2), 417–458.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *The American Economic Review*, *94*(2), 247-252.

Rothstein, J. (2009). Student sorting and bias in value-added estimation: Selection on observables and unobservables. *Education Finance and Policy*, *4*(4), 537–571.

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *The Quarterly Journal of Economics*, *125*(1), 175-214.

Rothstein, J. (2017). Measuring the impacts of teachers: Comment. *American Economic Review*, *107*(6), 1656-1684.

Rowan, B., Correnti, R., & Miller, R. (2002). What large scale survey research tells us about teacher effects on student achievement: Insights from the Prospects study of elementary schools. *Teachers College Record*, *104*, 1525-1567.

Ruediger, W. C., & Strayer, G. D. (1910). The qualities of merit in teachers. *Journal of Educational Psychology*, *1*(15), 272-278.

Ruzek, E. A., Domina, T., Conley, A. M., Duncan, G. J., & Karabenick, S. A. (2015). Using value-added models to measure teacher effects on students' motivation and achievement. *The Journal of Early Adolescence*, *35*(5-6), 852-882.

Sanders, W. L., & Horn, S. P. (1994). The Tennessee Value-Added Assessment System (TVAAS): Mixed-model methodology in educational assessment. *Journal of Personnel Evaluation in Education*, *8*, 299-311.

Sanders, W. L., & Horn, S. P. (1998). Research findings from the Tennessee Value-Added Assessment System (TVAAS) database: Implications for educational evaluation and research. *Journal of Personnel Evaluation in Education*, *12*(3), 247-256.

Sanders, W. L., & Rivers, J. C. (1996). *Cumulative and residual effects of teachers on future student academic achievement* (Tech. Rep.). Knoxville, TN: University of Tennessee Value-Added Research and Assessment Center.

Schacter, J., & Thum, Y. M. (2004). Paying for high- and low-quality teaching. *Economics of Education Review*, *23*(4), 411-430.

Steinberg, M. P., & Donaldson, M. L. (2016). The new educational accountability: Understanding the landscape of teacher evaluation in the post-NCLB era. *Education Finance and Policy*, *11*(3), 340-359.

Steinberg, M. P., & Garrett, R. (2016). Classroom composition and measured teacher performance: What do teacher observation scores really measure? *Educational Evaluation and Policy Analysis*, *38*(2), 293-317.

Steinberg, M. P., & Kraft, M. A. (2017). The sensitivity of teacher performance ratings to the design of teacher evaluation systems. *Educational Researcher*, *46*(7), 378-396.

Stone, Z. (2017). *The rise of student growth portfolio models in Tennessee* (Tech. Rep.). Nashville, TN: Tennessee Department of Education.

Sun, M., Loeb, S., & Grissom, J. A. (2017). Building teacher teams: Evidence of positive spillovers from more effective colleagues. *Educational Evaluation and Policy Analysis*, *39*(1), 104-125.

TDOE. (2018). *2017-18 observation guidelines*.

Tyack, D. B., & Cuban, L. (1995). *Tinkering toward utopia: A century of public school reform*. Cambridge, MA: Harvard University Press.

Veir, C. A., & Dagley, D. L. (2002). Legal issues in teacher evaluation legislation: A study of state statutory provisions. *BYU Educ. & LJ*, 1.

Whitehurst, G. J. R., Chingos, M. M., & Lindquist, K. M. (2014). Evaluating teachers with classroom observations. *Brown Center on Education Policy: Brookings Institute*.

Wise, A. A., Darling-Hammond, L., McLaughlin, M. W., & Bernstein, H. T. (1984). *Teacher evaluation: A study of effective practices*. Santa Monica, CA: RAND.