Identifying the Effects of Classroom Observations on Teacher Performance

By

Seth Baxter Hunter

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

Leadership and Policy Studies

June 30, 2018

Nashville, Tennessee

Approved:

Robert D. Ballou, Ph.D.

Jason A. Grissom, Ph.D.

Matthew G. Springer, Ph.D.

John H. Tyler, Ph.D.

For my amazing wife, Amy, who inspired, challenged, and supported me

and

For my son, Simon, who unknowingly provided invaluable perspective

ACKNOWLEDGEMENTS

Over the past year, I shared much of this research with the Association for Education Finance and Policy, Society for Research on Educational Effectiveness, Tennessee Education Research Alliance, and Tennessee Department of Education. To some degree I incorporated feedback from each presentation into this dissertation. I am especially indebted to the latter two organizations for insights regarding the study context.

I greatly appreciate the feedback from my committee members: Jason Grissom, Matthew Springer, and John Tyler. None of these committee members had to join my committee, and I am sure each had more pressing responsibilities. I sincerely appreciate the time each gave for my development.

I am eternally grateful for the mentorship of Dale Ballou, my advisor and dissertation chair. Over the course of my doctoral training Dale and I argued over identification strategies, the importance of conceptual frameworks, human nature, music, food, and more. He thinks he knows how important he is to me, but he has no idea. Dale shaped my standards for research, analytical skills, writing, teaching, and approach to mentoring. He is one of a kind and I am fortunate to call him my mentor.

Finally, I would not have been able to write this without the support of my family. My grandparents and parents always encouraged me to pursue my dreams and do my best. They may not realize it, but in some way, each passed on skills or beliefs that made this dissertation possible. I am most thankful for my lovely, talented, and caring wife, Amy, and my son, Simon, who kept reminding me what was important.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

INTRODUCTION

In the late 2000s, a confluence of factors led to substantial changes in educator evaluation systems. Whereas evaluation systems under No Child Left Behind (NCLB) largely focused on school performance (Manna, 2011; Mehta, 2013), these new evaluation systems focused on the teacher (Steinberg & Donaldson, 2016). Research produced over the 2000s suggested this new focus was warranted because researchers found teachers had a substantial impact on student achievement (Rivkin, Hanushek, & Kain, 2005; Rockoff, 2004) and teacher effectiveness varied substantially within schools (Aaronson et al., 2007; Rivkin et al., 2005). At the same time, educators in the late 2000s reported teacher evaluation was not helping them improve (Weisberg, Sexton, Mulhern, & Keeling, 2009). Soon after these findings became known, the federal Race to the Top competition incentivized state education agencies (SEA) to design new evaluation systems to improve teacher performance (US Department of Education, 2009).

Many changes to teacher evaluation fell into two broad categories: the introduction of student outcomes as a measure of teacher effectiveness, and introduction of standards-based observation protocols. The three most widely adopted means by which student achievement has been incorporated into teacher evaluations are value-added measures, student learning objectives, and student growth percentiles (Steinberg & Donaldson, 2016). In general, each of these three measures aims to differentiate teacher effectiveness based on student performance. This differentiation can inform personnel decision-making (e.g. retention) and teacher professional development plans. The second broad change to teacher evaluation systems concerned observations. While the practice of classroom observation had existed for decades (Brophy & Good, 1986), many classroom observation systems preceding the 2010s often used observation

rubrics that were not firmly grounded in research (Steinberg & Donaldson, 2016) and expected tenured teachers to be observed once every few years (Weisberg et al., 2009). Now, most modern teacher evaluation systems rely on standards-based observation protocols and expect most teachers to receive multiple observations per year (Steinberg & Donaldson, 2016). Conventional wisdom suggests these changes to teacher observation systems should improve teacher performance. That is, a higher frequency of observations using standards-based protocols should generate information educators can use to improve teacher performance.

While a great deal of recent research has investigated the effects of introducing student outcomes into teacher evaluation, the effects of observation-related changes have received less attention (Cohen & Goldhaber, 2016). Yet, there are at least three reasons to be concerned about these latter changes: burdens placed on administrators, costs of these new systems, and the importance of teachers to student outcomes. Recent research suggests new teacher evaluation systems demand more time from administrators (Kraft & Gilmour, 2016a; Neumerski et al., 2014), and that administrators believe these systems are quite burdensome (Rigby, 2015). These findings are not surprising considering that many teachers used to be observed once every few years, but are now expected to receive approximately four observations per year (Steinberg & Donaldson, 2016). Second, new observation systems are the most expensive component of modern teacher evaluation systems (Stecher et al., 2016). One hopes this large budget item is money well-spent, but to make this determination we need to know more about the relationship between observations and teacher performance. Finally, research finds students taught by more effective teachers experience better short- and long-term outcomes (Chetty, Friedman, & Rockoff, 2014). Although conventional wisdom posits teachers receiving a higher frequency of observations should improve, little research has examined whether this occurs.

I address the need to understand more about the effects of changes to observation systems by answering the following research question: what is the impact of more frequent observations on teacher performance? Some work has identified the effects of introducing modern observation *systems* on teacher performance (Steinberg & Sartain, 2015; Taylor & Tyler, 2012). In brief, this work suggests the adoption of modern observation systems (i.e. standards-based observation protocols) improves teacher performance. However, little, if any, research has examined whether undergoing more observations within these systems improves teacher performance. The administrative burdens and costs associated with more frequent observations may be worthwhile if observations appreciably improve teacher effectiveness. Alternatively, anything less than a positive relationship between observations and teacher performance would imply something about observation-related policies and/ or practices needs to change.

Treating variation in the number of observations received per year as exogenous is problematic (e.g. observers may observe less effective teachers more because they want to closely monitor their teaching). To overcome this endogeneity problem, I exploit policy-imposed discontinuities in the assignment of classroom observations. Policy-assigned observations depend on a continuous, observable measure of teacher performance. I identify exogenous variation in the number of observations received per year using a regression discontinuity design. Since educators have no control over policy-assigned observations, observations brought about by policy inducement are plausibly exogenous. I apply this research design to multiple years of administrative data from Tennessee to answer my research question.

To preview my findings, there is no evidence more observations improve teacher performance as measured by student growth scores. This holds true over teachers with varying levels of experience, at different grade levels, and who are evaluated by observers with greater or

lesser degrees of expertise. In Chapters 1and 2 I discuss the study context and review relevant research, respectively. Chapter 3 discusses methodology and data, and Chapter 4 threats to internal validity. Chapter 5 includes findings. In Chapter 6 I end with a discussion of study limitations and offer policy solutions that may improve the effectiveness of observations.

CHAPTER 1

STUDY CONTEXT

Introduction

In the early 2010s, the Tennessee Department of Education (TDOE) implemented sweeping changes to its teacher evaluation system. The primary goal of the new system, called the Tennessee Educator Acceleration Model (TEAM), is formative evaluation (Tennessee Board of Education, 2013). The secondary goal of TEAM concerns personnel decisions more broadly. Specifically, state policy urges educators to use TEAM evaluations when making decisions about hiring, tenure and dismissal, and compensation (Tennessee Board of Education, 2013).

In broad terms, TEAM aims to reach its primary goal by providing educators with measures of teacher performance that can be used to support teacher professional development, broadly defined (Alexander, 2016). TEAM generates "qualitative and quantitative" measures of teacher performance (Alexander, 2016). The qualitative measure is a summative observation score generated by trained observers using standards-based observation protocols. After an observation, observers should provide teachers with feedback that the teacher, or the teacher and observer, can use to improve teacher performance. Quantitative measures are based on student outcomes and are not generated until after the end of a school year (e.g. graduation rates, achievement, and value-added scores).

In what follows I briefly describe the calculation of growth and achievement scores because these scores partially determine a teacher's level of effectiveness (LOE), the running variable in regression discontinuity designs (RDDs). I then discuss the design of the classroom

observation process. This chapter ends with a brief explanation of the assignment of observations, the predictor of interest. I return to this assignment process in Chapter 3 when discussing instruments used in fuzzy RDDs.

Teacher Performance Measures Based on Student Outcomes

Two of three[1] TDOE measures of teacher effectiveness are based on student outcomes: the achievement and growth scores. The achievement measure is a measure of district- / school-wide student outcomes including student achievement scores, and graduation or attendance rates, etc. Teacher growth scores are based on student academic outcomes, but growth score options depend on whether the teacher teaches a tested subject.

A teacher and her school administrator[2] choose an achievement measure at the beginning of each school year from a TDOE approved list of measures (Tennessee State Board of Education, 2013). Students in a teacher's school or district generate scores produced by each of these measures. Achievement measures are based on aggregations of grade-, department-, school-, or district-wide student outcomes. Once an educator selects her achievement measure, teachers and their school administrator develop measurable performance criteria based on student outcomes produced by the achievement measure if such criteria do not exist. The teacher and her school administrator develop performance criteria aligned with the chosen achievement measure and map these criteria onto an integer scale of [1, 5]. For example, if a high school teacher and her school administrator select the graduation rate as the achievement measure, the performance criteria could assign percentage point changes in the graduation rate of (-100, -3), [-

[1] Some Tennessee districts use a fourth LOE determinant: study perception surveys. I exclude these districts from the analysis because they use alternative observation systems (i.e. non-TEAM observation systems).

[2] Not all school administrators serve as teacher evaluators, nor are all teacher evaluators school administrators. Nevertheless, more than 85% of teacher evaluators are principals or assistant principals (i.e. school administrators).

3, -1), [-1, 1), [1, 3), (3, 100] to levels 1, 2, 3, 4, and 5, respectively. (Tennessee Board of Education, 2013)

In principle, each teacher could develop her own achievement criteria, but in practice this does not happen. New performance criteria need approval from TDOE. Teachers and school administrators tend to avoid this approval process by using previously approved criteria. Teachers in the same school/ grade/ subject who choose the same school-/ grade-/ subject-level achievement measure tend to use the same achievement criteria.

The second quantitative TDOE measure of teacher effectiveness is the growth score. All teachers receive a growth score, however, the source of the score depends on whether the teacher teaches a tested subject. The Tennessee Value-Added Assessment System (TVAAS) estimates the impact of tested teachers on their students' test scores relative to the impact of the hypothetical average teacher on her students' test scores (SAS, 2016). The Tennessee suite of statewide accountability exams (e.g. End of Course exams, Tennessee Comprehensive Achievement Program) produces scores used by TVAAS.

A TVAAS *score* is converted into a TVAAS *index* by dividing a teacher's value-added estimate (i.e. TVAAS score) by the associated standard error (SAS, 2015). Thus, teachers with the same TVAAS score may have different TVAAS indices if their standard errors differ. The TVAAS index is continuous and ranges between -24 and 39 in the population of Tennessee teachers during the study period (2012-13 through 2014-15). These continuous scores are transformed into integer values in the range [1, 5] before use in the LOE calculation[3].

[3] Continuous growth indices in the ranges $(-\infty, -2)$, $[-2, -1)$, $[-1, 1)$, $[1, 2)$, and $[2, \infty)$ are respectively assigned integers 1, 2, 3, 4, and 5 (SAS, 2015).

TVAAS scores/ indices/ levels are only estimated for teachers of tested subjects, who represented less than 50% of all Tennessee teachers during the study period. TVAAS does not furnish a growth score for other Tennessee teachers[4]. The majority (80%) of the latter are assigned growth scores based on school-wide value-added scores. As the name implies, school level value-added scores are based on scores generated by students at a teacher's school[5]. The growth scores of all teachers are transformed into an integer in the interval [1, 5].

Teacher Performance Measures Based on Observational Ratings

Teacher LOE is determined by quantitative and qualitative measures. Qualitative measures are determined by the observation system. The Tennessee Board of Education (TBOE) and TDOE have issued multiple directives concerning classroom observations conducted as part of TEAM. For example, classroom observers must meet re/ certification expectations, prior performance determines the minimum number of observations teachers should receive, and teachers should always receive timely, post-observation feedback. In this section I describe key design elements of the TEAM classroom observation system.

There are multiple types of observations. Observations can be walkthroughs or non-walkthroughs. Both refer to a classroom visit by a TDOE-certified observer and either may generate post-observation feedback. However, a walkthrough should not result in any observation scores while non-walkthroughs must result in scores used for TEAM purposes. An observation is also unannounced or announced. The key difference between announced and

[4] During the study period local education agencies (LEA) assigned teachers of untested subjects one of three growth scores: (1) student portfolio scores for teachers of the Fine Arts, World Languages, and Physical Education, (2) K-2 student assessment scores for 1st and 2nd grade teachers, or (3) school-wide value-added scores for all non-tested teachers (Tennessee State Board of Education, 2013). Growth scores based on student portfolios or K-2 assessments are based on assessments of the teacher's content area.

[5] School level value-added scores are not typically based on assessments of the content taught by teachers of untested subjects. For example, no music assessment results are used in the estimation of school-wide value-added, although music teachers in some LEAs may be assigned the school-wide value-added score.

unannounced observations is that teachers know about the former observation in advance. TBOE policy states at least half of a teacher's observations must be unannounced (Tennessee Board of Education, 2013). Finally, observers can rate teachers with respect to one or two "domains" on the TEAM rubric during a single observation (I discuss domains in the following section).

For purposes of this dissertation, an observation refers to an announced or unannounced, non-walkthrough, single classroom visit conducted for the TEAM observation system. This operationalization means there is a one-to-one[6] correspondence between an observation, post-observation feedback session, and "observation cycle." An observation cycle includes a pre-observation conference if appropriate (see below), classroom visit that may result in multiple scored domains, a post-observation feedback session, and the in/ formal design and/ or refinement of a teacher improvement plan. TDOE expects observers to submit a single area for professional improvement after each classroom visit, no matter how many domains scored. Thus, according to my operationalization, the number of observations received by a teacher is the same as the number of post-observation feedback and improvement plan sessions received. For this reason, I use the terms "observations", "classroom visits", and "observation cycles" interchangeably.

[6] Observations could have been defined as the number of domains rated per year. For example, if an observer generated scores for two domains during a single classroom visit this could count as two observations instead of one. My definition of an observation represents a one-to-one-to-one-to-one relationship among classroom visits, observations, post-observation feedback sessions, and subsequent teacher performance improvement plans. The domain-based definition would represent a one-to-many-to-one-to-one relationship. This clouds the definition of treatment. Additionally, the domain-based definition should produce attenuated treatment effects. Notwithstanding these two points, I estimate the effects of domain-based observations in subsequently described RDDs. First-stage instruments are strong joint predictors of both versions of observations (all F-test p values < 0.001). Furthermore, all treatment effects behave as expected: estimates based on the number of domain-based observations received are attenuated relative to my definition.

Observation Rubrics: Characteristics and Scoring

Observations must be conducted using a TDOE approved rubric measuring at least three[7] domains: Instruction, Environment, and Planning (Tennessee Board of Education, 2013). While local education agencies (LEAs) could use their own rubrics, over 80% used the state-adopted TEAM rubric (see Appendix B) and accompanying TEAM observation system during the study period (Tennessee Department of Education, 2016). This dissertation focuses only on classroom observations in the TEAM observation system, given its widespread adoption and clear policies regarding the frequency of observations. TDOE and the National Institute for Excellence in Teaching co-developed the TEAM rubric, which is based on Charlotte Danielson's Framework for Teaching (Alexander, 2016). The Instruction domain includes twelve indicators, the most of any domain (see Appendix B). Some Instruction indicators are content neutral (e.g. Motivating Students, Grouping Students), but many indicators measure arguably content-specific teacher or teacher-student interactions (e.g. Academic Feedback, Presenting Instructional Content, Questioning). The second domain is the Environment domain, consisting of four indicators: Expectations, Managing Student Behavior, Environment, and Respectful Culture. The third domain is Planning, which includes the fewest indicators: Instructional Plans, Student Work, and Assessment. If the observation is announced, observers can judge a teacher with respect to the Planning domain during the pre-conference. Thus, only behaviors occurring during a lesson should inform the scoring of Instruction and Environment domains (see Appendix B).

Within the confines of TEAM observation policy, observers assign integer scores of [1, 5] with respect to each indicator. "Significantly Above Expectations" refers to exemplary

[7] TDOE expects observation rubrics to include a fourth domain, the Professionalism domain, but this domain is not scored during classroom visits. Observers use the Professionalism domain to judge the extent to which a teacher engages in extra-instructional practices such as data use and professional growth (see Appendix B for a copy of the TEAM rubric).

teacher/ student behaviors with respect to each indicator. These behaviors receive a score of five. "Significantly Below Expectations" refers to undesirable behaviors and receive a score of one. An observer could also generate a rating of three, indicating the behavior was "At Expectations" (see Appendix B). TDOE expects observers to generate Instruction, Environment, and Planning scores based on the preponderance of evidence observed. If the observed evidence does not place a teacher squarely into one of the three levels of performance, an observer can assign a rating of two (four) for a preponderance of behaviors straddling the lowest and middle (and highest) categories. (Alexander, 2016)

Observer Certification

Classroom observations should only be conducted by TDOE certified observers (Tennessee Board of Education, 2013). First-time observers must attend an initial TEAM training in which they hone the accuracy of their observation scores and learn about the TEAM teacher evaluation system. Attendees also receive some annual training (less than six hours) regarding pre- and post-observation conferences. Below, I describe these conferences in detail. Following this initial training, all prospective observers must demonstrate mastery regarding their knowledge of the teacher evaluation system and accurate rating of teacher/ student behaviors by taking an online certification exam administered by the National Institute for Excellence in Teaching. (Alexander, 2016)

Pre- and Post-Observation Conferences

Teachers should participate in post-observation conferences (a conference held after each observation) to discuss teacher strengths, weaknesses, and plans for improvement (Tennessee

Board of Education, 2013). Observers should also initiate a pre-observation conference when the observation is announced (Alexander, 2016). The purpose of the pre-conference is for the observer to learn more about the upcoming lesson. Observers may take this opportunity to score a teacher/ lesson with respect to the Planning domain and/ or provide suggestive formative feedback to the teacher in order to improve lesson outcomes (Alexander, 2016). Observers and teachers are expected to engage in a post-observation conference within one week of an announced or unannounced TEAM observation (Alexander, 2016; Tennessee Board of Education, 2013). Observers should collect their observational notes and ratings and be prepared to share their feedback with the teacher during the post-observation conference (Tennessee Board of Education, 2013).

TDOE expects observers to share more than observation scores with a teacher during the post-observation conference. During the post-observation conference the teacher and observer should draw on observation scores and notes to identify a teacher's area of greatest strength and weakness (Alexander, 2016). Specifically, the teacher and observer should select one indicator (e.g. one of the twelve Instruction indicators) as the teacher's area of greatest strength and one indicator as a teacher's greatest weakness after each classroom visit. During the post-observation conference the observer and teacher should develop a set of actionable next steps to develop the teacher's practice with respect to their area of weakness (Alexander, 2016). Observers should enter brief notes about a teacher's greatest strength and weakness into the TDOE administrative data system. In subsequent observations, observers should monitor a teacher's area of weakness to support instructional improvement.

Determination of Summative Observational Ratings

TEAM summative observational scores are the mean indicator score across all ratings received within a school year. TEAM scores ("summative TEAM scores" and "TEAM scores/ ratings" used interchangeably) are approximately continuous and range from [1, 5].

Teacher Level of Effectiveness (LOE)

LOE is a composite measure of teacher effectiveness. Different factors scale each of the three LOE determinants. Throughout the study period (2012-13 through 2014-15) achievement scores were scaled by a factor of 15. In 2012-13 the growth scores (i.e. growth scores measured in 2011-12) of all teachers were scaled by 35, but this factor was lowered to 25 for teachers of untested subjects beginning in the 2012-13 year (i.e. this changed first affected growth scores in 2013-14). Finally, the TEAM score of all teachers was scaled by a factor of 50 in 2012-13 school year, but beginning in the 2012-13 school year the summative observation score for teachers of untested subjects was scaled by a factor of 60. The sum of these three scale scores produces a variable I will call LOE-cont, an approximately continuous measure ranging from 100 to 500 (Tennessee State Board of Education, 2013).

TDOE uses LOE-cont to assign teachers to one of five discrete LOE categories (LOE). Teachers whose LOE-cont is within [100, 200), [200, 275), [275,350), [350, 425), or [425, 500] are respectively assigned LOE scores of 1, 2, 3, 4, or 5 (Tennessee Department of Education, n.d.-b).

There are two exceptions to the rule assigning LOE-cont to LOE. If a teacher's TVAAS level is three or greater, and greater than her achievement score, she can use the TVAAS level in lieu of the achievement score. This substitution is made at the discretion of the LEA. Secondly,

an LEA can opt into a rule where a TVAAS level of four or greater can override an LOE of three or lower. (Tennessee Department of Education, 2015)

Assignment of Observations

In this chapter I briefly explain the assignment of observations. I return to this discussion in Chapter 3 when discussing the predictor of interest and instrument used in regression discontinuity designs. There are three broad factors affecting the number of observations teachers receive each year: certification status, lagged LOE, and educator discretion. Broadly, "certification status" identifies whether a teacher has taught for less than four years or more than three years. For most teachers, TBOE assigns teachers with a lagged LOE1 four classroom visits and LOE5 one classroom visit. TBOE policy assigns teachers with LOE2 – LOE4 four or two classroom visits depending on their certification status. TBOE assigns a minimum number of observations, but districts/ schools/ teachers can add to these minima.

Expected Observation Practices

TBOE policy also describes expected observation practices with regard to the time spent on an observation, frequency of rating domains, combinations of domains observed during a single observation, and the number of domains scored during a semester. First, no matter how many observations a teacher receives per year, TBOE policy generally expects each observation (i.e. the classroom visit) to take approximately 15 minutes (Tennessee Board of Education, 2013). The remaining three expectations depend on the minimum number of observations assigned to a teacher.

The minimum number of policy-assigned observations assigned to a teacher affects how often an observer should rate a lesson with respect to the Instruction, Environment, and Planning domains. TDOE expects observers to rate LOE1 teachers at least three, two, and two times relative to the Instruction, Environment, and Planning domains, respectively. Teachers observed at least twice should be observed at least two, one, and one times with respect to the Instruction, Environment, and Planning domains, respectively. Teachers receiving a single observation should be scored on each of these three domains during that visit. Teachers are also scored with respect to a Professionalism rubric, but this part of their evaluation does not involve a classroom visit or post-observation conference. (Tennessee Board of Education, 2013)

Finally, there are expectations regarding the number of domains observed per semester. Teachers assigned a minimum of four observations must be observed with respect to all three domains each semester, while teachers assigned a minimum of two observations should be observed relative to at least two domains per semester. Teachers assigned a minimum of one observation are observed relative to all three domains in the first semester. (Tennessee Board of Education, 2013)

Summary

The TEAM teacher evaluation system aims to provide educators with information that will improve teacher performance. Classroom observations and subsequent post-observation feedback are the linchpins to teacher improvement within the Tennessee teacher evaluation system. The TEAM observation system is designed to improve teacher performance via frequent classroom observations conducted by TDOE certified observers and timely post-observation

feedback. Most importantly, a TEAM observation and post-observation conference should result in clear, actionable next steps for educators to follow on the path towards improvement.

In the next chapter, I discuss relevant literature with respect to the following: reasons why observations should improve teacher performance, challenges to the implementation of evaluation systems, and potential heterogeneity in the effects of observations on teacher performance.

CHAPTER 2

DISCUSSION OF RELATED LITERATURE

Introduction

This discussion of related literature first describes the theory of action undergirding modern teacher evaluation systems. This theory of action formalizes the conventional wisdom that more observations per year should improve teacher performance. I then discuss some characteristics of modern teacher evaluation systems that could inhibit the effectiveness of observations as a tool for improving teacher performance. I return to some of these inhibitors when discussing implications in the last chapter. The final section frames the examination of heterogeneous effects described in later chapters.

Modern Teacher Observation Systems: A Theory of Action

One distinguishing feature of modern teacher observation systems is the adoption of standards-based protocols (Steinberg & Donaldson, 2016). Theoretically, the adoption of these protocols (i.e. observation rubrics) should lead to higher student achievement because these rubrics describe teacher/ student behaviors prior work has linked to student learning. Observers should rate lessons/ teachers with respect to behaviors described in these rubrics, then share feedback with teachers. The provision of this feedback should highlight why the teacher did not exhibit exemplary behavior and provide information the teacher, or teacher and observer, can use to move teacher/ student behaviors towards those behaviors linked with higher student

achievement. If these things happen, more frequent observations should improve teacher performance as measured by student achievement scores.

It seems the TEAM teacher observation system was built with this theory of action in mind. Prior work has linked the teacher/ student behaviors described in the TEAM rubric to higher student achievement scores (Daley & Kim, 2010; Danielson Group, n.d.). TEAM observers should provide teachers with feedback during post-observation conferences so the teacher, or teacher and observer, can use the feedback to improve. Moreover, TDOE expects more frequent observations to improve subsequent observation scores, student achievement, and TVAAS (i.e. student growth) (Alexander, 2016; Tennessee Department of Education, 2016). While it is possible classroom observations can generate information used exclusively for personnel decisions (e.g. retention), TBOE emphasizes the use of observations as a teacher performance improvement tool (Tennessee Board of Education, 2013).

Widespread Challenges to the Effectiveness of Classroom Observations

Despite the use of observations as a tool to improve teacher performance, there are at least four widespread challenges to the effectiveness of classroom observations as a formative evaluation tool (i.e. tool to improve teacher performance). I consider a challenge "widespread" if it is likely to affect the typical observer and/ or teacher. Two of these challenges are systemic. Observers, the majority of whom are school administrators, do not have enough time to effectively manage the teacher evaluation process (Kraft & Gilmour, 2016a), and it is challenging for educators to match teachers needing formal professional development to appropriate professional learning opportunities (Curtis & Wiener, 2012; Weisberg et al., 2009). The other two challenges concern observer expertise: prior work suggests the typical classroom

observer lacks expertise in the facilitation of post-observation conferences and in the teacher's content area (Kraft & Gilmour, 2016a).

Modern teacher evaluation systems require substantially more time to manage than previous systems, which could make it difficult for school administrators to provide effective formative evaluation. In the early 2010s, teacher evaluation systems underwent substantial changes, including increasing the frequency of observations (Georgia Department of Education, 2012; Kraft & Gilmour, 2016b; Tennessee Board of Education, 2013). This change alone would increase the time school administrators spend on observations, *ceteris paribus*. Prior work also implies principals responded to these time demands by engaging in satisficing behaviors, such as conducting brief observations (Halverson, Kelley, & Kimball, 2004; Sartain et al., 2011) and brief post-observation conferences (Kimball, 2003; Kraft & Gilmour, 2016a). Either of these satisficing behaviors could impede the success of observations as formative evaluation tools.

The second widespread challenge concerns teacher professional learning opportunities, an ostensibly integral component of formative evaluation for some teachers. Some teachers may not need anything more than feedback to improve, but others may require subsequent professional development. Practitioners note teacher evaluation systems do a poor job matching teachers needing professional development with appropriate professional learning opportunities (Curtis & Wiener, 2012; Weisberg et al., 2009). This poor matching would likely inhibit the effectiveness of observations for some teachers. Moreover, research suggests teachers are less likely to act on post-observation feedback if they do not believe they can access appropriate professional learning opportunities in response to this feedback (Cherasaro, Brodersen, Reale, & Yanoski, 2016). At least two state educational agencies have recognized this problem within modern evaluation systems and have attempted to address it by supporting flexible, context-

specific, and ongoing school-based forms of professional development (Georgia Department of Education, 2012; Tennessee Department of Education, 2018).

The other two widespread challenges concern observer characteristics. Since the typical classroom observer is a school administrator (Kraft & Gilmour, 2016a), these two challenges amount to challenges concerning school administrator characteristics. For decades, research outside education has suggested observer expertise regarding an employee's core work is positively related to improvements in employee performance (Ilgen, Fisher, & Taylor, 1979). Observers with expertise in the employee's work are better equipped to provide reliable observation scores and more useful feedback (Ilgen et al., 1979; Levy & Williams, 2004). The provision of such scores and feedback is ostensibly important to formative evaluation. Research suggests employees are more likely to dismiss observation scores and post-observation feedback if they believe scores are inaccurate and feedback is not useful (Cherasaro et al., 2016; Jawahar, 2010). Education researchers have also found that classroom observers with expertise in the content area of the teachers they observe are better able to support teacher improvement (Hill & Grossman, 2013). Some research reports even principals believe their lack of content expertise makes it difficult for some teachers to accept their feedback (Kraft & Gilmour, 2016a). It is practically impossible for the typical observer/ school administrator to possess content expertise relative to all the classrooms she observes. In conjunction with research concerning the importance of observer expertise, the limited expertise of the typical classroom observer challenges the successful formative evaluation of teachers.

Ineffective facilitation of post-observation conferences is the fourth widespread challenge to the effectiveness of classroom observations. A meta-analysis by Kluger and DeNisi concerning the relationship between provision of post-observation feedback and employee

performance found mixed results (Kluger & DeNisi, 1996). While prior work and conventional wisdom suggest post-observation feedback should improve employee performance, the Kluger and DeNisi meta-analysis found there were circumstances under which feedback worsened employee performance (1996). Specifically, there was a negative relationship between the provision of post-observation feedback and employee performance when the feedback was mostly positive and the employee's core work was complex (negative feedback identifies an employee's area of weakness). Teaching is arguably complex and recent research finds principals report being uncomfortable providing teachers with negative feedback (Kraft & Gilmour, 2016a).

Although new observation systems include standards-based protocols, more frequent observations, and post-observation conferences (Alexander, 2016; Tennessee Department of Education, 2016), there are widespread challenges to the effectiveness of observations as a tool to improve teacher performance. Despite these widespread challenges research also suggests more observations can benefit some teacher subgroups more than others (i.e. heterogeneous effects exist).

Heterogeneity in the Effectiveness of Observations

In this section I discuss some teacher and school characteristics research suggests moderates the effectiveness of observations as a formative evaluation tool. It is important to discuss and examine heterogeneous treatment effects for at least two reasons. First, heterogeneous effects would imply SEAs/ policymakers could strategically reallocate the number of observations conducted by observers across teacher subgroups. Subgroups benefitting from observations could receive relatively more observations. Second, SEAs could target supports for

observers working with subgroups of teachers for whom observations are not relatively beneficial.

Prior work implies the relationship between the provision of post-observation feedback and changes in teacher performance depends in part on the presence of enabling work conditions. Specifically, researchers outside education find workplaces in which employees report receiving higher quality feedback or believe evaluation systems are formative have stronger, positive relationships between observations and employee performance (London & Smither, 2002). Similarly, teachers working in schools where colleagues approve of the implementation of the evaluation system report they are more likely to use information generated by the evaluation system to improve performance (Sun, Mutcheson, & Kim, 2016).

Prior research also suggests grade levels are a potential moderator of the relationship between observations and teacher performance. In the previous section, I discussed the importance of observer expertise. In the early 2000s, Kimball found high school teachers were especially likely to dismiss feedback generated by an observer without content expertise (Kimball, 2003). Perhaps high school teachers believe there is a wider gap between their content knowledge and the knowledge of the typical observer. Regardless, Kimball's work implies observations may be a more effective tool for formative evaluation in earlier grade levels.

The last potential moderator is a teacher level moderator: years of experience. Relative to mid- and late-career teachers, new teachers report they are more likely to perceive feedback provided by their observer as credible (Kimball, 2003). Psychological research confirms the conventional wisdom that if an employee does not perceive feedback as credible, she is less likely to productively respond to the feedback (Ilgen et al., 1979; Jawahar, 2010).

Summary

While evaluation/ observation systems can inform personnel decision-making and formative evaluation, it seems modern teacher observation systems focus on the latter. More importantly, the TEAM theory of action implies more frequent observations should improve measures of student achievement. However, there are widespread challenges that may inhibit the effectiveness of observations as a formative evaluation tool. At the same time, prior work suggest observations may be especially beneficial for certain teacher subgroups, which could allow teachers in these subgroups to overcome these widespread challenges.

CHAPTER 3

METHODOLOGY AND DATA

Introduction

In this chapter I discuss observation cycles in greater detail, and methods and data used to answer my research question. In brief, I draw on Tennessee Department of Education administrative data from 2012-13 through 2014-15 and use a 2SLS local regression discontinuity design (RDD) to generate my main findings.

The main findings concern the relationship between the number of observations received per year (sometimes referred to as "frequency of observations") and contemporaneous teacher performance. I also estimate longer-term and heterogeneous effects. It is plausible it takes time for observations to affect teacher performance, while prior research suggests teachers may respond to observations in different ways. The outcomes of interest are teacher level mean student achievement scores in math and reading/ language arts (RLA) and teacher value-added scores (TVAAS). As discussed in Chapters 1 and 2, the TEAM theory of action asserts educators will use feedback based on the standards-based TEAM observation rubric to move teacher/ student behaviors towards rubric-defined exemplary behaviors. Prior work has linked these exemplary behaviors to higher student achievement. Thus, more frequent observations should improve student achievement scores. I use teacher level mean (TLM) student achievement as the outcome of interest when identifying the effect of observations on student achievement. Since additional observations are expected to raise student achievement, more observations should also

affect value-added measures. I explore this by identifying the effect of more frequent observations on TVAAS scores.

One might expect that undergoing more observations would also increase scores teachers receive when observed again in the future. I do not discuss this relationship any further because there is strong evidence observers score teachers differently as a function of the number of observations they have been assigned. This becomes confounded with the causal impact of observations on teacher performance when observation scores are employed as an outcome of interest. Appendix C documents the existence of observer bias and explores some possible explanations.

In the remainder of this chapter I discuss the assignment of observations, models, and data characteristics.

Policy Assignment of Observations

According to state policy, two variables determine the minimum number of observations a teacher is supposed to undergo in a given year: prior-year LOE and teacher certification status—i.e., whether a teacher holds Apprentice or Professional certification[8]. The primary difference between Apprentice and Professional teachers is years of experience: the former has

[8] All teachers new to Tennessee traditional public schools receive an initial certification lasting three years. Certifications permit teachers to work in the Tennessee public education system as classroom teachers. TDOE should revoke the license of an Apprentice if she does not successfully complete an approved educator preparation program during her initial three years. Teachers who complete an educator program but do not advance to a Professional certification at the end of the initial three-year period retain their Apprentice certification status. If an Apprentice does not advance to Professional status by the end of the second three-year cycle her certification is non-renewed. Tested and non-tested teachers may advance from Apprentice to Professional status if they have not received a LOE1 (TVAAS1 for tested teachers) during any of the three years prior to their advancement request. Unlike an Apprentice certification, a Professional certification lasts six years. TDOE renews a Professional certification if a teacher has not received an LOE1 (or TVAAS1 for tested teachers) in any of the three years preceding her renewal request, otherwise the Professional teacher becomes an Apprentice. Classroom teachers with a Professional certification who do not meet performance expectations (i.e. assigned to LOE1 or tested teachers assigned to TVAAS1) should be placed on a one year "review status" during which they are evaluated as though they are a first-year educator. If a Professional teacher on review status does not receive LOE1 (or TVAAS1 for tested teachers) at the end of their review status year, she retains her Professional status, otherwise the certification is non-renewed. Policy assigns tested Apprentice LOE2 – LOE4 (i.e. LOE-cont [200, 425)) teachers a minimum of four observations per year, and tested Professional teachers assigned to LOE2 – LOE4 a minimum of two observations. (Tennessee State Board of Education, 2013a)

less than four, the latter four or more. State policy requires LOE5 (i.e. LOE-cont $\geq$ 425) teachers undergo a minimum of one observation and LOE1 (i.e. LOE-cont < 200) a minimum of four. Rules governing the minimum number of observations for LOE2 – LOE4 (i.e. LOE-cont $\geq$ 200 and LOE < 425) depend on certification status. Policy assigns an LOE2 – LOE4 Apprentice (Professional) teacher a minimum of four (two) observations. Districts or schools may adopt policies adding to these state-assigned minima. (Tennessee Department of Education, n.d.-b)[9]

Educator Compliance with TBOE Assignment of Observations

Table 1 displays the distribution of observations received by the population of teachers with different certification-LOE combinations. Percentages represent the percentage of teachers with a certain certification and LOE (e.g. 72.54% of Apprentice teachers with a LOE1 received four classroom visits) receiving some number of observations. The percentages within each box total to 100. Bold cells indicate the minimum number of state-prescribed classroom visits a teacher should receive.

Table 1 shows there is acceptable and unacceptable non-compliance from the standpoint of state policy. Educators can add to the state minimum number of observations, but teachers should never receive less than the minima. The largest degree of unacceptable non-compliance exists for Apprentice teachers with an LOE2 – LOE4: almost 25% of these teachers receive less than the expected four classroom visits. Roughly 10% (14%) of Professional (Apprentice) teachers with an LOE1 receive less than the assigned four observations. Unexpectedly, a sizeable

[9] As noted in Chapter 2, some teachers of tested subjects may have their LOE determined entirely by their TVAAS scores and not by LOE-cont. For most of the analyses that follow, the estimation sample drops such teachers. Even when they are retained, LOE-cont remains strongly predictive of the number of observations a teacher undergoes.

minority of teachers with an LOE5 receive more than the state prescribed minimum of one observation: approximately 20% of these teachers receive two observations.

There are at least three reasons why teachers may not receive the policy-assigned number of observations. First, there is some ambiguity regarding what is to count as an observation. Some administrators may define an observation in terms of domains rated. For example, if an observer rates a teacher with respect to the Instruction and Environment domains during a single classroom visit, some interpret this as two "observations." Second, districts/ schools may choose to evaluate some teachers more frequently than is required. Third, school administrators may deviate from state-assigned minima for reasons of their own.

Non-compliance regarding the number of observations received is plausibly endogenous. There are several potential sources of endogeneity. Teacher motivation is plausibly related to teacher performance and the number of observations received (i.e. treatment). School administrators may observe less motivated teachers more often to closely monitor teacher behaviors, negatively biasing estimated effects. Alternatively, school administrators may observe more motivated teachers more often because these teachers are receptive to feedback, easing the observation process for administrators. Yet, the performance of these teachers will likely improve independent of observations received, introducing positive bias. Unobserved school administrator effectiveness is another potential source of bias. Effective school administrators likely improve teacher performance independent of treatment (e.g. positive work climate, adopting academic programs positively influencing student growth). It is plausible these administrators can improve teacher performance via observations using less than the policy-assigned number of observations, positively biasing estimates. In a final example, endogeneity may also exist because of student behaviors/ characteristics. One year a teacher may be assigned

a group of students whose poor behavior affects teacher performance. School administrators may observe this teacher more often to help her devise strategies to mitigate poor student behavior, negatively biasing estimates.

Methodology

To identify the relationships of interest I employ 2SLS local RDDs. I use plausibly exogenous, local variation in lagged LOE-cont on either side of the 200 (425) threshold as an instrument to predict the number of observations received in year *t*. There are four instruments: whether an Apprentice teacher lies to the left or right of the 200 threshold; whether an Apprentice teacher lies to the left or right of the 425 threshold; and two analogous instruments for Professional teachers.

I use the following model when the outcome is TVAAS:

$$(1)\quad y_{ijt} = \beta_0 + \delta\widehat{obs}_{ijt} + \boldsymbol{A}f(\cdot) + \boldsymbol{BX}_{ijt} + \boldsymbol{CS}_{jt} + \gamma_t + e_{ijt}, \qquad \left|LOE_{ijt}\right| \leq w$$

$$(2)\ obs_{ijt} = \ddot{\beta}_0 + \boldsymbol{\theta}g(\cdot) + \ddot{\boldsymbol{A}}f(\cdot) + \ddot{\boldsymbol{B}}\boldsymbol{X}_{ijt} + \ddot{\boldsymbol{C}}\boldsymbol{S}_{jt} + \gamma_t + u_{ijt}, \qquad \left|LOE_{ijt}\right| \leq w$$

where y_{ijt} represents the TVAAS score of teacher *i* in school *j* in year *t*. obs_{ijt} is the number of observations received in year *t*, and $\widehat{obs}_{it}$ the predicted number of observations from the first stage equation. *f* is a second order polynomial of the running variable interacted with teacher certification status (i.e. Professional/ Apprentice), and *g* the vector of four instruments. $\boldsymbol{X}_{ijt}$ is a vector of covariates including teacher race/ ethnicity, certification status, gender, years of teaching experience, certification status, and level of education. $\boldsymbol{S}_{jt}$ is a vector of school level measures controlling for the distribution of teacher effectiveness in school *j* in year *t*, including the mean, standard deviation, and skewness of LOE-cont. γ_t is a year fixed effect, and e_{ijt} and

u_{ijt} are idiosyncratic error terms. A 2SLS estimator produces estimates for cases where $|LOE_{ijt}|$ is less than or equal to bandwidth *w*. I include teacher- and school level controls because they may have an independent influence on the outcomes of interest and should increase the precision of estimates. The local average treatment effect (LATE) of interest, δ, represents the effect of an additional observation per year on TVAAS scores[10]. Year fixed effects account for secular trends in observational practices.

To estimate the relationship between observations and changes in student achievement I use grade-subject standardized math and reading/ language arts (RLA) scores from students in grades 4 – 12. These models differ from equations 1 and 2 in three ways. First, these two outcomes are the teacher level mean (TLM) math or RLA achievement scores for students taught by teacher *i* in year *t*. Second, I add a fourth order polynomial of the average lagged score[11] of the same group of students as a control. Stated differently, the lagged polynomial is composed of the mean achievement score measured in year *t* – 1 that is associated with the group of students taught by teacher *i* in year *t*. Third, I add more controls to the vector $\boldsymbol{X}_{ijt}$. This vector now includes the proportion of students taught by: race/ ethnicity, free/ reduced price lunch status, ESL status, gender, and immigrant status. All other controls and model specifications in equations 1 and 2 remain unchanged. I add the new controls because they increase the precision of δ and may have an independent influence[12] on TLM math or RLA achievement scores.

[10] LATEs produced by RDDs assume linearity. For example, the LATE of going from two the three observations is the same as going from four to five. However, the effect of an additional observation may be non-linear (i.e. depend on number of observations received). I return to this point after discussing my main findings.

[11] If a student taught by teacher *i* in year *t* did not have an achievement score in year *t* – 1, the student was not included in the analysis.

[12] TVAAS scores use multiple years of prior achievement data, but do not include student or contextual controls (SAS, 2015). Theoretically, these controls are not needed to estimate unbiased teacher effects given the use of multiple years of prior achievement data, however, prior empirical work finds TVAAS estimate are sensitive to the inclusion of these controls in some circumstances (McCaffrey, Lockwood, Koretz, & Hamilton, 2003). There are at least two reasons why the sensitivity of TVAAS scores does not threaten the internal validity of my models. When RDD assumptions are met, controls are not needed to address bias. To the extent these assumptions are not met, the sensitivity of TVAAS scores could only bias estimated treatment effects if treatment is correlated with the circumstances affecting the sensitivity of TVAAS scores. There is no reason to believe this is the case. While student controls could be omitted from TLM models on these same grounds, these models use much smaller samples, decreasing statistical power. I offset some of the loss in power by using student controls in TLM models.

I examine TLM and TVAAS scores for three reasons: robustness checks, power, and policy-relevance. By controlling for lagged TLM scores in models where the outcome is TLM math/ RLA scores, the model is effectively measuring the achievement gains of the typical (i.e. mean) student. Models in which the outcome is TVAAS (with no lagged outcome) also effectively measure gains in student achievement. However, estimates produced by these models are based on different samples. Briefly, TVAAS scores use data from all students taught by a teacher. But, some students receive all content instruction from one teacher, others do not (e.g. 20% of a student's math instruction may come from one teacher, 80% from a second). TLM samples are restricted to the former. For reasons described below, this restriction removes some ambiguities about the receipt of treatment. Second, more than just math/ RLA teachers receive TVAAS scores. This, substantially increases the size of TVAAS samples. Third, in the Tennessee study context, TLM and TVAAS are two of the most policy-relevant measures of student growth.

Longer-Term and Heterogeneous Treatment Effects

My main findings concern the extent to which the number of observations a teacher receives per year affects her contemporaneous performance. However, there are other policy relevant effects, namely, longer-term and heterogeneous effects.

Longer-Term Effects

Prior research and conventional wisdom implies observations should affect contemporaneous employee (i.e. teacher) performance (Guerin, 1993; Kraft & Gilmour, 2016a; Murphy & Cleveland, 1995). However, the effects of observations conducted during an

academic year may not materialize until subsequent academic years (extended effect). It is also plausible that observations do not affect teacher performance until teachers receive some accumulated number of observation cycles over time (cumulative effect).

Extended and cumulative effects are especially plausible if educators use post-observation feedback to identify areas of weakness, then engage in in/ formal professional development to improve performance. Implicitly, it takes time for this process to affect teacher performance. Therefore, teachers may not have time to appreciably improve their performance by the end of the school year during which they received their observations. To explore extended and cumulative effects I modify equations 1 and 2.

I estimate the extended effect using the following RDD:

$$(3)\quad y_{ijt} = \beta_0 + \delta\widehat{obs}_{ij,t-1} + \boldsymbol{A}f(\cdot) + \boldsymbol{BX}_{ijt} + \boldsymbol{CS}_{jt} + \gamma_t + e_{ijt}, \qquad \left|LOE_{ij,t-1}\right| \leq w$$

$$(4)\ obs_{ij,t-1} = \ddot{\beta}_0 + \boldsymbol{\theta} g_{t-1}(\cdot) + \ddot{\boldsymbol{A}}f(\cdot) + \ddot{\boldsymbol{B}}\boldsymbol{X}_{ijt} + \ddot{\boldsymbol{C}}\boldsymbol{S}_{jt} + \gamma_t + u_{ijt}, \qquad \left|LOE_{ij,t-1}\right| \leq w$$

The key difference between equations 3 and 4, and equations 1 and 2, is the year during which observations and the running variable were measured. In the original models, the number of observations a teacher received per year and outcomes of interest were both measured in year *t* (i.e. the predictor over the year, outcomes after completion of all observations). In equation 1, I predicted the number of policy-assigned observations a teacher received using certification status from year *t* and lagged LOE-cont (i.e. LOE measured in year *t* – 1). In equations 3 and 4, the predictor of interest and instruments are lagged by an additional year, but the outcomes are not. The instruments are certification in *t* – 1 and LOE-cont measured in year *t* – 2, and the predictor of interest is the number of observations received in year *t* – 1. All controls and outcomes from equations 1 and 2 remain unchanged. Thus, in equations 3 and 4, δ represents the effect of an additional observation per year on the outcomes measured one year after treatment.

I use a full-sample RDD (i.e. RDD not restricted to local bandwidths) instead of the local RDD to estimate the effects of the total number of observations received through year *t* (i.e. cumulative observations) because few teachers remain in the local RDD bandwidths over time. To estimate the cumulative effect, I use the following model:

$$(5)\quad y_{ijt} = \beta_0 + \delta\widehat{c\_obs}_{ijt} + \boldsymbol{A}\check{f}(\cdot) + \boldsymbol{BX}_{ijt} + \boldsymbol{CS}_{jt} + \gamma_t + e_{ijt}$$

$$(6)\ c\_obs_{ijt} = \ddot{\beta}_0 + \lambda c\_pol_{ijt} + \ddot{\boldsymbol{A}}\check{f}(\cdot) + \ddot{\boldsymbol{B}}\boldsymbol{X}_{ijt} + \ddot{\boldsymbol{C}}\boldsymbol{S}_{jt} + \gamma_t + u_{ijt}$$

Equations 5 and 6 include two new variables: c_obs_{ijt} and c_pol_{ijt}. The cumulative observations received (c_obs_{ijt}) and policy-assigned observations (c_pol_{ijt}) represent the number of observations received since the 2011-12 school year and number of policy-assigned observations assigned over this same period, respectively. The number of cumulative policy-assigned observations assigned to a teacher is based on the TDOE observation assignment schedule[13]. Thus, the full-sample RDD isolates exogenous variation in cumulative observations received using policy-assigned cumulative observations.

Equations 5 and 6 also include previously seen variables: y_{ijt}, $\boldsymbol{X}_{ijt}$, $\boldsymbol{S}_{jt}$ and γ_t refer to the same quantities as in previous equations. $\check{f}$ is a higher order LOE polynomial interacted with certification status. As a full-sample RDD, it is important $\check{f}$ is correctly specified. I added higher degrees of LOE-cont until the last term added was insignificant, which happened after adding the cubic term. Thus, $\check{f}$ is a second order polynomial[14].

[13] I assign teachers of tested subjects fewer observations when there are discrepancies between their discrete lagged LOE and TVAAS.
[14] I estimated full-sample RDDs using up to a fourth-degree polynomial. Results are insensitive to higher order polynomials.

Heterogeneous Effects by Teacher and School Characteristics

Research reviewed in the *Heterogeneity in the Effectiveness of Observations* section implies the effectiveness of observations may depend on some teacher, observer, and school characteristics. Specifically, the reviewed literature suggests there may be heterogeneous effects with respect to: teacher years of experience; teacher perceptions about the utility and fairness of the evaluation/ observation system and perceptions about observer expertise; and grade levels (e.g. elementary, middle).

It is also plausible that there are heterogeneous effects with respect to school administrator effectiveness. Prior research finds the effectiveness of observations as tools for formative evaluation depend on observer expertise with respect to content (Hill & Grossman, 2013) and facilitation of post-observation feedback sessions (Cherasaro et al., 2016). While TDOE does not directly measure these constructs, it collects multiple measures of school administrator effectiveness. I also explore heterogeneous effects with respect to these measures of administrator effectiveness (for more information these measures see Appendix H).

Data

This dissertation uses TDOE administrative and survey data from 2012-13 through 2014-15. I use data from the Tennessee Educator Survey[15] (TES) in robustness tests and to estimate some heterogeneous effects.

To construct the predictor of interest I draw on a rich set of TEAM observation data from 2012-13 through 2014-15. Associated with each observation record are unique teacher, observer,

[15] The annual survey administered by TDOE is now the "Tennessee Educator Survey." During the study period this survey was called the "Race to the Top Survey" and the "First to the Top Survey."

and school identifiers, and observation event characteristics including whether the observation was unannounced/ announced/ a walkthrough, and the date on which observers entered these data into the TDOE administrative data system. These data include 45 duplicated records. I retain one copy of each duplicate. I calculate the number of observations a teacher receives within a school year using the total number of these unique records per teacher.

Control variables include teacher demographics/ characteristics and evaluation scores. Teacher demographic data come from 2012-13 through 2014-15 certification files, and survey and staff files from the same academic years. Professional certification files contain certification information (i.e. Professional/ Apprentice). TDOE staff[16] files include teacher: years of experience, education level, race/ ethnicity, and gender. Administrative data include lagged and then-current achievement, growth, and observation scores from 2012-13 through 2014-15; these data are used to calculate LOE-cont. Administrative data also include the TVAAS scores, summative TEAM observation scores for all teachers, and unique teacher and school identifiers.

I combine student data with the aforementioned datasets using unique student and teacher identifiers[17]. Test score data include scaled math and RLA accountability test scores in grades three through eight, and high school end of course assessments in English I, II, and III, and algebra I and II. I standardize achievement scores by grade-subject. Student data also include multiple variables measuring student demographics/ characteristics.

Some robustness and moderating analyses use TES data. All Tennessee teachers receive the TES in late spring of each academic year. Response rates exceeded 50% during the study period.

[16] Some teachers were missing demographic data. If these teachers took the TES they often supplied demographic information. In cases where staff file information was missing and TES demographics were not I overwrote the former with the latter.
[17] All analyses exclude students who took alternative achievement tests designed for students with exceptional needs. This does not mean analyses exclude special education students. Special education students can take the standard achievement test.

Sample Restrictions

Analytical samples are restricted in different ways for different analyses. I exclude teachers whose LOE-cont scores do not place them in the expected LOE (e.g. a record shows an LOE-cont of 420 but LOE5) from the sample. Over the study period there were 1730 such records, or 0.9% of the population of evaluated Tennessee teachers. Discrepancies between a discrete and LOE-cont can exist for at least three reasons: override rules (e.g. 3/4/5 or 4/5 overrides), appeals, or clerical errors[18].

I restrict analytical samples used to estimate the relationship between observations and TLM scores in ways that do not apply to TVAAS samples. I discard students if a single teacher did not claim 100% of the student's math or RLA instructional time, and retain high school students only if their end-of-course (EOC) exam occurred in the spring. I make both restrictions due to concerns about receipt of treatment.

Due to data limitations, it is ambiguous when some students were assigned to a given teacher, so it is unclear how many observations a teacher received when teaching certain students. Teachers of tested subjects must claim some percentage of content (e.g. math, RLA) instructional time for each student they teach. If 40% of Student X's RLA instruction comes from Teacher A and 60% from Teacher B, Teachers A and B would claim 40% and 60% RLA instructional time, respectively. Such claims could mean Teacher A taught Student X for the first 40% of the school year and Teacher B taught Student X the remainder of the year. However, it is not necessarily the case that the time over which a teacher claims a student's instructional time

[18] Districts can adopt override rules allowing teachers of tested subjects with growth scores higher than their LOE to override their LOE with the higher growth score. It is also possible for teachers to appeal for a higher LOE on the basis that something about their evaluations were conducted incorrectly. Finally, there are bound to be some errors due to clerical oversight or programming.

be unbroken. Student X could start and end the year with Teacher A, while receiving instruction during the middle of the year from Teacher B. Since I do not know when a student taught by multiple teachers in the same content area was taught by a particular teacher, I do not know how many observations the teacher received while the student was assigned to them. For example, if a teacher claimed 20% of a student's content instructional time, this does not necessarily mean the teacher received 20% or 80% of her observations while the student was assigned to them. Not knowing when a student was assigned to a teacher clouds the relationship between treatment, teacher behaviors, and student learning. I forego making potentially dubious assumptions about these relationships and drop students if they did not receive 100% of their content instruction from a single teacher. Of the approximately 2,412,000 (3,011,000) records associated with grades 4-12 math (RLA) students during the study period, approximately 391,000 (1,016,000) records were dropped because of this sample restriction.

I drop high school students who did not take spring EOC exams for similar reasons. Nearly all students take an EOC exam at the end of fall or spring. Fall administrations tend to occur near the mid-point of the academic year. Thus, high school students taking fall EOC exams spend at least the first half of an academic year with one content teacher (i.e. one teacher because of the previous sample restriction). Ideally, I could tabulate the number of observations received by high school teachers of fall EOC test-takers through the fall semester of each study year to determine how many observations these teachers received to that point. Unfortunately, I cannot precisely determine when teachers received their observations within the academic year. TDOE administrative data include "observation dates," but I have learned these are the dates on which observers entered these data into the TEAM system, not dates on which observations necessarily occurred. This raises questions about the number of observations teachers of fall test-takers

received prior to the exam[19]. I therefore discard high school students who took their EOC exam in the fall. This sample restriction resulted in dropping an additional 64,000 (133,000) fall EOC test-takers in math (RLA) records over the study period.

Since I know more about the relationships among observations received, time a teacher spent with a student, and EOC test scores, I check the sensitivity of my main findings to this last sample restriction, retaining all high school EOC fall test-takers. Results produced by this less restrictive sample are statistically indistinguishable from results produced by the restricted sample.

Descriptive Statistics

Table 2 includes some descriptive statistics for the sample used in a bandwidth of 40 (i.e. the largest bandwidth used, see Chapter 5 for details) when the outcome is TVAAS scores. In this sample the typical teacher is a white female, holds more than a BA/ BS degree, and has approximately 13 years of experience. The typical TEAM score is 4.07 on a scale of [1, 5] and the standard deviation of the annual change in TEAM scores is 0.40. The average TVAAS score is 2.12, meaning the typical teacher in my sample contributed 2.12 normal curve equivalents to her students' achievement compared to what these students would have scored if they were taught by the hypothetical average Tennessee teacher (SAS, 2016). Moreover, a TVAAS score of 2.12 represents an approximate gain of 0.10 on the scale of standardized student achievement scores. The standard deviation of the annual change in these TVAAS is 6.15 normal curve equivalents, or approximately 0.29 on the scale of standardized student achievement scores.

[19] For example, suppose a teacher receives four observations over the course of the school year and administrative data suggest the second observation occurred January 10. Now suppose this teacher's students took an achievement test on December 10. Due to ambiguity in the timing of observations received, it is unclear if this teacher received 25% or 50% of treatment prior to the achievement tests administered on December 10.

Table 2 shows approximately 90% of the pooled local sample in a bandwidth of 40 includes teachers originally from the 425 threshold, and about 85% of this sample is Professional teachers.

Table 3 shows descriptive statistics for records when the outcomes are TLM math or RLA scores and the running variable is restricted a bandwidth of 40. The typical teacher in both samples resembles the typical teacher in the sample described in Table 2. Roughly 50% of a math/ RLA teacher's students are female, 14% are black, 77% are white, and 7% Hispanic. Slightly over half a teacher's students have FRPL status, 8% are ESL, and 2% have immigrant status. Again, almost 95% of the sample in a bandwidth of 40 when the outcomes are TLM math or RLA scores is Professional teachers near the 425 threshold.

Considering that a very large percentage of the analytical samples come from records surrounding the 425 threshold, it is useful to know the teacher population distribution of LOE. During the study period, LOE1, LOE2, LOE3, LOE4, and LOE5 teachers comprised < 1%, 9%, 22%, 33%, and 36% of the population of Tennessee teachers, respectively. Thus, almost 70% of Tennessee teachers are in LOE4 or LOE5.

Summary

I draw on multiple years of Tennessee Department of Education administrative data and use two-stage regression discontinuity designs to estimate the effects of an additional observation on: value-added (TVAAS) scores, and teacher-year level mean standardized student scores in math and reading/ language arts. I also estimate longer-term effects, and heterogeneous effects based on teacher and school characteristics.

In the next chapter, I present evidence regarding threats to internal validity.

CHAPTER 4

THREATS TO INTERNAL VALIDITY

Introduction

In this chapter I present evidence concerning the assumptions of a 2SLS RDD research design. I share findings regarding manipulation of the running variable and the imbalance of covariates at the 200 and 425 thresholds. I also present some evidence concerning the validity of the instruments.

Manipulation of the Running Variable

Manipulation of the running variable is a standard threat to causal inference in RDDs. The concern is that LOE-cont scores have been manipulated such that teachers possessing some confounding characteristic are strategically placed just to one side of the cut score in the running variable. For example, observers might wish to place teachers with historically high student achievement scores just above the 425 threshold regardless of what the observer sees during classroom observations. Given how LOE-cont is determined, such manipulation is practically infeasible. Observers would need a keen prescience concerning teacher growth and achievement levels to situate LOE scores just to one side of the LOE 200 (or 425) threshold since teachers do not receive their achievement and growth scores until the completion of all observations. Thus, observers wanting to manipulate LOE in this fashion would have to rely on historic measures of teacher performance to guess current year values. However, the polychoric correlation between growth (achievement) levels from year t and $t - 1$ is 0.50 (0.37). These conditions suggest it is

practically impossible for an observer to strategically place a teacher just to one side of a threshold.

Researchers often empirically test for manipulation. Conventional tests of manipulation identify relatively large discontinuities in the probability density function of the running variable as evidence of manipulation. However, given that the LOE-cont variable comprises three components, two of which take integer values, the probability density function of this variable exhibits spikes at multiples of five (see Appendix D). I therefore test for manipulation of the observation score assuming observers were prescient and knew teacher achievement and growth scores in advance, an unrealistic assumption. If there is no evidence of manipulation under this assumption, manipulation of LOE-cont under more realistic conditions is even more implausible[20].

I empirically test for manipulation using robust-bias correction approaches developed by Cattaneo, Jannson, and Ma (2016). There is no evidence observers systematically manipulated scores at the 425 threshold, and this finding is robust to the use triangular and epanechnikov kernel functions. Furthermore, there is no evidence of manipulation at the 200-threshold using the epanechnikov kernel function (robust-bias corrected p-value = 0.68). There is evidence of manipulation at the 200 threshold when using the triangular kernel function (p-value = 0.0001). However, it is important to remember this finding is predicated on the assumption of observer prescience. Even if one suspects there may have been manipulation at this threshold, there is no such evidence at the 425 threshold. For this reason, as well as others described below, I present results separately in which the analysis sample has been restricted to teachers in the neighborhood of the 425 threshold.

[20] I do this by dropping the achievement and growth LOE-cont components because these components only take integer values. I check for manipulation in the approximately continuous observation scores.

Covariate Balance

In an RDD it is important that there are no discontinuities in pre-treatment and time-invariant observables at the cut point (Morgan & Winship, 2007; Murnane & Willett, 2011). Such discontinuities raise concerns about discontinuities in unobserved confounding variables even after controlling for observables. For example, suppose teachers just below the 425-threshold have substantially lower prior TLM scores than teachers above the threshold. Prior TLM scores may be negatively related to teacher motivation, which almost certainly has a positive relationship with teacher performance (i.e. student growth). Indeed, school administrators may assign students with lower prior achievement scores to highly motivated teacher because of this relationship. However, school administrators may observe more motivated teachers less frequently, negatively biasing estimates.

Because variation in the instrumented number of observations received is based on discontinuities at the 200 and 425 thresholds for both Apprentice and Professional teachers, I conduct balance checks for each of these four groups. To conduct these tests, I use the equation:

$$(7)\quad x_{ijt} = \pi_0 + \ddot{\boldsymbol{\theta}} g'(\cdot) + \ddot{\boldsymbol{A}} f(\cdot) + \ddot{\boldsymbol{B}} \boldsymbol{X'}_{ijt} + \ddot{\boldsymbol{C}} \boldsymbol{S}_{jt} + \gamma_t + p_{ijt}, \qquad \left|LOE_{ijt}\right| \leq w$$

X' is the original ***X*** vector but does not include covariate *x*. All other variables refer to the same quantities described in previous sections. I estimate equation 7 once at each of the 200 and 425 thresholds, once for each of the five elements in x_{ijt} and lagged TVAAS, once each for Professional and Apprentice teachers, and once for each of the RDD bandwidths of 20, 30, and 40. This results in a total of 72 estimates. Balance tests at the 200 threshold suggest there are no systematic imbalances (see Table 4). However, there is evidence of a systematic imbalance in

teacher race at the 425 threshold for Apprentice teachers (see Table 5). To the extent this is evidence of bias, I estimate some effects using only Professional teachers at the 425 threshold.

I also check the balance of covariates from TLM models. In results not shown, tests at the 200 threshold showed substantial imbalance, weakening the claim that teachers on either side of the 200 threshold are equivalent. To the extent bias exists at the 200 threshold, I separate effects by threshold.

Table 6 displays results from balance tests at the 425 threshold using samples when the outcomes are TLM student achievement (156 more tests). At face value, there appear to be some discontinuities in the covariates at the 425 threshold. It appears Professional teachers just below the threshold have at least 1.5 fewer years of experience in the bandwidths of 20. However, outliers drive this discontinuity: a few Professional teachers above the 425 threshold have more than 50 years of experience. Table 6 includes a balance test using censored years of experience, in which I assign[21] years of experience greater than 10 a value of 10. There is no discontinuity in censored years of experience, so I conclude teacher years of experience does not pose a threatening imbalance. After ruling out the significant imbalance in years of experience, there are five more imbalances for Apprentice teachers and three for Professional teachers. To the extent the imbalance of Apprentice teacher covariates introduces bias, I estimate some samples using only 425-Professional teachers. The only significant result in Table 6 associated with 425-Professional teachers is a negligible imbalance in the proportion of a teacher's students assigned to ESL. When a Professional teacher is just below the 425 threshold, the proportion of her students who are ESL is 0.01 lower than Professional teachers just above 425.

[21] Prior work suggests returns to experience after 10 years on the job are relatively small (Harris & Sass, 2011; Papay & Kraft, 2013).

Validity of Instrumental Variables

If alternative treatments exist at the 200 (LOE1/ LOE2) or 425 (LOE4/ LOE5) thresholds, this would threaten the internal validity of the instruments. No other state policies exist at the 200 or 425 thresholds[22], but some state policies depend on other thresholds (e.g. LOE2/ LOE3 triggers teacher tenure). While policy-assigned threats are implausible, psychological effects could threaten the internal validity of my instruments. Each of these threats exist because crossing the 200 or 425 thresholds may induce behaviors affecting the outcomes of interest, independent of observations. There are three potential, threatening psychological effects. Teachers assigned to lower LOE may face an impetus to improve independent of the observation process ("impetus to improve"). Lower performing teachers, especially LOE1 teachers, may depart the profession via self-exit or dismissal, and knowledge of an impending departure may lead to demoralization, negatively affecting the outcomes of interest ("impending departure"). Finally, assignment to a higher LOE, particularly assignment to LOE5, may induce a substantial boost in teacher self-efficacy resulting in improved performance independent of the observation process ("psychological boost"). If these psychological threats exist, the first would positively bias the estimates of interest and the rest would induce negative bias.

Below I discuss some findings concerning the impetus to improve, impending departure, and psychological boost. There is no systematic evidence any of these psychological effects threatens the validity of the instruments. After presenting main findings in the next chapter, I present more results concerning the psychological boost, and present new findings from a generic falsification test for threatening psychological effects.

[22] While it is possible crossing the 200 or 425 threshold triggers district policies, I am unaware of any strong district incentives at these thresholds.

Impetus to Improve

Teachers assigned to lower LOE may face more of an impetus to improve than teachers in higher LOE, independent of the observation process. This impetus may exist due to socio-professional pressure (e.g. teachers in higher LOE have more prestige), to avoid punitive consequences, or other reasons. This impetus could cause LOE1 (LOE4) teachers to improve more than LOE2 (LOE5) teachers, inducing upward bias.

If such an impetus exists, evidence to that effect should appear in teachers' responses to items on the annual educator survey (the TES) asking about their efforts to improve. Teachers were asked to report the number of hours they spent in professional development focused on approximately ten different aspects of teaching (e.g. pedagogy, classroom management). I added these reported hours together, generating a sum of the total number of hours teachers reported engaging in PD (*PDhrs*). A second set of items asked teachers to list the number of times they collaborated (*tchcollab*) with other teachers for various purposes (e.g. improve instruction). An impetus to improve may have driven such collaboration. A third item asked about the total amount of time teachers spent preparing for classroom observations (*obshrs*). A fourth item asked a similar question concerning the amount of time teachers spent trying to improve their instruction (*insthrs*). Finally, fourteen items asked teachers about the extent to which they exerted more time or effort (*effortsum*) on various activities (e.g. lesson prep, reflecting on teaching). I dichotomized each of the 14 responses, then found the sum of the binary responses to produce the *effortsum* outcome. Appendix E contains these items and their original scales, descriptions of how I transformed items from their original scales and lists the years during which items were administered on the TES. Table 7 contains some descriptive statistics of these measures.

Using OLS[23] I regress these survey outcomes on: the four instruments and other covariates from equations 1 and 2. F-tests of the joint significance of the instruments are presented in Table 8. In a bandwidth of 20 (30) the instruments jointly predict *tchcollab* and *effortsum* (*obshrs*), but the instruments do not predict these outcomes in other bandwidths. Closer inspection reveals the 200-Professional (β[121.16], SE [33.096]) and 200-Apprentice (β[161.07], SE [53.458]) instruments positively predict *tchcollab* in a bandwidth of 20. The 200-Apprentice instrument negatively predicts *effortsum* in a bandwidth of 20 (β[-32.24], SE [10.244]). And, the 200-Professional and 425-Professional instruments positively predict *obshrs* in a bandwidth of 30 (respectively, β[1.18] SE [0.483] and β[0.21] SE [0.095]).

Considering the aforementioned threats to internal validity at the 200-threshold, I estimate new RDDs restricted to teachers in bandwidths surrounding the 425 threshold. Results from these models are in Table 9. The 425-threshold instruments do not jointly predict any of the impetus to improve outcomes.

There is little reason to believe the instruments are threatened by an impetus to improve. Instruments in the pooled sample (Table 8) do not predict survey outcomes across bandwidths. In the 425-threshold sample the instruments are not jointly predictive of any survey outcomes in any bandwidth. To the extent 200-threshold instruments are related to improvement efforts, I estimate some effects only using teachers surrounding the 425-threshold.

It is possible test results are sensitive to the operationalization of survey items (e.g. adding all PD items together). I explore the sensitivity of findings presented in this section to different operationalizations (see Appendix F). Sensitivity analyses produce qualitatively similar results.

[23] When treating survey outcomes as ordinal or multinomial there was no evidence the proportional-odds assumption was valid and multinomial logit models failed to converge.

Impending Departure

Assignment to a lower LOE (especially LOE1) may lead a teacher to quit or lead school administrators to dismiss the teacher. In either case, it is plausible that knowledge of an impending teacher departure/ dismissal could lead the teacher and/ or administrator to abandon teacher improvement efforts. Such abandonment could threaten the validity of the instrument and negatively bias subsequent estimates of the relationship between observations and teacher performance.

To explore the potential threat of an impending departure, I again use equation 2. I regress whether a teacher quits teaching (i.e. not present in administrative data the subsequent year) on the set of right-hand side variables from equation 2 (i.e. one-stage local RDD). The predicted outcomes and 95% confidence intervals (CI) produced by these local RDDs are represented in Figure 1. Figure 1 shows the 200-threshold point estimates exhibit expected patterns: LOE1 teachers are more likely to quit than LOE2 teachers. However, these point estimates are imprecisely estimated and statistically insignificant. There is little difference in the estimates of LOE4 and LOE5 teachers.

Psychological Performance Boost

LOE5 teachers may experience a substantial psychological boost enhancing their performance relative to LOE4 teachers. Although recent research finds teacher self-efficacy is not predictive of changes in student achievement (Jackson, Rockoff, & Staiger, 2014), earlier education research claims the opposite: enhancing employee self-efficacy is associated with teacher performance improvements (Hoy & Woolfolk, 1993; Raudenbush, Rowan, & Cheong,

1992). Thus, a boost in the self-efficacy of LOE5 teachers increase student performance for reasons independent of the observation process, introducing negative bias.

In this section, I present tests concerning the effects of crossing the 425 threshold on teacher perceptions about the observation/ evaluation system. The basis for a teacher's psychological boost is a score produced by the TEAM system. Research in social psychology finds people (i.e. teachers) typically perceive the world in ways reinforcing their self-image (Baumeister, 1998). I hypothesize that if the TEAM system brings about a psychological boost for teachers, these teachers will have a more positive view of the evaluation/ observation system. I characterize these tests as checks for "reinforcing perceptions."

Five TES items arguably measure some reinforcing perceptions. The first item measures whether a teacher believes evaluations will improve teaching (*imprvtch*). TES also asked teachers if their post-observation feedback was useful (*fbuseful*) and if their observers were qualified to conduct observations (*obsqual*). The fourth and fifth items asked teachers if they changed their teaching due to evaluations (*chngtch*) and if evaluations were fair (*faireval*). Responses to each of these items were originally on a four-point scale from Strongly Disagree to Strongly Agree. I dichotomized responses so Agree/ Strongly Agree became one and Disagree/ Strongly Disagree became zero (see Appendix E for details). None of these items directly measure the construct of interest (i.e. perceptions of the TEAM observation/ evaluation that reinforce positive self-image due to LOE5 assignment). But, it seems unlikely that teachers whose self-image depends on a perceived trait established by the TEAM evaluation system would have more negative views of this system than teachers in lower levels of effectiveness. To explore this hypothesis, I use tests similar to those in the *Impetus to Improve* section. Results

from these new tests are in Table 10. There is no evidence the instruments are threatened by a psychological performance boost.

Again, results from reinforcing perceptions tests could be sensitive to the operationalization of survey items. I explore the sensitivity of findings presented in this section to different operationalizations (see Appendix F). Sensitivity tests yield qualitatively similar results.

Summary

In this chapter I presented evidence concerning internal validity. I argued that manipulation of the running variable is implausible and devised a test able to rule out manipulation at the 425-threshold. Balance tests found some covariates are imbalanced. To the extent these tests suggest bias is present, I estimate effects using samples meeting the assumptions of regression discontinuity designs.

I also presented evidence concerning the validity of the instruments. Over 126 tests (90 impetus to improve, six impending departure, 30 reinforcing perceptions), five detected relationships between individual instruments and potentially threatening psychological effects, no more than expected by chance when accepting a Type I error rate of 5%. These findings hold over different operationalizations of survey items (see Appendix F). There is strong evidence supporting the validity of the instruments.

CHAPTER 5

FINDINGS

Introduction

In this chapter I present evidence concerning the effects of classroom observations on teacher performance. To identify these effects, I use multiple years of statewide administrative data from Tennessee with fuzzy regression discontinuity designs (RDD). Fuzzy RDDs are needed because educators do not comply with observation assignment schedules produced by the state. And, for reasons previously discussed, it is plausible non-compliance is endogenous.

In the preceding chapter I presented some evidence supporting the validity of my instruments. After discussing my main findings, I present further evidence corroborating the validity of instruments. I end this chapter with a discussion of heterogeneous effects.

Main Findings

I first discuss findings from models when the outcome is TVAAS scores. I generate RDD estimates in bandwidths of 20, 30, and 40. The Imbens-Kalyanaraman (IK) bandwidth selector (2012) estimated optimal bandwidths of approximately 35 and 20, respectively, at the 200 and 425 thresholds. My bandwidths bracket the optimal bandwidths[24]. Table 11 presents findings from the pooled RDD in the left panel. To the extent the assumptions of a fuzzy RDD are not met at the 200-threshold or for 425-Apprentice teachers, I present estimates using only

[24] I also estimated optimal bandwidths using the cross-validation (CV) method proposed by Ludwig and Miller (2007), but this estimator yielded optimal bandwidths of 75 at the 200 and 425 thresholds. The difference from one LOE to the next is 75, which is why I ignored the CV estimated bandwidth.

Professional teachers surrounding the 425 threshold in the right panel. As seen in the bottom-left and bottom-right panels, the instruments strongly predict the number of observations received. Second stage estimates in the top-left panel show an additional observation leaves TVAAS scores unchanged[25]. The insignificant effect of an additional observation on TVAAS scores hovers near zero. The top-right panel of Table 11 shows qualitatively similar results: there is no evidence an additional observation changes TVAAS scores.

Estimates produced by the pooled local RDD when the outcomes are TLM student achievement scores in Table 12. Second stage estimates in the top panel of Table 12 suggest the marginal observation lowers teacher level mean math achievement by approximately -0.10 units or roughly one-fifth of a standard deviation in the change of these scores (see Table 3 for descriptive statistics). The estimated effects on RLA scores are also negative, but insignificant. Table 13 shows second stage estimates for all teachers surrounding the 425 threshold, and results for Professional teacher surrounding these thresholds for reasons similar to the sample restriction in Table 11. All controls are the same as those used in models that produced the results presented in Table 12. The top panel of Table 13 shows estimates for math/ RLA teachers surrounding the 425 threshold. These point estimates are qualitatively similar to those in Table 12, but none of the estimates in Table 13 are significant. These results hold after restricting the sample to Professional teachers surrounding the 425 threshold (see bottom panel Table 13).

[25] LATEs produced by pooled and unpooled local RDDs assume linearity. For example, the LATE of going from two to the three observations is the same as going from four to five. However, the effect of an additional observation may be non-linear (i.e. depend on number of observations received). I return to this point after discussing results produced by full-sample RDDs.

Longer-Term Effects

Although I have not found any evidence that an additional observation improves contemporaneous teacher performance, it is plausible observation-driven improvement takes time. In this section I present findings related to two longer-term effects of interest: the extended and cumulative effects of observations on teacher performance.

Table 14 presents the extended effects of observations. The top panel displays results using the pooled sample and all four instruments. I also present estimates produced by restricted samples in the bottom panel for aforementioned reasons. The extended effects resemble my main findings: across outcomes and bandwidths most point estimates are negative and all are insignificant. Estimated cumulative effects (see Table 15) are qualitatively similar to previous results, but, for the first time, all estimates are significantly negative.

Remaining Psychological Threats to the Validity of Local RDD Instruments

In the previous chapter I discussed three threats to the validity of instruments used in the local RDD: an impetus to improve, impending departure, and psychological performance boost. In this section I discuss one more test examining the threat of a psychological boost and present findings from a generic falsification test of psychological effects. This new evidence further corroborates the validity of my instruments.

Psychological Performance Boost, Performance Loss

In theory, a teacher should not receive an LOE5 unless they exceed expectations. Conferral of LOE5 may lead a teacher to believe she is exceptional, boosting self-efficacy to the point that she takes instructional risks ultimately leading to higher student achievement. If

assignment to LOE5 boosts performance via psychological gains, behavioral economics implies the loss of LOE5 should cause even greater psychological losses (Tversky & Kahneman, 1991). This implies the effect of the marginal observation for LOE4 teachers assigned to LOE5 in the previous year should be more negative than LOE4 teachers whose previous LOE assignment was less than LOE5.

I estimate LOE5-loss moderated effects by returning to equations 1 and 2. First, I create a dummy variable (LOE5to4) taking a value of one if an LOE4 teacher in year *t* (i.e. once-lagged LOE) was LOE5 in year $t - 1$ (i.e. twice-lagged LOE), otherwise the variable is zero. This dummy variable is interacted with 425-threshold instruments and the endogenous variable. Aside from these interactions, the LOE5-loss sample differs from the original local RDD samples in two ways. I restrict the sample to teachers surrounding the 425 threshold because this psychological threat only pertains to LOE5 teachers. And, the LOE5-loss samples are smaller due to my use of a twice-lagged variable.

I present the LOE5-loss moderated effects on TVAAS and TLM scores in Table 16. None of the interactions in any model are significant, meaning there is no evidence of a productivity-losing LOE5 loss. In conjunction with evidence presented in Chapter 4, these new results effectively rule out threats to the validity of my instruments due to a psychological boost.

Falsification Tests for Generic Psychological Effects Related to LOE Assignment

To some degree, all potentially threatening psychological effects rest on teacher assignment to a relatively lower LOE (LOE1/ LOE2 or LOE4/ LOE5). The impetus to improve, impending departure, and psychological performance boost are supposedly brought about by assignment to a lower (higher in the case of the psychological performance boost) LOE. Across

these psychological effects, the threat to internal validity is that assignment to a lower LOE causes a change in teacher performance independent of the observation process. If such psychological effects exist at the 200 or 425 thresholds, they should exist at other thresholds without discontinuities in the assignment of observations. Thus, assignment to LOE2 instead of LOE3 (LOE3 instead of LOE4) should affect teacher performance if assignment to a lower threshold induces a psychological effect.

I test for the presence of generic psychological effects due to LOE assignment using equation 8:

$$(8)\quad y_{ijt} = \beta_0 + \delta d_{ijt} + \boldsymbol{A}m(\cdot) + \boldsymbol{BX}_{ijt} + \boldsymbol{CS}_{jt} + \gamma_t + e_{ijt}, \qquad \left|LOE_{ijt}\right| \leq w$$

which resembles equations 1 and 2. The key differences between equation 8, and equations 1 and 2, is the use of different instruments and thresholds in the running variable. Equation 8 does not represent a pooled RDD, but two separate RDDs.

The results of this falsification test suggest a generic psychological effect related to LOE assignment does not bias the main findings. Table 17 shows crossing the 275 or 350 threshold is unrelated to TVAAS or TLM math or RLA scores. Furthermore, the point estimates are near zero.

Heterogeneous Effects by Teacher and School Characteristics

I estimate heterogeneous effects by: grade levels (e.g. elementary); teacher years of experience; teacher perceptions about the utility and fairness of the evaluation/ observation system and perceptions about observer expertise; and school administrator effectiveness. To create these estimates I interact moderators with the instruments and endogenous variables (see Appendix H for details). There is no evidence more observations improve student growth for

teachers in any subgroup and weak evidence of heterogeneous effects. Many point estimates vary by subgroups as hypothesized, but moderated effects are often statistically indistinguishable from one another (i.e. 95% confidence intervals overlap across subgroups).

Previous research implies more observations would be most useful for teachers in earlier grade levels because the typical upper grade observer is unlikely to possess expertise in the observed content area of upper grade teachers. Findings in Table 18 partially support this hypothesis. Across all outcomes the largest negative point estimates tend to be associated with high school teachers. Effects on TVAAS for elementary and middle teachers are close in magnitude and significantly less negative than high school estimates (see confidence intervals in Table 18). None of the other point estimates are statistically different from one another.

Prior work also suggests observations would benefit early career teachers the most. There is weak evidence supporting this hypothesis when the outcome is TVAAS: point estimates become more negative as experience increases (see Table 19). However, the moderated effects on TVAAS are statistically indistinguishable over years of experience. When the outcomes are TLM scores point estimates do not support the hypothesized moderated relationship (Table 19).

In broad terms, I hypothesized observations would improve teacher performance more for teachers perceiving the evaluation/ observation as relatively more useful or credible. There is little evidence supporting this hypothesis (see Appendix H).

Heterogeneity by Measures of Administrator Effectiveness

The last set of analyses explore if working in schools with more effective administrators, as measured by the administrator TEAM system, moderates the LATEs. Like teachers, administrators receive a composite effectiveness rating (i.e. LOE) based on observations and

student outcomes (e.g. school level value-added scores). I forego a detailed discussion of the determinants of administrator TEAM scores and LOE. What is important is the administrator evaluation system generates observation scores and LOE allowing me to rank order administrators using these measures, the de facto measures of administrator effectiveness in Tennessee. If these annual measures differentiate administrator skills/ effectiveness, it is plausible teachers working in schools with more skilled/ effective administrators may receive relatively more beneficial observations.

Theoretically, administrator TEAM and LOE scores measure observation-related and non-observation-related effectiveness. If heterogeneity in the effects of observations on teacher performance are only sensitive to observation-related administrator skills, *admLOE* and *admTEAM* may not differentiate among effective observers. At face value, the administrator TEAM rubric measures some observation-specific skills. I identify administrator TEAM rubric indicators (see Table 34 in Appendix H) describing behaviors directly related to teacher: evaluation (*admTE*) and professional learning (*admPL*).

Ideally, I would moderate LATEs by the *admTE*, *admPL*, *admTEAM,* and *admLOE* of administrators who observed teachers, but the data do not support this type of linkage. Administrative data include an "observer identification" variable, but this variable captures who enters observation data into the information management system, which is not necessarily the person who conducted the observation. For example, a principal may conduct an observation but assign an assistant principal to enter observation data into the system. Considering this limitation, I calculate the school-year mean *admTE*, *admPL*, *admTEAM,* and *admLOE*, rank order schools by these annual measures, then assign schools to quartiles. These quartiles serve as the moderators. Thus, the heterogeneous effects with respect to these moderators estimate if the

LATEs of an additional observation on teacher performance depend on the skills/ effectiveness of the hypothetically typical (i.e. mean) administrator within the teacher's school.

Results in Table 20 resemble other moderated findings: all point estimates are negative and tend to move in the hypothesized direction, but none of the point estimates are statistically distinguishable from one another. Across all outcomes, point estimates become less negative as *admLOE* rises (see top panel). This is also true when the outcome is TVAAS scores and moderator is *admTEAM* (see bottom panel). But, the moderation of treatment on TLM scores by *admTEAM* does not follow such clear patterns. Point estimates associated with observers in the fourth quartile of *admTEAM* tend to be the least negative, but some of the most negative LATEs are associated with observers in the third quartile of *admTEAM*.

Evidence is inconsistent regarding the hypothesis that administrator observation-specific effectiveness, as measured by the administrator TEAM rubric, moderates the effect of observations on student growth. Similar to results in Table 20, all estimates are negative and statistically indistinguishable from one another. The top panel of Table 21 shows that the most negative effects on TVAAS and TLM RLA scores are associated with teachers in schools with the least skilled observers regarding teacher evaluation. It is also true that the least negative effects on TLM math scores are associated with teachers in schools where observers are the most skilled teacher evaluators. These patterns support the hypothesized relationships. However, other estimates moderated by *admTE* do not follow hypothesized patterns.

The bottom panel of Table 21 shows similarly inconsistent results. As hypothesized, the least negative effects on TLM math scores are associated with teachers in schools with the most highly rated observers regarding *admPL*. And, the most and least negative effects on TLM RLA are associated with teachers in schools with the least and most highly rated observers concerning

their skills in supporting teacher professional learning, respectively. However, other point estimates move in unexpected directions. Like previous results, all point estimates are negative and there is no strong evidence of heterogeneity.

It is possible the scores underlying these four moderators do not measure the type of administrator effectiveness that truly matters when it comes to conducting teacher observations for formative evaluation. This is especially plausible for *admLOE* and *admTEAM*, both of which may measure broad, ambiguous administrator behaviors. Yet, the most consistent evidence supporting the hypothesis administrator skills moderates the effects of observations on student growth came from the *admLOE* moderator. Moreover, some of the weakest evidence supporting this hypothesis is produced by the two moderators that arguably measure observation-specific effectiveness: *admTE* and *admPL*. A thorough investigation into these patterns is beyond the scope of this dissertation but could help practitioners and researchers better understand the measurement properties of administrator LOE and TEAM scores. Nevertheless, the results of these moderation analyses are clear: despite any strong evidence of heterogeneity, it is clear more observations do not improve student growth for teachers in any of the examined subgroups.

Sensitivity Tests: Full-Sample RDDs

Thus far, there is no evidence observations improve teacher performance. However, it is possible the results do not generalize beyond the local RDD bandwidths and/ or lack enough power to detect effects. I address these possibilities by estimating RDDs using records from across the LOE-cont spectrum ("full-sample RDD"). I use the following models:

$$(9)\ obys_{ijt} = \ddot{\beta}_0 + \delta\widehat{obs}_{ijt} + \ddot{\boldsymbol{A}}\check{f}(\cdot) + \ddot{\boldsymbol{B}}\boldsymbol{X}_{ijt} + \ddot{\boldsymbol{C}}\boldsymbol{S}_{jt} + \gamma_t + u_{ijt}$$

$$(10)\ obs_{ijt} = \ddot{\beta}_0 + \boldsymbol{\theta}pol\_obs_{ijt} + \ddot{\boldsymbol{A}}\check{f}(\cdot) + \ddot{\boldsymbol{B}}\boldsymbol{X}_{ijt} + \ddot{\boldsymbol{C}}\boldsymbol{S}_{jt} + \gamma_t + u_{ijt}$$

Where y_{ijt} is the TVAAS score, $\check{f}$ a higher order polynomial of unpooled lagged LOE-cont interacted with teacher certification status[26], and all other controls are the same as those used in equations 1 and 2. obs_{ijt} is the endogenous number of observations received. The instrument in equation 10 differs from the instruments in previous equations. pol_obs_{ijt} is the minimum number of observations assigned to the *i*th teacher in school *j* in year *t* based on the TDOE observation assignment schedule[27].

Estimates produced by the full-sample RDD are qualitatively similar to results produced by local RDDs, but all estimates are significant (see left panel of Table 22). Thus, local RDD estimates generalize beyond the local bandwidths. The right panel of Table 22 includes full-sample extended effects estimates (equations 3 and 4 represent models estimating local RDD extended effects). Local RDD extended effects were insignificant (see Table 14). Full-sample extended effects are either negative and significant, or insignificant.

Despite the overwhelming evidence that more observations do not improve contemporaneous or future teacher performance, all estimates assume the effects of observations on teacher performance are linear. However, it is plausible the effect of an additional observation depends on the number of observations received. The receipt of one or two observations per year may provide teachers with a manageable amount of feedback they can use to improve. But, the receipt of too many may lead to an overwhelming amount of feedback, confusing improvement efforts. Alternatively, the receipt of too few observations may not provide enough guidance to be

[26] Higher degree terms of LOE-cont were added to a polynomial of LOE-cont until the last term was no longer significant. This happened after adding the third-degree term, so the polynomial in the full-sample RDD is of the second order.

[27] All teachers of untested subjects with a lagged LOE5 or LOE1 have pol_obs_{ijt} values of one or four, respectively. All Apprentice (Professional) teachers of untested subjects with a lagged LOE below five (above one but below five) should receive a minimum of four (two) observations. The minimum number of observations assigned to teachers of tested subjects is somewhat ambiguous. I assign teachers of tested subjects the minimum number of observations possible based on their lagged LOE and TVAAS. I first assign a teacher of tested subjects the maximum of her lagged discrete TVAAS and LOE, then assign policy-assigned observations using this maximum and their certification status. This results in fewer assigned observations when possible.

useful. I tested for the presence of non-linear effects in both local and full-sample RDDs in Appendix G. There is no evidence of non-linearities.

Summary

Evidence presented in this chapter clearly supports the conclusion that more frequent observations do not improve contemporaneous or future teacher performance. Neither was it true that the accumulation of observations over time improved teacher performance. Furthermore, moderation analyses failed to find any evidence of positive effects across different groups of teachers. In the next and final chapter, I discuss what might account for these findings and the implications of my work.

CHAPTER 6

CONCLUSIONS AND IMPLICATIONS

Introduction

In this final chapter, I recapitulate my findings and describe the limitations of my work. In brief, I find no evidence that more observations improve teacher performance, but the generalizability of this inference may be limited. After interpreting my findings, I speculate why my negative and null findings might exist. In broad terms, I hypothesize that the ineffectiveness of observations as a formative evaluation tool may rest on the design of the Tennessee observation system. I end with implications for policy.

Review of Results and Limitations

Throughout the previous chapter I discussed multiple findings about the relationship between the number of observations per year and teacher performance. Using RDDs I estimated contemporaneous, longer-term, cumulative, and heterogeneous effects. While observer bias prevents me from drawing conclusions about the effects of observations on summative observation scores (see Appendix E), I found no evidence that the marginal observation improves measures of teacher performance based on student growth scores.

The collection of effects on TVAAS scores imply more observations lower student growth or leave it unchanged. Contemporaneous effects were insignificant (Table 11) or negative (Table 22), longer-term effects were also insignificant (Table 14) or negative (Table 22), and the cumulative effect of all observations received through year *t* were negative (Table 15).

Moderated effects (Tables 18 – 21) resemble previous findings: all were insignificant or negative. Only one moderation analysis found moderately strong evidence of heterogeneous effects: the effect on the TVAAS scores of high school teachers is significantly more negative than the effects on TVAAS for teachers in other grade levels.

Observations change teacher-level mean (TLM) math and reading/ language arts (RLA) scores in much the same way. Contemporaneous (Tables 12, 13 and 22), longer-term (Tables 14 and 22), and cumulative effects (Table 15) were either insignificant or negative. While moderation analyses also produced negative or null findings, there was only weak evidence of heterogeneous effects on TLM scores.

Despite some threats to internal validity, none of these threats challenge the implication that more observations do not improve teacher performance. In TVAAS models there was weak evidence some Apprentice teacher covariates are imbalanced (Tables 4 and 5). In TLM models there was strong evidence of imbalance at the 200-threshold and some evidence of Apprentice teacher covariate imbalance at the 425-threshold (Table 6). To the extent these imbalances introduce bias, I estimated effects using samples meeting the assumptions of an RDD. Estimates in restricted and unrestricted samples are qualitatively similar, no matter the restriction. Finally, there is strong evidence supporting the validity of the instruments.

Limitations

There are three potential limitations concerning the generalizability of my estimates. First, local RDD estimates were based on variation predominantly coming from Professional teachers surrounding the 425 threshold, and in some instances these teachers comprised the entirety of the analytical sample. If effects produced by these samples are sample-specific, this

would threaten the generalizability of my findings. I explored this by comparing estimates produced by 425-Professional teachers to all teachers surrounding the 425 threshold, all teachers surrounding the 200 and 425 threshold, and all teachers (i.e. full-sample RDDs). In all circumstances, estimates produced by different models and different samples were qualitatively similar. Moreover, even if inferences are restricted to Professional teachers on either side of the 425 threshold, these conclusions would apply to over 60% of Tennessee teachers.

The second limitation concerns potential heterogeneity in effects due to the source of identifying variation. Observations occur for two broad reasons: policy-assignment and/ or educator discretion. All estimated relationships are LATEs, representing the effect of the marginal policy-assigned observation on teacher performance. These LATEs may not capture the effect of the marginal observation on teacher performance if there is heterogeneity in identifying variation.

Finally, measures of teacher performance are restricted to test-based outcomes (Appendix C examines observational ratings, but there is strong evidence of bias in these models). Teachers are responsible for more than improving student achievement. Moreover, observations may affect other outcomes since rubrics often describe more than just academic behaviors. Future work may address other outcomes of interest including student: disciplinary infractions, attendance, and course-taking.

Potential Explanatory Mechanisms

In this section I speculate why more observations do not improve teacher performance. Broadly, observers may be ill-equipped to facilitate teacher professional learning via observations, and some policies may undermine credibility of the TEAM observation system.

Observers may not possess the expertise to recognize high-leverage, content-specific teaching behaviors needing improvement. In Chapter 2 I reviewed research suggesting observer expertise is an antecedent to effective observations. Considering that the typical observer is a school administrator, it is highly unlikely teachers are observed by a content expert. This is especially the case in upper grade levels where teachers are not generalists. Indeed, the strongest evidence of heterogeneous effects existed when treatment was moderated by grade levels. It is plausible observers without content expertise provide misguided post-observation feedback. But, to appease their observers/ evaluators, teachers may incorporate this misguided feedback into practice, worsening student achievement or leaving it unchanged.

Second, Tennessee classroom observers may be ill-equipped to effectively facilitate post-observation feedback conferences. During the study period annual observer training devoted at most six hours to the provision of feedback and facilitation of post-observation conferences. Six hours per year may not be enough time to fully develop these skills. Moreover, some observers may not participate in this training every year. Observers who pass an annual online certification exam can test out of annual training. After speaking with some TDOE employees I learned most observers pass their annual re-certification tests, so most receive little cumulative training. Yet, post-observation feedback is the ostensible linchpin of formative evaluation via observations. Considering the lack of training, it is plausible feedback provided by these observers is ineffective and potentially damaging to teacher performance. Under this explanation, teachers receiving more observations would receive more ineffective feedback (e.g. incoherent, inaccurate). Prior work implies the accumulation of ineffective feedback leads employee (i.e. teacher) performance astray (Jawahar, 2010; Kluger & DeNisi, 1996).

In addition to these observer weaknesses, some observation policies may weaken credibility of the TEAM observation system. A great deal of prior work suggests employees (i.e. teachers) do not act on feedback they believe lacks credibility (Cherasaro et al., 2016; Ilgen et al., 1979; Kimball, 2003; Kinicki, Prussia, Wu, & McKee-Ryan, 2004). There are at least three reasons why teachers may not act on feedback generated by the TEAM observation system. First, the typical classroom observation is expected to last 15 minutes (see Chapter One). Teachers may not trust feedback based on such short observations, leading them to ignore it. Second, prior work finds observers engage in satisficing behaviors (e.g. shortened post-observation conferences) to manage the demands of modern teacher evaluations systems (see Chapter 2). Multiple interactions with Tennessee school administrators echo this sentiment: observers say they are overly burdened by the TEAM observation system. In conjunction with the brevity of classroom observations, short post-observation conferences may further teacher perceptions that post-observation feedback is not credible. Third, requiring observers to identify a teacher weakness, or area of refinement, after every observation (see Chapter One) may erode the importance of post-observation feedback. This policy may raise doubt in a teacher's mind as to whether her area of refinement represents a genuine weakness or an ignorable, compliance-driven requirement. Moreover, some observers frequently apologize for identifying an area of refinement, acknowledging the only reason they identify the weakness is to comply with policy. Thus, the design and implementation of the area of refinement policy may lead teachers to perceive post-observation feedback as perfunctory and ignorable.

Implications

The potential explanations for my findings imply some policy changes could improve the effectiveness of observations. First, teacher performance may be more likely to improve if observers possess content-expertise. But, it is unreasonable to expect school administrators to become experts in all observed content areas. Instead, districts could form collectives, pool resources, and hire cross-district content-specific observers. These observers would conduct observations instead of school administrators. It seems reasonable that the typical school administrator would welcome this change since prior work suggests school administrators are overly burdened by modern teacher observation systems (see Chapter 2). However, I have interacted with some Tennessee school and district leaders who value the teacher performance information gained from classroom observations. It seems these leaders place special value on this information for personnel decision-making (e.g. teacher retention). Therefore, I am not proposing school administrators never visit classrooms. School administrators could continue visiting classrooms for summative teacher evaluation. At the same time, observers with content expertise, shared by multiple districts, can conduct classroom observations for formative teacher evaluation.

The second policy implication concerns the demands on school administrator time, and potential brevity of observation cycles. Currently, teachers assigned to the highest (lowest) category of effectiveness are assigned one (four) observation(s) per year. Teachers assigned to the middle categories of effectiveness (i.e. LOE2 – LOE4) are assigned two or four observations. Policies could change so school administrators spend more or the same time with less effective teachers, and less time with more effective teachers, potentially reducing the total number of observations conducted. For example, teachers assigned the highest category of effectiveness

could be observed once over two years, and all teachers in the middle categories of effectiveness could be assigned two observations. These new schedules would address two problems described in the previous section. A reduction in the total number of observations school administrators conduct each year means a single observation can last more than 15 minutes, and administrators can devote more time to each post-observation conferences.

Policymakers may be able to improve the importance of post-observation feedback with a minor change to observation policy: remove the requirement all observations must identify an area of weakness. This requirement may be in place so reticent school administrators can blame state policy when identifying a teacher's area of weakness. The Tennessee Board of Education can still provide cover for these administrators by requiring observers to identify an area of improvement for the least effective teachers. Or, observers could be expected to identify a single area of weakness after multiple observations.

Adopting any of the proposed policy solutions in this section may improve teacher performance. However, the second proposition may be the least controversial: shift observer foci to those teachers needing the most assistance. To the extent more observations worsen teacher performance or leave it unchanged, reducing the number of observations received by more effective teachers should do no harm. It is also plausible that the performance of less effective teachers could improve if observers can devote more time to each observation cycle. While this dissertation does not offer a clear solution, it does identify a clear problem: teacher classroom observations are not working as intended.

APPENDIX

A. Tables and Figures

Figure 1: Predictive Margins of Quitting with 95% CIs

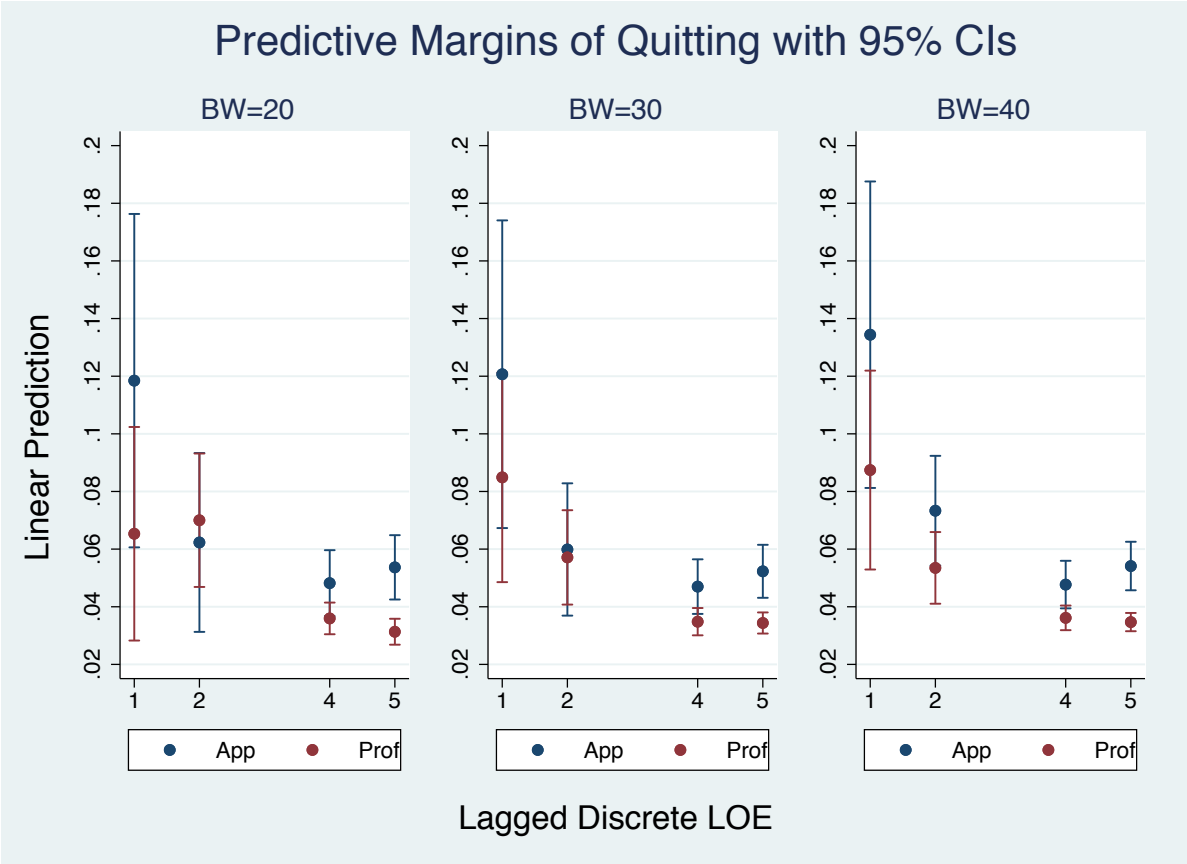


Table 1: Distribution of Observations by Certification and Prior LOE

	Prior LOE 1		Prior LOE 2 – 4		Prior LOE 5	
	Freq	%	Freq	%	Freq	%
Apprentice Teachers Classroom Visits						
1	3	1.55	409	3.16	**2167**	**63.59**
2	9	4.66	1840	14.20	625	18.34
3	11	5.70	1089	8.40	263	7.72
4	**140**	**72.54**	**8241**	**63.58**	276	8.10
5 or more	30	15.54	1383	10.67	77	2.26
Professional Teachers Classroom Visits						
1	4	1.22	2906	4.74	**25264**	**68.90**
2	25	7.62	**43643**	**71.15**	8382	22.86
3	18	5.49	4067	6.63	2265	6.18
4	**229**	**69.82**	9516	15.51	697	1.90
5 or more	52	15.85	1209	1.97	57	0.16

Note: Percentages represent the percentage of all Tennessee teachers with a certain certification status and LOE (e.g. 72.54% of Apprentice teachers with a prior LOE1 received four classroom visits) receiving a given number of observations per year. The percentages within each bold box total to 100. Bold cells indicate the minimum number of policy-assigned classroom visits a teacher should receive.

Table 2: Sample Descriptive Statistics. DV= TVAAS

TVAAS Score	1.51	(7.46)
TEAM Score	4.03	(0.51)
Female	0.83	.
BA+	0.59	.
Years Experience	12.18	(9.01)
Black_White	0.05	.
% Sample from Teachers Surrounding LOE 200	7.27%	
% Sample from Apprentice Teachers	13.8%	

Note: Sample descriptive statistics use data from the analytical sample associated with the largest bandwidth of 40. Standard deviations in parentheses. BA+ is an indicator signaling if teacher has earned more than a BA/ BS degree. Black_White is an indicator signaling if teacher is black instead of white. Nearly all TN teachers are black or white.

Table 3: Sample Descriptive Statistics. DV=TLM Math and RLA Teachers

	Math Mean	Math SD	RLA Mean	RLA SD
TLM Math Scores	0.02	(0.70)	.	.
Annual Change in TLM Math Scores	0.10	(0.51)	.	.
TLM RLA Scores	.	.	0.03	(0.67)
Annual Change in TLM RLA Scores	.	.	0.10	(0.46)
Female	0.83	(0.38)	0.90	(0.3)
BA+	0.57	(0.49)	0.61	(0.49)
Years Experience	11.47	(8.69)	12.10	(9.07)
Black_White	0.05	(0.22)	0.06	(0.23)
% Sample from Teachers Surrounding LOE 200	6.2	.	6.9	.
% Sample from Apprentice Teachers	15.8	.	14.3	.
Proportion of Students Taught with Characteristics				
Female	0.49	(0.12)	0.48	(0.13)
Black	0.14	(0.19)	0.14	(0.19)
White	0.77	(0.24)	0.77	(0.24)
Asian	0.02	(0.04)	0.02	(0.04)
Hispanic	0.07	(0.1)	0.07	(0.1)
FRPL	0.56	(0.25)	0.54	(0.25)
ESL	0.08	(0.12)	0.07	(0.13)
Immigrant	0.02	(0.05)	0.01	(0.06)

Note: Sample descriptive statistics use data from the analytical sample associated with the largest bandwidth of 40. Standard deviations in parentheses. An annual change in teacher level mean scores uses achievement data linked to students taught by a teacher in year *t*: it is the difference in teacher level mean scores received by these students in year *t* and the scores these students received in year $t - 1$. BA+ is an indicator signaling if teacher has earned more than a BA/ BS degree. Black_White is an indicator signaling if teacher is black instead of white. Nearly all TN teachers are black or white. Proportions convey the proportion of students a teacher taught with a given characteristic.

Table 4: Covariate Balance Tests at LOE 200 Threshold. DV= TVAAS

Covariate	$w = 20$	$w = 30$	$w = 40$
Experience: App	1.83	1.21	0.78
	[1.292]	[1.157]	[1.161]
Experience: Prof	0.33	0.43	0.23
	[1.929]	[1.623]	[1.459]
Female: App	-0.10	-0.13	-0.09
	[0.133]	[0.110]	[0.097]
Female: Prof	-0.06	-0.11	-0.15
	[0.109]	[0.091]	[0.083]
BA+: App	-0.06	-0.19	-0.18*
	[0.119]	[0.099]	[0.089]
BA+: Prof	0.13	0.04	0.06
	[0.103]	[0.087]	[0.080]
Black: App	0.07	0.07	0.05
	[0.084]	[0.072]	[0.068]
Black: Prof	-0.05	0.02	0.01
	[0.082]	[0.068]	[0.064]
Lag TVAAS: App	-3.43	-1.52	-1.53
	[2.849]	[2.519]	[2.068]
Lag TVAAS: Prof	0.88	-0.47	-0.20
	[1.954]	[1.667]	[1.556]
Lag TEAM: App	0.02	0.04	0.04
	[0.060]	[0.050]	[0.044]
Lag TEAM: Prof	0.02	> -0.01	-0.04
	[0.073]	[0.059]	[0.055]
N (Tch-Yrs)	602	972	1507

Note: Estimates represent the total predicted change in the outcome. Standard errors in brackets, clustered at the teacher level. OLS estimator employed to estimate all coefficients. BA+ is a binary variable indicating if a teacher reported having a degree higher than a BA/ BS. Black is an indicator signaling whether the teacher reported her

ethnicity/ race as Black or White. Lagged TVAAS not included in any model, but balance is still checked. Tests for balance of lagged TVAAS use samples of 533, 853, and 1307. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 5: Covariate Balance Tests at LOE 425 Threshold. DV= TVAAS

Covariate	$w = 20$	$w = 30$	$w = 40$
Experience: App	0.22	0.15	-0.17
	[0.304]	[0.265]	[0.214]
Experience: Prof	-0.37	0.35	0.75
	[0.574]	[0.477]	[0.414]
Female: App	0.05	0.06	0.02
	[0.073]	[0.058]	[0.050]
Female: Prof	< 0.01	-0.01	0.01
	[0.025]	[0.020]	[0.017]
BA+: App	0.04	0.03	0.06
	[0.085]	[0.068]	[0.058]
BA+: Prof	0.03	0.02	-0.01
	[0.033]	[0.027]	[0.023]
Black: App	-0.09*	-0.08**	-0.04
	[0.036]	[0.028]	[0.025]
Black: Prof	-0.01	-0.01	-0.01
	[0.015]	[0.012]	[0.010]
Lag TVAAS: App	1.15	-0.47	-0.64
	[1.033]	[0.837]	[0.725]
Lag TVAAS: Prof	0.18	0.20	0.31
	[0.381]	[0.305]	[0.264]
Lag TEAM: App	0.06	0.10*	0.07
	[0.059]	[0.048]	[0.041]
Lag TEAM: Prof	-0.03	-0.01	0.02
	[0.027]	[0.023]	[0.019]
N (Tch-Yrs)	9412	14278	19224

Note: Estimates represent the total predicted change in the outcome. Standard errors in brackets, clustered at the teacher level. OLS estimator employed to estimate all coefficients. BA+ is a binary variable indicating if a teacher reported having a degree higher than a BA/ BS. Black is an indicator signaling whether the teacher reported her

ethnicity/ race as Black or White. Lagged TVAAS not included in any model, but balance is still checked. Tests for balance of lagged TVAAS use samples of 6520, 10264, and 14135. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 6: Covariate Balance Tests at LOE 425 Threshold. DV=TLM Math and RLA Teachers

	TLM Math			TLM RLA		
Covariate	$w = 20$	$w = 30$	$w = 40$	$w = 20$	$w = 30$	$w = 40$
Experience: App	-0.10	0.11	-0.02	0.41	0.62	-0.02
	[0.501]	[0.417]	[0.341]	[0.501]	[0.467]	[0.343]
Experience: Prof	-1.76*	0.01	0.75	-2.53**	-0.46	0.06
	[0.873]	[0.715]	[0.617]	[0.961]	[0.775]	[0.657]
Censored Exp: App	-0.02	-0.04	-0.13	0.29	0.20	0.02
	[0.277]	[0.234]	[0.207]	[0.316]	[0.268]	[0.232]
Censored Exp: Prof	-0.33	0.07	0.17	-0.21	0.13	0.25
	[0.258]	[0.211]	[0.181]	[0.252]	[0.203]	[0.175]
Female: App	0.20*	0.19*	0.09	0.02	0.01	0.02
	[0.095]	[0.080]	[0.069]	[0.085]	[0.075]	[0.064]
Female: Prof	-0.06	-0.04	-0.02	-0.03	-0.03	-0.01
	[0.042]	[0.034]	[0.029]	[0.031]	[0.025]	[0.022]
BA+: App	-0.14	-0.09	-0.02	-0.12	-0.05	0.03
	[0.117]	[0.096]	[0.083]	[0.129]	[0.105]	[0.088]
BA+: Prof	0.04	0.02	0.02	0.06	0.02	0.01
	[0.051]	[0.042]	[0.036]	[0.049]	[0.040]	[0.035]
Black: App	-0.10	-0.09	-0.01	-0.13*	-0.05	-0.03
	[0.060]	[0.050]	[0.044]	[0.062]	[0.056]	[0.050]
Black: Prof	-0.04	-0.02	-0.01	-0.02	0.01	< 0.01
	[0.025]	[0.021]	[0.018]	[0.025]	[0.020]	[0.018]
Prior TLM Student Score: App	> -0.01	0.10	0.03	0.04	0.13	0.11
	[0.230]	[0.160]	[0.134]	[0.266]	[0.194]	[0.159]
Prior TLM Student Score: Prof	< 0.01	0.01	> -0.01	-0.03	< 0.01	0.02
	[0.083]	[0.067]	[0.056]	[0.073]	[0.058]	[0.049]

Proportion of Students Taught with Characteristics						
Female: App	0.04	0.01	0.02	-0.01	0.02	0.01
	[0.040]	[0.031]	[0.026]	[0.045]	[0.034]	[0.029]
Female: Prof	-0.02	> -0.01	> -0.01	-0.01	> -0.01	-0.01
	[0.013]	[0.011]	[0.009]	[0.015]	[0.012]	[0.010]
Black: App	> -0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
	[0.005]	[0.004]	[0.003]	[0.005]	[0.004]	[0.003]
Black: Prof	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	> -0.01
	[0.001]	[0.001]	[0.001]	[0.002]	[0.001]	[0.001]
White: App	> -0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
	[0.005]	[0.004]	[0.003]	[0.005]	[0.004]	[0.003]
White: Prof	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	> -0.01
	[0.001]	[0.001]	[0.001]	[0.002]	[0.001]	[0.001]
Asian: App	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
	[0.004]	[0.003]	[0.003]	[0.005]	[0.004]	[0.003]
Asian: Prof	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	> -0.01
	[0.001]	[0.001]	[0.001]	[0.001]	[0.001]	[0.001]
Hispanic: App	> -0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
	[0.005]	[0.004]	[0.003]	[0.005]	[0.004]	[0.003]
Hispanic: Prof	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
	[0.001]	[0.001]	[0.001]	[0.001]	[0.002]	[0.001]
FRPL: App	0.12*	0.08*	0.05	0.11	0.08	0.05
	[0.051]	[0.041]	[0.035]	[0.061]	[0.047]	[0.038]
FRPL: Prof	0.02	0.01	> -0.01	0.02	0.01	> -0.01
	[0.021]	[0.017]	[0.015]	[0.020]	[0.016]	[0.013]
ESL: App	-0.03	-0.02	-0.02	-0.01	0.01	< 0.01
	[0.028]	[0.021]	[0.016]	[0.016]	[0.013]	[0.011]
ESL: Prof	-0.01	-0.01*	-0.01	-0.01	-0.01*	-0.01*
	[0.006]	[0.004]	[0.004]	[0.006]	[0.005]	[0.004]

Immigrant: App	0.02	0.01	0.01	0.01	0.01	< 0.01
	[0.019]	[0.015]	[0.012]	[0.021]	[0.018]	[0.015]
Immigrant: Prof	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
	[0.004]	[0.004]	[0.003]	[0.004]	[0.003]	[0.003]
N(Tch-Yrs)	3348	5205	7197	3143	4906	6783

Note: Estimates represent the total predicted change in the outcome. Standard errors in brackets. Standard errors clustered at the teacher level. OLS estimator employed to estimate all coefficients. BA+ is a binary variable indicating if a teacher reported having a degree higher than a BA/ BS. Black is an indicator signaling whether the teacher reported her ethnicity/ race as Black or White. Censored experience is years of experience censored after ten years. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 7: Sample Descriptive Statistics. DV = Survey Items

	Mean	SD
Sum: Svy Hrs in PD (*PDhrs*)	2.56	[17.46]
Sum: Svy Tch Collab (*tchcollab*)	11.91	[41.10]
Sum: Svy Exerted More Effort (*effortsum*)	0.31	[1.54]
Svy Hrs Improved Instruction (*insthrs*)	37.27	[33.74]
Sum: Svy Hrs Prepped for Obs (*obshrs*)	2.25	[1.30]
Eval Will Improve Teaching (*imprvtch*)	0.59	[0.492]
Post-Obs FB is Useful (*fbuseful*)	0.87	[0.336]
Evals Change My Teaching (*chngtch*)	0.72	[0.449]
Observer Qualified (*obsqual*)	0.78	[0.413]
Evaluations Are Fair (*faireval*)	0.64	[0.480]

Note: Sample descriptive statistics use data from samples associated with the largest bandwidth of 40. Standard deviations in brackets, number of teacher-year records in bandwidths of 40 in parentheses. See Appendix E to read about the original survey items and scales these variables are based on and how I created these variables.

Table 8: Tests of Joint Significance Concerning the Impetus to Improve

	$w = 20$	$w = 30$	$w = 40$
Sum: Svy Hrs in PD (*PDhrs*)	1.59	1.55	1.46
	[0.175]	[0.185]	[0.211]
	(2350)	(3601)	(4664)
Sum: Svy Tch Collab (*tchcollab*)	5.79***	1.12	1.82
	[< 0.001]	[0.343]	[0.122]
	(1048)	(1607)	(2109)
Sum: Svy Exerted More Effort (*effortsum*)	3.57**	1.87	0.59
	[0.007]	[0.112]	[0.671]
	(1427)	(2183)	(2802)
Sum: Svy Hrs Improved Instruction (*insthrs*)	0.20	0.45	0.44
	[0.940]	[0.776]	[0.782]
	(3701)	(5740)	(7555)
Sum: Svy Hrs Prepped for Obs (*obshrs*)	0.82	3.22*	2.37
	[0.511]	[0.012]	[0.051]
	(8480)	(12958)	(16604)

Note: p-values in brackets, number of teacher-year records in sample in parentheses. All models include teacher demographics, certification status, controls for the distribution of teacher effectiveness at the school level, second order polynomial of LOE interacted with teacher certification status, and year fixed effects. * ($p < 0.05$), ** ($p < 0.01$), *** ($p < 0.001$)

Table 9: Impetus to Improve: Testing Joint Significance of 425-Threshold Instruments

	$w = 20$	$w = 30$	$w = 40$
Sum: Svy Hrs in PD (*PDhrs*)	0.13 [0.879] (1698)	1.08 [0.339] (2526)	0.72 [0.487] (3318)
Sum: Svy Tch Collab (*tchcollab*)	1.57 [0.208] (709)	0.85 [0.429] (1087)	1.12 [0.326] (1439)
Sum: Svy Exerted More Effort (*effortsum*)	0.92 [0.399] (1084)	0.12 [0.886] (1589)	< 0.01 [0.999] (2046)
Sum: Svy Hrs Improved Instruction (*insthrs*)	0.13 [0.881] (2721)	0.18 [0.831] (4181)	0.64 [0.526] (5591)
Sum: Svy Hrs Prepped for Obs (*obshrs*)	0.89 [0.410] (6417)	0.76 [0.466] (9417)	0.21 [0.814] (12174)

Note: Ibid.

Table 10: Tests of Joint Significance Concerning Reinforcing Perceptions

	$w = 20$	$w = 30$	$w = 40$
Evals Improve Teaching (*imprvtch*)	1.20	0.59	0.45
	[0.302] (12227)	[0.552] (18676)	[0.640] (24069)
Post-Observation Feedback is Useful (*fbuseful*)	0.23	1.31	0.17
	[0.792] (7796)	[0.270] (12000)	[0.845] (15745)
Changed Teaching Due to Evals (*chngtch*)	0.11	0.64	0.43
	[0.897] (3239)	[0.529] (4986)	[0.650] (6530)
Observers are Qualified (*obsqual*)	0.33	0.56	0.19
	[0.722] (7730)	[0.570] (11768)	[0.826] (14957)
Evaluations are Fair (*faireval*)	1.05	0.58	0.89
	[0.350] (7735)	[0.562] (11772)	[0.410] (14970)

Note: p-values in brackets, number of teacher-year records in sample in parentheses. All models include teacher demographics, certification status, controls for the distribution of teacher effectiveness at the school level, second order polynomial of LOE interacted with teacher certification status, and year fixed effects.

Table 11: Pooled and 425-Only Local RDD. DV=TVAAS

	Pooled RDD			425-Prof RDD		
	$w = 20$	$w = 30$	$w = 40$	$w = 20$	$w = 30$	$w = 40$
2nd Stage: Number of Observations	-0.77	-0.05	0.10	0.02	0.42	0.67
	[0.623]	[0.543]	[0.444]	[0.941]	[0.724]	[0.587]
1st Stage:						
App Below LOE 200	1.31***	1.32***	1.59***	.	.	.
	[0.318]	[0.281]	[0.299]	.	.	.
Prof Below LOE 200	1.76***	1.85***	1.80***	.	.	.
	[0.226]	[0.171]	[0.152]	.	.	.
App Below LOE 425	0.86***	0.94***	0.89***	.	.	.
	[0.258]	[0.203]	[0.171]	.	.	.
Prof Below LOE 425	0.57***	0.57***	0.62***	0.48***	0.51***	0.54***
	[0.070]	[0.057]	[0.049]	[0.054]	[0.044]	[0.038]
Prof	-0.28	-0.24	-0.24*	.	.	.
	[0.166]	[0.131]	[0.109]	.	.	.
Intercept	3.12***	3.03***	2.95***	2.75***	2.77***	2.73***
	[0.287]	[0.229]	[0.197]	[0.212]	[0.174]	[0.149]
N(Teachers-Year)	7053	11117	15622	8150	12489	16936

Note: Ibid. Each 1st stage estimate represents the total effect of crossing a threshold for each teacher group, none of the 1st stage estimate are interactions. * ($p < 0.05$), ** ($p < 0.01$), *** ($p < 0.001$)

Table 12: Pooled Local RDD Main Results. DV=TLM Math and RLA Teachers

	TLM Math			TLM RLA		
	$w = 20$	$w = 30$	$w = 40$	$w = 20$	$w = 30$	$w = 40$
2nd Stage: Number of Observations	-0.11*	-0.10*	-0.08*	-0.04	-0.03	-0.05
	[0.049]	[0.042]	[0.035]	[0.043]	[0.032]	[0.029]
1st Stage:						
App Below LOE 200	1.26	1.15	1.70***	1.34***	1.01***	1.08***
	[0.679]	[0.0595]	[0.486]	[0.285]	[0.227]	[0.197]
Prof Below LOE 200	2.28***	2.27***	2.17***	2.06***	2.57***	2.63***
	[0.257]	[0.197]	[0.201]	[0.473]	[0.454]	[0.457]
App Below LOE 425	1.31***	1.30***	1.16***	0.97*	1.28***	1.17***
	[0.360]	[0.279]	[0.239]	[0.394]	[0.302]	[0.259]
Prof Below LOE 425	0.50***	0.50***	0.57***	0.61***	0.64***	0.64***
	[0.101]	[0.081]	[0.071]	[0.101]	[0.080]	[0.068]
Prof	-0.28	-0.15	-0.17	-0.24	-0.24	-0.15
	[0.219]	[0.175]	[0.149]	[0.296]	[0.233]	[0.202]
Intercept	4.31***	4.33***	3.79***	2.67**	2.73***	2.79***
	[0.920]	[0.081]	[0.714]	[0.973]	[0.564]	[0.546]
N(Teachers-Year)	3348	5205	7197	3143	4906	6783

Note: Teacher-clustered standard errors in brackets. All models include a polynomial of the teacher-level mean of prior achievement scores for students taught in year t (e.g. the 2011-12 scores of students taught in 2012-13), proportion of students taught holding various characteristics, teacher demographics including certification status, controls for the distribution of teacher effectiveness at the school level, a second order polynomial of pooled LOE interacted with teacher certification status, and year fixed effects. Each 1st stage estimate represents the total effect of crossing a threshold for each teacher group, none of the 1st stage estimate are interactions. * ($p < 0.05$), ** ($p < 0.01$), *** ($p < 0.001$)

Table 13: Effects by Thresholds. DV=TLM Math or RLA

	TLM Math			TLM RLA		
	$w = 20$	$w = 30$	$w = 40$	$w = 20$	$w = 30$	$w = 40$
	Teachers Surrounding LOE 425 Only					
2nd Stage: Number of Observations	-0.13	-0.09	-0.08	-0.07	-0.04	-0.06
	[0.088]	[0.079]	[0.067]	[0.083]	[0.055]	[0.048]
N (Tch-Yrs)	3175	4909	6746	2975	4625	6315
	Professional Teachers Surrounding LOE 425 Only					
2nd Stage: Number of Observations	-0.29	-0.17	-0.08	-0.07	-0.03	-0.02
	[0.160]	[0.118]	[0.085]	[0.082]	[0.060]	[0.052]
N (Tch-Yrs)	2662	4187	5814	2558	4025	5539

Note: Ibid.

Table 14: Extended Effects of Observations

	TVAAS			TLM Math Scores			TLM RLA Scores		
	$w = 20$	$w = 30$	$w = 40$	$w = 20$	$w = 30$	$w = 40$	$w = 20$	$w = 30$	$w = 40$
2nd Stage: Number of Lagged Observations	-0.83	-1.01	-0.20	0.03	-0.15	-0.07	-0.04	0.01	-0.04
	[0.979]	[0.789]	[0.754]	[0.139]	[0.115]	[0.091]	[0.093]	[0.069]	[0.055]
N(Tch-Yrs)	3875	5831	7686	2136	3228	4386	1989	3004	4059
	425-Professional Sample								
2nd Stage: Number of Lagged Observations	0.38	-0.05	0.76	-0.03	-0.18	-0.09	-0.18	-0.10	-0.12
	[2.066]	[1.522]	[1.256]	[0.144]	[0.117]	[0.092]	[0.130]	[0.095]	[0.078]
N(Tch-Yrs)	3332	5067	6715	1789	2734	3752	1711	2621	3564

Note: Teacher-clustered standard errors in brackets. Bottom panel estimates based only on Professional teacher at 425 threshold. * ($p < 0.05$), ** ($p < 0.01$), *** ($p < 0.001$)

Table 15: Cumulative Effects of Observations

	TVAAS	TLM Math	TLM RLA
2nd Stage: Cumulative Number of Observations	-1.53***	-0.02***	-0.01*
	[0.045]	[0.005]	[0.004]
1st Stage: Cumulative Number of Policy-Assigned Observations	0.49***	0.83***	0.83***
	[0.015]	[0.025]	[0.021]
N (Tch-Yrs)	46412	13210	14204

Note: Teacher-clustered standard errors in brackets. Estimates produced by fuzzy full-sample RDDs using same controls as previous models. The predictor and instrument are, respectively, the total number of observations received and assigned through year *t* beginning in 2012-13. * ($p < 0.05$), ** ($p < 0.01$), *** ($p < 0.001$)

Table 16: Robustness Tests Concerning Loss of LOE5

	TVAAS			TLM Math			TLM RLA		
	$w = 20$	$w = 30$	$w = 40$	$w = 20$	$w = 30$	$w = 40$	$w = 20$	$w = 30$	$w = 40$
2nd Stage: Number of Observations	-0.29	0.16	1.09	-0.19	-0.03	-0.11	-0.05	-0.07	-0.08
	[1.113]	[0.913]	[0.778]	[0.242]	[0.177]	[0.105]	[0.120]	[0.079]	[0.065]
2nd Stage: From LOE 5 to 4 Interaction	0.80	0.28	0.82	-0.20	0.21	0.11	-0.07	-0.08	-0.12
	[1.472]	[1.184]	[1.135]	[0.528]	[0.435]	[0.383]	[0.136]	[0.123]	[0.112]
N (Tch-Yrs)	4037	6453	9138	1563	2506	3590	1506	2431	3401

Note: Teacher-clustered standard errors in brackets. Models include previously discussed controls. The instruments are crossing the 425 threshold for Apprentice and Professional teachers, and the interaction between these instruments and a dummy variable indicating if a teacher with an LOE of 4 in year $t - 1$ had an LOE of 5 in year $t - 2$. The second row of coefficients are interactions, not main effects. * ($p < 0.05$), ** ($p < 0.01$), *** ($p < 0.001$)

Table 17: Effects of Crossing LOE at Other Thresholds

	TVAAS			TLM Math			TLM RLA		
	w = 20	*w* = 30	*w* = 40	*w* = 20	*w* = 30	*w* = 40	*w* = 20	*w* = 30	*w* = 40
Crossing Prior 275 LOE Threshold	0.29	0.43	0.78	-0.01	-0.01	0.02	0.05	0.04	0.02
	[0.626]	[0.523]	[0.437]	[0.046]	[0.038]	[0.033]	[0.032]	[0.027]	[0.024]
N (Tch-Yrs)	5166	7745	10196	1415	2099	2748	1796	2756	3680
Crossing Prior 350 LOE Threshold	0.19	-0.02	-0.09	0.01	0.04	0.03	-0.01	< 0.01	> -0.01
	[0.393]	[0.327]	[0.286]	[0.034]	[0.029]	[0.025]	[0.020]	[0.017]	[0.015]
N (Tch-Yrs)	8646	12835	16896	2307	3438	4556	3729	5509	7191

Note: Teacher-clustered standard errors in brackets. Models include previously discussed controls, but the predictor of interest is crossing the LOE-cont 275 and 350 thresholds. * ($p < 0.05$), ** ($p < 0.01$), *** ($p < 0.001$)

Table 18: Heterogenous Effects by Grade Level

	TVAAS			TLM Math			TLM RLA		
	$w = 20$	$w = 30$	$w = 40$	$w = 20$	$w = 30$	$w = 40$	$w = 20$	$w = 30$	$w = 40$
K-5	-1.61*	-1.28**	-0.73	-0.07	-0.06	-0.03	-0.07	-0.04	-0.05
	[-2.85,-0.38]	[-2.21,-0.34]	[-1.48,0.02]	[-0.16,0.01]	[-0.14,0.01]	[-0.10,0.03]	[-0.15,0.01]	[-0.10,0.02]	[-0.10,0.00]
5-8	-1.52*	-1.23*	-0.77	-0.06	-0.09*	-0.09*	-0.03	-0.02	-0.03
	[-2.90,-0.14]	[-2.27,-0.19]	[-1.57,0.02]	[-0.15,0.03]	[-0.18,-0.01]	[-0.16,-0.02]	[-0.11,0.06]	[-0.08,0.05]	[-0.08,0.03]
9-12	-6.43***	-6.14***	-5.23***	-0.09*	-0.12**	-0.08*	-0.05	-0.04	-0.04
	[-8.47,-4.39]	[-7.66,-4.62]	[-6.41,-4.05]	[-0.17,-0.00]	[-0.20,-0.04]	[-0.15,-0.01]	[-0.13,0.03]	[-0.11,0.02]	[-0.10,0.01]
K-8	-1.10	-0.63	-0.26	-0.16*	-0.10	-0.07	-0.04	-0.01	-0.01
	[-2.33,0.13]	[-1.52,0.26]	[-1.00,0.48]	[-0.31,-0.01]	[-0.21,0.01]	[-0.16,0.02]	[-0.13,0.05]	[-0.08,0.06]	[-0.07,0.05]
K-12	-1.58	-2.83*	-1.46*	0.01	-0.04	-0.05	0.11	0.06	0.02
	[-4.20,1.04]	[-5.10,-0.56]	[-2.80,-0.12]	[-0.12,0.15]	[-0.17,0.10]	[-0.15,0.06]	[-0.16,0.39]	[-0.17,0.29]	[-0.05,0.09]
N (Tch-Yrs)	7053	11117	15622	3348	5205	7197	3143	4906	6783

Note: Standard errors clustered at teacher level. 95% confidence intervals in brackets. Models use previously discussed controls. * ($p < 0.05$), ** ($p < 0.01$), *** ($p < 0.001$)

Table 19: Heterogeneous Effects by Teacher Experience

	TVAAS			TLM Math			TLM RLA		
	w = 20	*w* = 30	*w* = 40	*w* = 20	*w* = 30	*w* = 40	*w* = 20	*w* = 30	*w* = 40
YrsExp [0, 5)	-2.52**	-2.47***	-2.19***	-0.13**	-0.09*	-0.07*	-0.05	-0.04	-0.06*
	[-4.02,-1.01]	[-3.74,-1.20]	[-3.25,-1.13]	[-0.21,-0.04]	[-0.16,-0.01]	[-0.13,-0.00]	[-0.13,0.03]	[-0.10,0.02]	[-0.11,-0.00]
YrsExp [5, 10)	-3.17**	-2.84***	-2.39***	-0.06	-0.02	-0.04	-0.04	-0.02	-0.03
	[-5.22,-1.12]	[-4.47,-1.20]	[-3.67,-1.11]	[-0.15,0.03]	[-0.10,0.06]	[-0.11,0.03]	[-0.13,0.03]	[-0.10,0.02]	[-0.11,-0.00]
YrsExp [10, .)	-3.28***	-3.00***	-2.35***	-0.10*	-0.08*	-0.07	-0.09	-0.05	-0.05
	[-5.08,-1.49]	[-4.32,-1.67]	[-3.47,-1.23]	[-0.19,-0.00]	[-0.17,-0.00]	[-0.15,0.01]	[-0.18,0.01]	[-0.12,0.02]	[-0.11,0.01]
N (Tch-Yrs)	7053	11117	15622	3348	5205	7197	3143	4906	6783

Note: Ibid.

Table 20: Heterogenous Effects by Administrator Effectiveness

	TVAAS			TLM Math			TLM RLA		
	w = 20	*w* = 30	*w* = 40	*w* = 20	*w* = 30	*w* = 40	*w* = 20	*w* = 30	*w* = 40
	School-Level Distribution of Admin LOE								
1st Qrt	-3.60***	-3.33***	-2.98***	-0.09	-0.10*	-0.08	-0.07*	-0.05	-0.04
	[-5.37,-1.84]	[-4.96,-1.69]	[-4.35,-1.62]	[-0.20,0.02]	[-0.19,-0.00]	[-0.17,0.00]	[-0.14,-0.01]	[-0.11,0.01]	[-0.09,0.02]
2nd Qrt	-3.08***	-2.67***	-2.26***	-0.07	-0.06	-0.06	-0.03	-0.03	-0.06
	[-4.76,-1.40]	[-3.94,-1.41]	[-3.34,-1.19]	[-0.15,0.02]	[-0.14,0.02]	[-0.14,0.01]	[-0.13,0.07]	[-0.10,0.04]	[-0.12,0.01]
3rd Qrt	-2.72**	-2.53***	-2.33***	-0.07	-0.07	-0.06	-0.01	-0.02	-0.02
	[-4.40,-1.04]	[-3.93,-1.12]	[-3.49,-1.17]	[-0.16,0.02]	[-0.15,0.02]	[-0.13,0.01]	[-0.10,0.07]	[-0.08,0.05]	[-0.08,0.04]
4th Qrt	-2.47**	-2.42***	-2.06***	-0.03	-0.02	-0.03	-0.01	-0.03	-0.04
	[-4.07,-0.88]	[-3.80,-1.04]	[-3.17,-0.94]	[-0.13,0.07]	[-0.11,0.07]	[-0.10,0.04]	[-0.09,0.08]	[-0.09,0.04]	[-0.10,0.02]
N (Tch-Yrs)	6968	10972	15390	3298	5125	7073	3100	4834	6676
	School-Level Distribution of Admin TEAM Scores								
1st Qrt	-3.46***	-2.97***	-2.48***	-0.07	-0.07	-0.05	-0.03	-0.03	-0.04
	[-5.14,-1.79]	[-4.37,-1.58]	[-3.65,-1.31]	[-0.15,0.01]	[-0.14,0.00]	[-0.12,0.01]	[-0.11,0.04]	[-0.09,0.03]	[-0.10,0.01]
2nd Qrt	-2.99**	-2.69***	-2.46***	-0.09	-0.07	-0.07	-0.03	-0.02	-0.03
	[-4.79,-1.19]	[-3.90,-1.49]	[-3.53,-1.39]	[-0.18,0.01]	[-0.15,0.02]	[-0.14,0.00]	[-0.12,0.05]	[-0.08,0.04]	[-0.09,0.02]
3rd Qrt	-3.17***	-2.66***	-2.19***	-0.06	-0.07	-0.07	-0.01	-0.02	-0.05
	[-4.87,-1.48]	[-4.00,-1.32]	[-3.32,-1.06]	[-0.16,0.03]	[-0.16,0.02]	[-0.14,0.01]	[-0.09,0.07]	[-0.08,0.05]	[-0.11,0.01]
4th Qrt	-2.74**	-2.38**	-2.06***	-0.03	-0.03	-0.02	-0.02	-0.02	-0.02
	[-4.73,-0.75]	[-3.87,-0.89]	[-3.29,-0.84]	[-0.13,0.06]	[-0.11,0.05]	[-0.09,0.05]	[-0.11,0.08]	[-0.09,0.05]	[-0.09,0.05]
N (Tch-Yrs)	6912	10881	15259	3265	5077	7008	3066	4782	6607

Note: Ibid.

Table 21: Heterogenous Effects by Administrator Skills

	TVAAS			TLM Math			TLM RLA		
	w = 20	*w* = 30	*w* = 40	*w* = 20	*w* = 30	*w* = 40	*w* = 20	*w* = 30	*w* = 40
	School-Level Distribution of Admin Skills as Teacher Evaluator								
1st Qrt	-3.55***	-2.90***	-2.26***	-0.11*	-0.08	-0.05	-0.08*	-0.06	-0.07*
	[-5.27,-1.83]	[-4.30,-1.50]	[-3.42,-1.09]	[-0.20,-0.02]	[-0.16,0.01]	[-0.12,0.02]	[-0.16,-0.01]	[-0.12,0.00]	[-0.13,-0.02]
2nd Qrt	-2.56**	-2.16**	-1.78**	-0.12*	-0.10*	-0.09**	-0.06	-0.05	-0.07*
	[-4.13,-0.98]	[-3.49,-0.83]	[-2.91,-0.65]	[-0.21,-0.03]	[-0.18,-0.02]	[-0.16,-0.03]	[-0.14,0.02]	[-0.12,0.02]	[-0.14,-0.00]
3rd Qrt	-2.72**	-2.43***	-1.76**	-0.06	-0.07	-0.07	-0.03	-0.04	-0.05
	[-4.44,-1.00]	[-3.86,-0.99]	[-2.93,-0.59]	[-0.17,0.04]	[-0.17,0.03]	[-0.15,0.01]	[-0.12,0.06]	[-0.11,0.03]	[-0.11,0.02]
4th Qrt	-2.89**	-2.17**	-1.64**	-0.07	-0.06	-0.04	-0.03	-0.05	-0.06
	[-5.03,-0.74]	[-3.59,-0.74]	[-2.82,-0.47]	[-0.17,0.02]	[-0.14,0.03]	[-0.12,0.03]	[-0.12,0.05]	[-0.12,0.02]	[-0.12,0.01]
N (Tch-Yrs)	6299	9948	14028	2973	4629	6412	2798	4373	6060
	School-Level Distribution of Admin Skills as Supporter of Teacher Professional Learning								
1st Qrt	-2.81***	-2.51***	-2.18***	-0.08	-0.07	-0.06	-0.10**	-0.07*	-0.08**
	[-4.36,-1.26]	[-3.81,-1.20]	[-3.30,-1.05]	[-0.17,0.00]	[-0.14,0.00]	[-0.12,0.01]	[-0.17,-0.02]	[-0.13,-0.00]	[-0.14,-0.02]
2nd Qrt	-3.90***	-3.24***	-2.56***	-0.08	-0.09*	-0.07*	-0.06	-0.04	-0.05
	[-5.82,-1.98]	[-4.79,-1.70]	[-3.88,-1.25]	[-0.17,0.01]	[-0.17,-0.01]	[-0.14,-0.00]	[-0.13,0.02]	[-0.10,0.03]	[-0.12,0.01]
3rd Qrt	-3.09**	-2.40***	-2.03***	-0.08	-0.07	-0.06	-0.07	-0.06	-0.07*
	[-4.98,-1.21]	[-3.63,-1.16]	[-3.11,-0.94]	[-0.19,0.02]	[-0.16,0.01]	[-0.13,0.01]	[-0.15,0.01]	[-0.13,0.01]	[-0.13,-0.01]
4th Qrt	-3.37**	-2.63**	-2.21**	-0.07	-0.06	-0.05	-0.04	-0.04	-0.05
	[-5.58,-1.16]	[-4.24,-1.02]	[-3.54,-0.89]	[-0.17,0.03]	[-0.14,0.02]	[-0.13,0.02]	[-0.13,0.05]	[-0.11,0.04]	[-0.12,0.02]
N (Tch-Yrs)	6360	10041	14159	2994	4663	6459	2816	4402	6100

Note: Ibid.

Table 22: Full-Sample RDD. Short-Term and Extended Effects

	TVAAS	TLM Math	TLM RLA	TVAAS	TLM Math	TLM RLA
	Contemporaneous Effects			Extended Effects		
2nd Stage: Number of Observations	-1.54***	-0.10***	-0.04*	-0.48*	-0.08**	< 0.01
	[0.212]	[0.022]	[0.017]	[0.232]	[0.024]	[0.019]
1st Stage: Number of Assigned Policy-Assigned Observations	0.49***	0.49***	0.50***	0.49***	0.48***	0.48***
	[0.015]	[0.026]	[0.025]	[0.018]	[0.031]	[0.028]
N (Tch-Yrs)	46412	15311	16470	24661	8976	9604

Note: Teacher-clustered standard errors in brackets. Models include previously discussed controls. The instrument is the minimum number of policy-assigned observations assigned to a teacher based on their lagged LOE and certification status. * ($p < 0.05$), ** ($p < 0.01$), *** ($p < 0.001$)

B. Tennessee Educator Acceleration Model Observation Rubrics

General Educator Rubric: Instruction

	Significantly Above Expectations (5)	At Expectations (3)	Significantly Below Expectations (1)
Standards and Objectives	• All learning objectives are clearly and explicitly communicated, connected to state standards, and referenced throughout lesson. • Sub-objectives are aligned and logically sequenced to the lesson's major objective. • Learning objectives are: (a) consistently connected to what students have previously learned, (b) known from life experiences, and (c) integrated with other disciplines. • Expectations for student performance are clear, demanding, and high. • There is evidence that most students demonstrate mastery of the daily objective that supports significant progress towards mastery of a standard.	• Most learning objectives are communicated, connected to state standards, and referenced throughout lesson. • Sub-objectives are mostly aligned to the lesson's major objective. • Learning objectives are connected to what students have previously learned. • Expectations for student performance are clear. • There is evidence that most students demonstrate mastery of the daily objective that supports significant progress towards mastery of a standard.	• Few learning objectives are communicated, connected to state standards, and referenced throughout lesson. • Sub-objectives are inconsistently aligned to the lesson's major objective. • Learning objectives are rarely connected to what students have previously learned. • Expectations for student performance are vague. • There is evidence that few students demonstrate mastery of the daily objective that supports significant progress towards mastery of a standard.
Motivating Students	• The teacher consistently organizes the content so that it is personally meaningful and relevant to students. • The teacher consistently develops learning experiences where inquiry, curiosity, and exploration are valued. • The teacher regularly reinforces and rewards effort.	• The teacher sometimes organizes the content so that it is personally meaningful and relevant to students. • The teacher sometimes develops learning experiences where inquiry, curiosity, and exploration are valued. • The teacher sometimes reinforces and rewards effort.	• The teacher rarely organizes the content so that it is personally meaningful and relevant to students. • The teacher rarely develops learning experiences where inquiry, curiosity, and exploration are valued. • The teacher rarely reinforces and rewards effort.
Presenting Instructional Content	Presentation of content always includes: • visuals that establish the purpose of the lesson, preview the organization of the lesson, and include internal summaries of the lesson; • examples, illustrations, analogies, and labels for new concepts and ideas; • effective modeling of thinking process by the teacher and/or students guided by the teacher to demonstrate performance expectations; • concise communication; • logical sequencing and segmenting; • all essential information; and • no irrelevant, confusing, or non-essential information.	Presentation of content most of the time includes: • visuals that establish the purpose of the lesson, preview the organization of the lesson, and include internal summaries of the lesson; • examples, illustrations, analogies, and labels for new concepts and ideas; • modeling by the teacher to demonstrate performance expectations; • concise communication; • logical sequencing and segmenting; • all essential information; and • no irrelevant, confusing, or non-essential information.	Presentation of content rarely includes: • visuals that establish the purpose of the lesson, preview the organization of the lesson, and include internal summaries of the lesson; • examples, illustrations, analogies, and labels for new concepts and ideas; • modeling by the teacher to demonstrate performance expectations; • concise communication; • logical sequencing and segmenting; • all essential information; and • relevant, coherent, or essential information.

1

General Educator Rubric: Instruction

	Significantly Above Expectations (5)	At Expectations (3)	Significantly Below Expectations (1)
Lesson Structure and Pacing 	• The lesson starts promptly. • The lesson's structure is coherent, with a beginning, middle, and end. • The lesson includes time for reflection. • Pacing is brisk and provides many opportunities for individual students who progress at different learning rates. • Routines for distributing materials are seamless. • No instructional time is lost during transitions.	• The lesson starts promptly. • The lesson's structure is coherent, with a beginning, middle, and end. • Pacing is appropriate and sometimes provides opportunities for students who progress at different learning rates. • Routines for distributing materials are efficient. • Little instructional time is lost during transitions.	• The lesson does not start promptly. • The lesson has a structure, but it may be missing closure or introductory elements. • Pacing is appropriate for less than half of the students and rarely provides opportunities for students who progress at different learning rates. • Routines for distributing materials are inefficient. • Considerable time is lost during transitions.
Activities and Materials 	• Activities and materials include all of the following: o support the lesson objectives, o are challenging, o sustain students' attention, o elicit a variety of thinking, o provide time for reflection, o are relevant to students' lives, o provide opportunities for student-to-student interaction, o induce student curiosity and suspense, o provide students with choices, o incorporate multimedia and technology, and o incorporate resources beyond the school curriculum texts (e.g., teacher-made materials, manipulatives, resources from museums, cultural centers, etc.). • In addition, sometimes activities are game-like, involve simulations, require creating products, and demand self-direction and self-monitoring. • The preponderance of activities demand complex thinking and analysis. • Texts and tasks are appropriately complex.	• Activities and materials include most of the following: o support the lesson objectives, o are challenging, o sustain students' attention, o elicit a variety of thinking; o provide time for reflection, o are relevant to students' lives, o provide opportunities for student-to-student interaction, o induce student curiosity and suspense; o provide students with choices, o incorporate multimedia and technology, and o incorporate resources beyond the school curriculum texts (e.g., teacher-made materials, manipulatives, resources from museums, cultural centers, etc.). • Texts and tasks are appropriately complex.	• Activities and materials include few of the following: o support the lesson objectives, o are challenging, o sustain students' attention, o elicit a variety of thinking, o provide time for reflection, o are relevant to students' lives, o provide opportunities for student to student interaction, o induce student curiosity and suspense, o provide students with choices, o incorporate multimedia and technology, and o incorporate resources beyond the school curriculum texts (e.g., teacher made materials, manipulatives, resources from museums, etc.).

General Educator Rubric: Instruction

	Significantly Above Expectations (5)	At Expectations (3)	Significantly Below Expectations (1)

Questioning	• Teacher questions are varied and high quality, providing a balanced mix of question types:
	  ○ knowledge and comprehension,
	  ○ application and analysis, and
	  ○ creation and evaluation.
	• Questions require students to regularly cite evidence throughout lesson.
	• Questions are consistently purposeful and coherent.
	• A high frequency of questions is asked.
	• Questions are consistently sequenced with attention to the instructional goals.
	• Questions regularly require active responses (e.g., whole class signaling, choral responses, written and shared responses, or group and individual answers).
	• Wait time (3-5 seconds) is consistently provided.
	• The teacher calls on volunteers and non-volunteers, and a balance of students based on ability and sex.
	• Students generate questions that lead to further inquiry and self-directed learning.
	• Questions regularly assess and advance student understanding.
	• When text is involved, majority of questions are text-based.
		• Teacher questions are varied and high quality providing for some, but not all, question types:
		  ○ knowledge and comprehension,
		  ○ application and analysis, and
		  ○ creation and evaluation.
		• Questions usually require students to cite evidence.
		• Questions are usually purposeful and coherent.
		• A moderate frequency of questions asked.
		• Questions are sometimes sequenced with attention to the instructional goals.
		• Questions sometimes require active responses (e.g., whole class signaling, choral responses, or group and individual answers).
		• Wait time is sometimes provided.
		• The teacher calls on volunteers and non-volunteers, and a balance of students based on ability and sex.
		• When text is involved, majority of questions are text-based.
			• Teacher questions are inconsistent in quality and include few question types:
			  ○ knowledge and comprehension,
			  ○ application and analysis, and
			  ○ creation and evaluation.
			• Questions are random and lack coherence.
			• A low frequency of questions is asked.
			• Questions are rarely sequenced with attention to the instructional goals.
			• Questions rarely require active responses (e.g., whole class signaling, choral responses, or group and individual answers).
			• Wait time is inconsistently provided.
			• The teacher mostly calls on volunteers and high-ability students.

Academic Feedback	• Oral and written feedback is consistently academically focused, frequent, high quality and references expectations. • Feedback is frequently given during guided practice and homework review. • The teacher circulates to prompt student thinking, assess each student's progress, and provide individual feedback. • Feedback from students is regularly used to monitor and adjust instruction. • Teacher engages students in giving specific and high-quality feedback to one another.	• Oral and written feedback is mostly academically focused, frequent, and mostly high quality. • Feedback is sometimes given during guided practice and homework review. • The teacher circulates during instructional activities to support engagement, and monitor student work. • Feedback from students is sometimes used to monitor and adjust instruction.	• The quality and timeliness of feedback is inconsistent. • Feedback is rarely given during guided practice and homework review. • The teacher circulates during instructional activities but monitors mostly behavior. • Feedback from students is rarely used to monitor or adjust instruction.

General Educator Rubric: Instruction

	Significantly Above Expectations (5)	At Expectations (3)	Significantly Below Expectations (1)
Grouping Students 	• The instructional grouping arrangements (either whole-class, small groups, pairs, individual; heterogeneous or homogenous ability) consistently maximize student understanding and learning efficiency. • All students in groups know their roles, responsibilities, and group work expectations. • All students participating in groups are held accountable for group work and individual work. • Instructional group composition is varied (e.g., race, gender, ability, and age) to best accomplish the goals of the lesson. • Instructional groups facilitate opportunities for students to set goals, reflect on, and evaluate their learning.	• The instructional grouping arrangements (either whole class, small groups, pairs, individual; heterogeneous or homogenous ability) adequately enhance student understanding and learning efficiency. • Most students in groups know their roles, responsibilities, and group work expectations. • Most students participating in groups are held accountable for group work and individual work. • Instructional group composition is varied (e.g., race, gender, ability, and age) most of the time to best accomplish the goals of the lesson.	• The instructional grouping arrangements (either whole-class, small groups, pairs, individual; heterogeneous or homogenous ability) inhibit student understanding and learning efficiency. • Few students in groups know their roles, responsibilities, and group work expectations. • Few students participating in groups are held accountable for group work and individual work. • Instructional group composition remains unchanged irrespective of the learning and instructional goals of a lesson.
Teacher Content Knowledge 	• Teacher displays extensive content knowledge of all the subjects she or he teaches. • Teacher regularly implements a variety of subject-specific instructional strategies to enhance student content knowledge. • The teacher regularly highlights key concepts and ideas and uses them as bases to connect other powerful ideas. • Limited content is taught in sufficient depth to allow for the development of understanding.	• Teacher displays accurate content knowledge of all the subjects he or she teaches. • Teacher sometimes implements subject-specific instructional strategies to enhance student content knowledge. • The teacher sometimes highlights key concepts and ideas and uses them as bases to connect other powerful ideas.	• Teacher displays under-developed content knowledge in several subject areas. • Teacher rarely implements subject-specific instructional strategies to enhance student content knowledge. • Teacher does not understand key concepts and ideas in the discipline and therefore presents content in a disconnected manner.
Teacher Knowledge of Students 	• Teacher practices display understanding of each student's anticipated learning difficulties. • Teacher practices regularly incorporate student interests and cultural heritage. • Teacher regularly provides differentiated instructional methods and content to ensure children have the opportunity to master what is being taught.	• Teacher practices display understanding of some student anticipated learning difficulties. • Teacher practices sometimes incorporate student interests and cultural heritage. • Teacher sometimes provides differentiated instructional methods and content to ensure children have the opportunity to master what is being taught.	• Teacher practices demonstrate minimal knowledge of students anticipated learning difficulties. • Teacher practices rarely incorporate student interests or cultural heritage. • Teacher practices demonstrate little differentiation of instructional methods or content.

General Educator Rubric: Instruction

	Significantly Above Expectations (5)	At Expectations (3)	Significantly Below Expectations (1)
Thinking 	• The teacher thoroughly teaches two or more types of thinking: ○ analytical thinking, where students analyze, compare and contrast, and evaluate and explain information; ○ practical thinking, where students use, apply, and implement what they learn in real-life scenarios; ○ creative thinking, where students create, design, imagine, and suppose; and ○ research-based thinking, where students explore and review a variety of ideas, models, and solutions to problems. • The teacher provides opportunities where students: ○ generate a variety of ideas and alternatives, ○ analyze problems from multiple perspectives and viewpoints, and ○ monitor their thinking to insure that they understand what they are learning, are attending to critical information, and are aware of the learning strategies that they are using and why.	• The teacher thoroughly teaches one or more types of thinking: ○ analytical thinking, where students analyze, compare and contrast, and evaluate and explain information; ○ practical thinking, where students use, apply, and implement what they learn in real-life scenarios; ○ creative thinking, where students create, design, imagine, and suppose; and ○ research-based thinking, where students explore and review a variety of ideas, models, and solutions to problems. • The teacher provides opportunities where students: ○ generate a variety of ideas and alternatives, and ○ analyze problems from multiple perspectives and viewpoints.	• The teacher implements no learning experiences that thoroughly teach any type of thinking. • The teacher provides no opportunities where students: ○ generate a variety of ideas and alternatives, or ○ analyze problems from multiple perspectives and viewpoints.
Problem-Solving 	The teacher implements activities that teach and reinforce three or more of the following problem-solving types: • Abstraction • Categorization • Drawing Conclusions/Justifying Solutions • Predicting Outcomes • Observing and Experimenting • Improving Solutions • Identifying Relevant/Irrelevant Information • Generating Ideas • Creating and Designing	The teacher implements activities that teach two of the following problem-solving types: • Abstraction • Categorization • Drawing Conclusions/Justifying Solution • Predicting Outcomes • Observing and Experimenting • Improving Solutions • Identifying Relevant/Irrelevant Information • Generating Ideas • Creating and Designing	The teacher implements no activities that teach the following problem-solving types: • Abstraction • Categorization • Drawing Conclusions/Justifying Solution • Predicting Outcomes • Observing and Experimenting • Improving Solutions • Identifying Relevant/Irrelevant Information • Generating Ideas • Creating and Designing

General Educator Rubric: Planning

	Significantly Above Expectations (5)	At Expectations (3)	Significantly Below Expectations (1)

Instructional Plans	Instructional plans include:
	• measurable and explicit goals aligned to state content standards;
	• activities, materials, and assessments that:
	  ○ are aligned to state standards,
	  ○ are sequenced from basic to complex,
	  ○ build on prior student knowledge, are relevant to students' lives, and integrate other disciplines, and
	  ○ provide appropriate time for student work, student reflection, and lesson unit and closure;
	• evidence that plan is appropriate for the age, knowledge, and interests of all learners; and
	• evidence that the plan provides regular opportunities to accommodate individual student needs.
		Instructional plans include:
		• goals aligned to state content standards,
		• activities, materials, and assessments that:
		  ○ are aligned to state standards,
		  ○ are sequenced from basic to complex,
		  ○ build on prior student knowledge, and
		  ○ provide appropriate time for student work, and lesson and unit closure;
		• evidence that plan is appropriate for the age, knowledge, and interests of most learners; and
		• evidence that the plan provides some opportunities to accommodate individual student needs.
			Instructional plans include:
			• few goals aligned to state content standards,
			• activities, materials, and assessments that:
			  ○ are rarely aligned to state standards,
			  ○ are rarely logically sequenced,
			  ○ rarely build on prior student knowledge, and
			  ○ inconsistently provide time for student work, and lesson and unit closure; and
			• little evidence that the plan provides some opportunities to accommodate individual student needs.

Student Work	Assignments require students to: • organize, interpret, analyze, synthesize, and evaluate information rather than reproduce it, • draw conclusions, make generalizations, and produce arguments that are supported through extended writing, and • connect what they are learning to experiences, observations, feelings, or situations significant in their daily lives both inside and outside of school.	Assignments require students to: • interpret information rather than reproduce it, • draw conclusions and support them through writing, and • connect what they are learning to prior learning and some life experiences.	Assignments require students to: • mostly reproduce information, • rarely draw conclusions and support them through writing, and • rarely connect what they are learning to prior learning or life experiences.
Assessment	Assessment plans: • are aligned with state content standards; • have clear measurement criteria; • measure student performance in more than three ways (e.g., in the form of a project, experiment, presentation, essay, short answer, or multiple choice test); • require extended written tasks; • are portfolio based with clear illustrations of student progress toward state content standards; and • include descriptions of how assessment results will be used to inform future instruction.	Assessment plans: • are aligned with state content standards; • have measurement criteria; • measure student performance in more than two ways (e.g., in the form of a project, experiment, presentation, essay, short answer, or multiple choice test); • require written tasks; and • include performance checks throughout the school year.	Assessment plans: • are rarely aligned with state content standards; • have ambiguous measurement criteria; • measure student performance in less than two ways (e.g., in the form of a project, experiment, presentation, essay, short answer, or multiple choice test); and • include performance checks, although the purpose of these checks is not clear.

General Educator Rubric: Environment

	Significantly Above Expectations (5)	At Expectations (3)	Significantly Below Expectations (1)
Expectations 	• Teacher sets high and demanding academic expectations for every student. • Teacher encourages students to learn from mistakes. • Teacher creates learning opportunities where all students can experience success. • Students take initiative and follow through with their own work. • Teacher optimizes instructional time, teaches more material, and demands better performance from every student.	• Teacher sets high and demanding academic expectations for every student. • Teacher encourages students to learn from mistakes. • Teacher creates learning opportunities where most students can experience success. • Students complete their work according to teacher expectations.	• Teacher expectations are not sufficiently high for every student. • Teacher creates an environment where mistakes and failure are not viewed as learning experiences. • Students demonstrate little or no pride in the quality of their work.
Managing Student Behavior 	• Students are consistently well behaved and on task. • Teacher and students establish clear rules for learning and behavior. • The teacher overlooks inconsequential behavior. • The teacher deals with students who have caused disruptions rather than the entire class. • The teacher attends to disruptions quickly and firmly.	• Students are mostly well behaved and on task, some minor learning disruptions may occur. • Teacher establishes rules for learning and behavior. • The teacher uses some techniques, such as social approval, contingent activities, and consequences, to maintain appropriate student behavior. • The teacher overlooks some inconsequential behavior, but at other times, stops the lesson to address it. • The teacher deals with students who have caused disruptions, yet sometimes he or she addresses the entire class.	• Students are not well behaved and are often off task. • Teacher establishes few rules for learning and behavior. • The teacher uses few techniques to maintain appropriate student behavior. • The teacher cannot distinguish between inconsequential behavior and inappropriate behavior. • Disruptions frequently interrupt instruction.
Environment 	The classroom: • welcomes all members and guests, • is organized and understandable to all students, • supplies, equipment, and resources are all easily and readily accessible, • displays student work that frequently changes, and • is arranged to promote individual and group learning.	The classroom: • welcomes most members and guests, • is organized and understandable to most students, • supplies, equipment, and resources are accessible, • displays student work, and • is arranged to promote individual and group learning.	The classroom: • is somewhat cold and uninviting, • is not well organized and understandable to students, • supplies, equipment, and resources are difficult to access, • does not display student work, and • is not arranged to promote group learning.
Respectful Culture 	• Teacher-student interactions demonstrate caring and respect for one another. • Students exhibit caring and respect for one another. • Positive relationships and interdependence characterize the classroom.	• Teacher-student interactions are generally friendly, but may reflect occasional inconsistencies, favoritism, or disregard for students' cultures. • Students exhibit respect for the teacher and are generally polite to each other. • Teacher is sometimes receptive to the interests and opinions of students.	• Teacher-student interactions are sometimes authoritarian, negative, or inappropriate. • Students exhibit disrespect for the teacher. • Student interaction is characterized by conflict, sarcasm, or put-downs. • Teacher is not receptive to interests and opinions of students.

Professionalism Rubric

	Significantly Above Expectations (5)	At Expectations (3)	Significantly Below Expectations (1)
Professional Growth and Learning 	• Uses feedback from observations and self-assessment to significantly improve performance in identified areas of need • Consistently prepared and highly engaged in professional learning opportunities • Engages in evaluation process with eagerness by seeking out feedback from both supervisors and colleagues • Consistently self-reflects on evidence of instruction, accurately matching evidence to the rubric in both areas of strength and areas of growth	• Uses feedback from observations and self-assessment to implement and reflect on personal improvement strategies • Prepared and engaged in professional learning opportunities • Engages in evaluation process with evidence of focus on improving practice and openness to feedback • Self-reflections on evidence on instruction largely match the expectations of the rubric	• Inconsistently uses feedback from observations to improve and demonstrates little evidence of growth on targeted indicators • Unprepared or disengaged in professional learning opportunities provided • Engages in evaluation process without evidence of focus on continuous improvement of practice. • Self-reflections do no match the expectations of the rubric or assessment of the evaluator
Use of Data 	• Systematically and consistently utilizes formative and summative school and individual student achievement data to: ○ Analyze the strengths and weaknesses of all his/her students, ○ Plan, implement, and assess instructional strategies to increase student achievement and decrease achievement gaps between subgroups of students ○ Plan future instructional units based on the analysis of his/her students' work ○ Reflect on use of instructional strategies that led or impeded student learning	• Utilizes student achievement data to address strengths and weaknesses of students and guide instructional decisions to increase student achievement • Analyzes student work to guide planning of instructional units	• Rarely utilizes student achievement data to address strengths and weaknesses of students to guide instructional decisions related to student achievement
School and Community Involvement 	• Regularly organizes and leads school activities and events that positively impact school results and culture • Always adheres to school and district personnel policies and serves as a leader and model for others • Regularly works with peers to contribute to a safe and orderly learning environment and actively facilitates improvement in school-wide culture	• Regularly supports and contributes to school activities and events • Regularly adheres to school and district personnel policies • Regularly works with peers to contribute to a safe and orderly learning environment	• Rarely supports school activities and events. • Inconsistently adheres to school and district personnel policies • Rarely works with peers to contribute to a safe and orderly learning environment

Professionalism Rubric

	Significantly Above Expectations (5)	At Expectations (3)	Significantly Below Expectations (1)
Leadership 	Actively and consistently contributes to the school community by assisting and/or mentoring others, including successful engagement in three or more of the following: • Collaborative planning with subject and/or grade level teams • Actively leading in a professional learning community • Coaching/mentoring • Supervising clinical experiences • Leading data-driven professional opportunities	Contributes to the school community by assisting others, including at least two of the following: • Collaborative planning with subject and/or grade level teams, • Actively participating in a professional learning community, • Coaching/mentoring • Supervising clinical experiences	Inconsistently contributes to the school community by assisting and/or mentoring others

C. Relationships Between Observations and TEAM Scores

Conventional wisdom and the TEAM theory of action implies more observations should improve observational scores (i.e. TEAM scores). However, TEAM scores are susceptible to a source of bias that cannot affect student achievement: observer bias. There are at least two reasons why observers may generate biased scores. First, observers have an incentive to show that teachers exhibit large within-year growth ("strategic rating"). Most observers are school administrators, and Tennessee school administrators receive their own observational ratings from an administration supervisor (Tennessee Board of Education, 2013). School administrator observational ratings are based in part on the performance of their teachers, and teacher TEAM scores are the portion of teacher performance a school administrator can directly control. Thus, school administrators may take advantage of this system and strategically rate teachers assigned more observations lower on the first observation, only to rate them higher on later observations. Doing so would suggest teachers experience within-year "growth." Second, observers almost certainly know teachers assigned more observations have relatively lower LOE. Observers may rate teachers in such a way that confirms an un/ conscious impression: teachers in lower LOE are worse teachers and their observational ratings should reflect this. In this section I explore these sources of observer bias and present evidence that observer bias is present.

Tests presented in this appendix use teachers of tested and untested subjects, whereas analyses in the body of the dissertation almost always use the former. In the event unobserved heterogeneity exists across teachers of tested and untested subjects, I check the balance of covariates in this new sample using equation 7. Tables 23 and 24 show no evidence of imbalanced covariates at either threshold.

Table 25 displays estimates produced by equations 1 and 2 when the outcome is TEAM scores. The association between observations and TEAM scores is approximately one-quarter of the standard deviation of the annual change in TEAM scores (i.e. -0.11 units on the TEAM scale). A teacher's lagged TEAM score is a right-hand side variable in all models in this appendix. Table 26 shows the relationship between observations and TEAM scores separated by threshold resemble the results in Table 25.

I test for observer bias by examining scores assigned during the first occasion that an observer evaluated a teacher. The question is whether this score is influenced by the number of times the teacher is supposed to be observed. It is important to point out that any such influence on the first score received cannot be a genuine treatment effect. The first observation score cannot be influenced by treatment effects because the teacher has not had time to respond to the first observation. Instead, these estimates pick up the effect of a teacher's *assignment* to receive more observations—i.e., they indicate that the instrument influences the outcome other than through the observations.

To explore observer bias I modify equations 1 and 2. First, I replace the outcome in equation 1 with the first observational rating a teacher receives in a school year. I also add two controls to equations 1 and 2: the month of the first observation and the domains rated on the first observation. It is plausible that the month during which a teacher receives her first observation is correlated with her performance (e.g. observers may want to postpone difficult observations). It may also be the case that observers tend to rate teachers in one domain more harshly than another. I also estimate some effects using a sample restricted to teachers receiving more than one observation. It may be unwise to compare the "first" observation received by a

teacher whose first observation is her only observation, to teachers who receive more than one observation. Thus, I drop teachers from some analytical samples if they receive one observation.

There is clear evidence suggesting observer bias exists. The top-left panel of Table 27 reprints estimates from Table 25, and the top-right panel of Table 27 uses equations 1 and 2 with the restricted sample. Estimates produced by the restricted sample resemble those produced by the unrestricted sample. The second panel of Table 27 shows the first rating generated for teachers receiving an additional observation is systematically lower than teachers assigned fewer observations. This is true in the unrestricted and restricted samples and strongly suggests rater bias is present.

As mentioned in the beginning of this appendix, there are at least two sources of observer bias: school administrator desire to strategically rate teachers so ratings exhibit large within-year growth, and the influence of LOE on rating behaviors. In the remainder of this appendix I explore whether evidence suggests these are indeed the sources of observer bias. For the purposes of this dissertation, these explorations are unnecessary. It is enough to know that TEAM scores should not be used as outcomes in my main analyses. However, identifying sources of observer bias is of interest in its own right. Evidence of either source would imply policy is biasing observation scores, something policymakers and practitioners would be eager to correct.

I test for biased, strategic rating (e.g. low to high ratings) by observers in two ways. I estimate the relationship between an additional observation and two new outcomes: the last observation score received, and the difference between the first and summative observation score. If observers initially rate teachers assigned more observations lower, only to systematically rate them higher on subsequent observations, the last observation received by

teachers assigned more observations should be higher than teachers assigned fewer observations. The second new outcome measures the atypicality of a first score relative to the summative TEAM score. A large positive (negative) value of this atypicality measure would suggest the first score was atypically lower (higher) than the average observational score. If administrators engage in low-to-high scoring behavior, the atypicality measure of teachers assigned more observations will be higher than the atypicality measure of teachers assigned fewer observations. To test these two hypotheses, I replace the outcome in equation 1 with each of these two new outcomes (these models control for lagged summative TEAM score). When the outcome is the last observation score a teacher received, I also control for the month of the last observation and domains rated during this observation. Similarly, when the outcome is the atypicality measure I control for the month of the first observation and domains rated during that observation.

The penultimate and bottom panels in Table 27 display results from these two tests using unrestricted and restricted samples (i.e. teachers receiving more than one observation). There is no evidence supporting my hypotheses that administrators engage in the low-to-high strategic rating behavior. The penultimate panel in Table 27 shows the association between the assignment of an additional observation and teacher's last rating is approximately is nearly the same as that estimated when the outcome was the first rating, suggesting the relationship between observer bias and the first observational rating may persist. The final panel shows the assignment of an additional observation is not predictive of a more atypical first rating score.

If observers consciously or unconsciously assume that teachers who received low LOE's in the past should receive lower ratings again, I should observe that teachers just below the 275 (350) lagged LOE-cont thresholds systematically receive a lower first observation score than teachers just above the threshold. Importantly, there are no policy-assigned discontinuities in the

observation schedule at these thresholds. The treatment of interest in this supplemental analysis is assignment to an adjacent, lower LOE. To assess the effect of assignment to a lower LOE I use the following model:

$$(D1)\quad TEAM1st_{ijt} = \lambda b_{ijt} + \breve{\boldsymbol{A}}h(\cdot) + \breve{\boldsymbol{B}}\boldsymbol{X}_{ijt} + \breve{\boldsymbol{C}}\boldsymbol{S}_{jt} + \gamma_t + v_{ijt}, \qquad \left|LOE_{ijt}\right| \leq w$$

Where $TEAM1st_{ijt}$ is the first observational rating received by teacher *i* in school *j* in year *t*. LOE_{ijt} is now centered at values of 275 (350) and *w* is the bandwidth. b_{ijt} is an indicator signaling if the teacher is below the 275 (350) threshold. All equation 1 and 2 controls are also in equation D1. But, equation D1 also includes a fixed effect for the domains rated on the first observation and fixed effect for the month of the first observation. λ is the coefficient of interest. Equation D1 does not represent a pooled RDD, but two separate one-stage RDDs. In results not shown, all covariates in $\boldsymbol{X}_{ijt}$ balance at the 275 and 350 thresholds.

If assignment to a lower LOE negatively influences observer bias, results in Table 28 should be negative. Instead, the effect of assignment to LOE2 instead of LOE3 on the first observation score is positive (see top panel), the opposite of what is expected if observer bias exists due to LOE assignment. As seen in the bottom panel of Table 28, there is also no evidence at the 375-threshold observer bias exists due to teacher LOE. If observer bias due to teacher assignment is not present at these two thresholds, it is unlikely the observer bias at the 200 and 425 thresholds is a response to lower teacher LOE.

Robustness tests presented in this appendix suggest observer bias is not a direct response to teacher lagged LOE per se. Furthermore, covariate balance tests presented in Tables 23 and 24 show teachers on either side of the 200 or 425 thresholds are similar with respect to time-invariant and pre-treatment characteristics. If teachers on either side of these thresholds appear

similar, and the most obvious difference between teachers on either side of these thresholds (i.e. different lagged LOE) does not account for observer bias, would could drive this source of bias?

Teachers on either side of the 200 or 425 thresholds differ in one more way: TDOE assigns teachers below these thresholds more observations than teachers above these thresholds. It may be the case that the TDOE assignment of observations is the driver of observer bias. Observers know that teachers just below these thresholds should receive more observations, and this information may un/ consciously lead observers to decide that teachers below the threshold deserve lower scores. This explanation is speculative, but evidence presented thus far supports this proposition.

There may be other explanations that can account for the relationships discussed in this appendix. A thorough investigation of the sources of observer bias are beyond the scope of this dissertation, but such work would almost certainly be of interest to practitioners. By pinpointing the sources of bias, practitioners may be able to develop policies/ interventions that can mitigate the problem.

Table 23: Covariate Balance Tests at LOE 200 Threshold. DV= TEAM

Covariate	$w = 20$	$w = 30$	$w = 40$
Experience: App	0.46	0.19	0.05
	[0.832]	[0.785]	[0.769]
Experience: Prof	-1.16	-0.63	-0.55
	[1.237]	[1.028]	[0.941]
Female: App	-0.16	-0.14	-0.11
	[0.094]	[0.080]	[0.072]
Female: Prof	0.01	-0.01	> -0.01
	[0.068]	[0.057]	[0.051]
BA+: App	-0.02	-0.14	-0.10
	[0.092]	[0.078]	[0.071]
BA+: Prof	0.05	0.06	0.06
	[0.069]	[0.057]	[0.052]
Black: App	0.10	0.12	0.09
	[0.069]	[0.060]	[0.056]
Black: Prof	-0.03	-0.02	0.01
	[0.057]	[0.047]	[0.043]
N (Tch-Yrs)	1322	2181	3384

Note: Estimates represent the total predicted change in the outcome. None of these estimates are interactions. Standard errors, clustered at teacher level, in brackets. Records are included in these covariate balance checks if used in models where the outcome is TEAM scores. OLS estimator employed to estimate all coefficients. BA+ is a binary variable indicating if a teacher reported having a degree higher than a BA/ BS. Black is an indicator signaling whether the teacher reported her ethnicity/ race as Black or White.

Table 24: Covariate Balance Tests at LOE 425 Threshold. DV=TEAM

Covariate	$w = 20$	$w = 30$	$w = 40$
Experience: App	0.13	0.12	0.03
	[0.297]	[0.228]	[0.192]
Experience: Prof	-0.02	0.09	0.21
	[0.357]	[0.295]	[0.257]
Female: App	-0.01	0.01	> -0.01
	[0.046]	[0.037]	[0.033]
Female: Prof	0.02	0.01	< 0.01
	[0.016]	[0.013]	[0.012]
BA+: App	-0.06	-0.05	-0.02
	[0.051]	[0.042]	[0.036]
BA+: Prof	0.01	0.01	-0.01
	[0.020]	[0.016]	[0.014]
Black: App	-0.02	-0.02	-0.01
	[0.021]	[0.018]	[0.016]
Black: Prof	> -0.01	< 0.01	< 0.01
	[0.009]	[0.008]	[0.006]
N (Tch-Yrs)	24872	36731	47817

Note: Ibid.

Table 25: Local RDD. DV=TEAM

	$w = 20$	$w = 30$	$w = 40$
2nd Stage: Number of Observations	-0.12***	-0.12***	-0.10***
	[0.017]	[0.014]	[0.012]
1st Stage:			
App Below LOE 200	0.86***	0.81***	1.00***
	[0.243]	[0.120]	[0.210]
Prof Below LOE 200	1.89***	1.91***	1.88***
	[0.152]	[0.131]	[0.121]
App Below LOE 425	1.26***	1.26***	1.36***
	[0.138]	[0.111]	[0.095]
Prof Below LOE 425	0.66***	0.65***	0.68***
	[0.034]	[0.027]	[0.023]
Prof	-0.50***	-0.45***	-0.34***
	[0.102]	[0.083]	[0.071]
Intercept	3.08***	2.94***	2.69***
	[0.175]	[0.143]	[0.124]
N(Tch-Yrs)	26194	38912	51201

Note: Teacher-clustered standard errors are in brackets. All models include the one-year lag of the outcome, teacher demographics including certification status, controls for the distribution of teacher effectiveness at the school level, a second order polynomial of pooled LOE interacted with teacher certification status, and year fixed effects. Each 1st stage estimate represents the total effect of crossing a threshold for each teacher group, none of the 1st stage estimate are interactions. * ($p < 0.05$), ** ($p < 0.01$), *** ($p < 0.001$)

Table 26: Effects by Thresholds. DV=TEAM

	$w = 20$	$w = 30$	$w = 40$
	Teachers Surrounding LOE 425 Only		
2nd Stage: Number of Observations	-0.10***	-0.10***	-0.09***
	[0.022]	[0.018]	[0.015]
N (Tch-Yrs)	24872	36731	47817
	Teachers Surrounding LOE 200 Only		
2nd Stage: Number of Observations	-0.14	-0.04	-0.08
	[0.117]	[0.105]	[0.092]
N (Tch-Yrs)	1322	2181	3384

Note: Ibid.

Table 27: Exploring Rater Bias. DV=First or Last TEAM Ratings Received

	$w = 20$	$w = 30$	$w = 40$	$w = 20$	$w = 30$	$w = 40$
	Any Number of Obs			Observations Received > 1		
	DV = TEAM Score					
2nd Stage: Number of Observations	-0.12***	-0.12***	-0.10***	-0.16**	-0.15***	-0.13***
	[0.017]	[0.014]	[0.012]	[0.058]	[0.044]	[0.037]
	DV = 1st Rating					
	-0.14***	-0.15***	-0.12***	-0.24**	-0.22***	-0.19***
	[0.035]	[0.029]	[0.023]	[0.077]	[0.059]	[0.047]
	DV = Last Rating					
	-0.15***	-0.16***	-0.13***	-0.25**	-0.26***	-0.22***
	[0.038]	[0.031]	[0.025]	[0.083]	[0.064]	[0.053]
	DV = 1st Rating – TEAM Score					
	-0.01	-0.02	-0.01	-0.05	-0.05	-0.05
	[0.020]	[0.017]	[0.013]	[0.044]	[0.035]	[0.028]

Note: Teacher-clustered standard errors are in brackets. All models include the one-year lag of the TEAM score, teacher demographics including certification status, controls for the distribution of teacher effectiveness at the school level, a second order polynomial of pooled LOE interacted with teacher certification status, and year fixed effects. Last three models restricted to teachers receiving more than one classroom visit. Last two models control for month of the first or last observation and the domains rated during that observation. These models use the pooled running variable and include teachers surrounding both the 200 and 425 thresholds. * ($p < 0.05$), ** ($p < 0.01$), *** ($p < 0.001$)

Table 28: Effects of Crossing 275 and 300 Thresholds. DV= 1st Observation Score

	DV=1st TEAM Score if Num Obs > 1		
	$w = 20$	$w = 30$	$w = 40$
Crossing Prior 275 LOE Threshold	0.04	0.05	0.05*
	[0.032]	[0.026]	[0.023]
N (Tch-Yrs)	9529	14227	18570
Crossing Prior 350 LOE Threshold	-0.01	-0.01	-0.01
	[0.023]	[0.018]	[0.016]
N (Tch-Yrs)	15211	22982	30528

Note: Teacher-clustered standard errors in brackets. All models include previously discussed controls. The predictor of interest is crossing the 275 or 350 thresholds. * ($p < 0.05$), ** ($p < 0.01$), *** ($p < 0.001$)

D. Discontinuities Surrounding LOE Multiples of Five

LOE-cont is the running variable is the RDD. If manipulation of the running variable is present, this threatens internal validity. However, standard tests for manipulation may falsely suggest manipulation of LOE-cont is present.

Conventional tests of manipulation compare the probability density function (PDF) of the running variable as it approaches a cut score from the left to the PDF of the running variable as it approaches the cut score from the right. A relatively large difference between these PDF estimates is evidence of manipulation. However, this type of discontinuity in LOE-cont is expected since it is approximately continuous, invalidating conventional tests.

LOE-cont scores are a weighted average determined by three components, and two components (achievement and growth scores) are integer variables in [1, 5]. Five is the least common multiple of weights applied to these two integer variables, so all linear combinations of these two components are multiples of five. Any LOE deviating from a multiple of five only does so due to summative observational ratings since this LOE component is composed of rational numbers. Thus, the distribution of summative observational ratings entirely determines the continuity of LOE scores. Moreover, it seems like there should be a preponderance of LOE scores at multiples of five.

Figure 2 is a histogram of continuous LOE. Figure 3 is the distribution of these same scores transformed via modulus five (LOEmod5). All LOE multiples of 5 are transformed to 0 in LOEmod5. Figure 3 shows concentrations of the PDF at multiples of five, as expected. Considering approximately continuous properties of LOE-cont, I remove the two integer components from LOE-cont before testing for manipulation.

Figure 2: Distribution of Lagged Continuous LOE

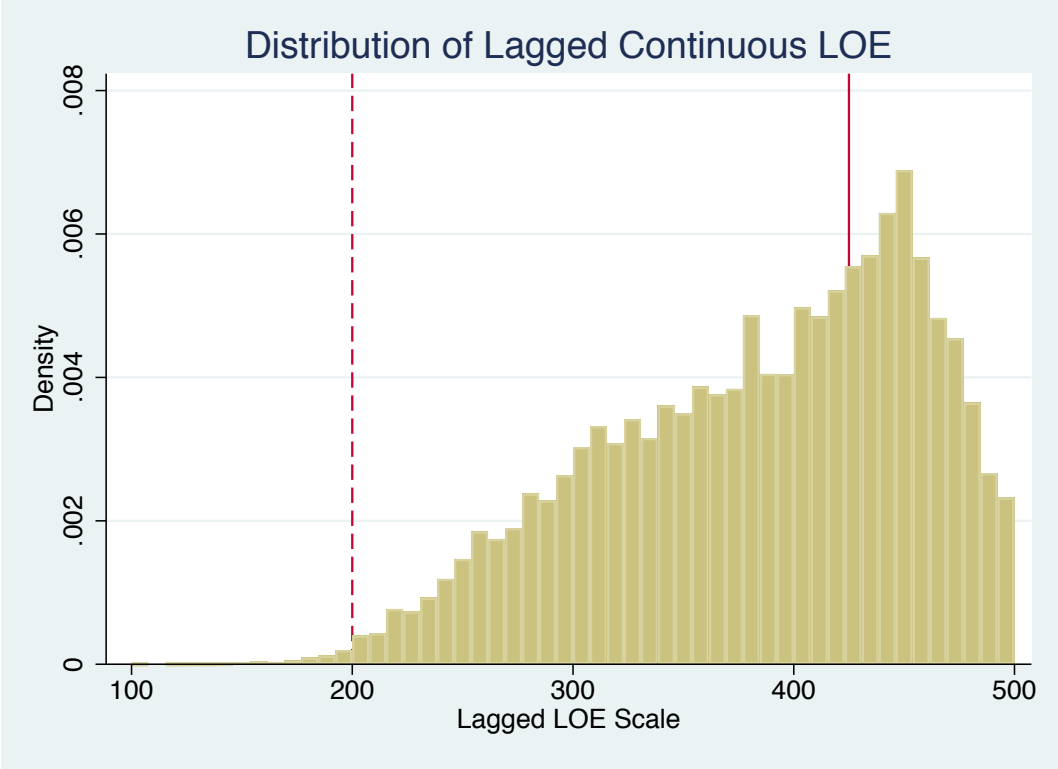
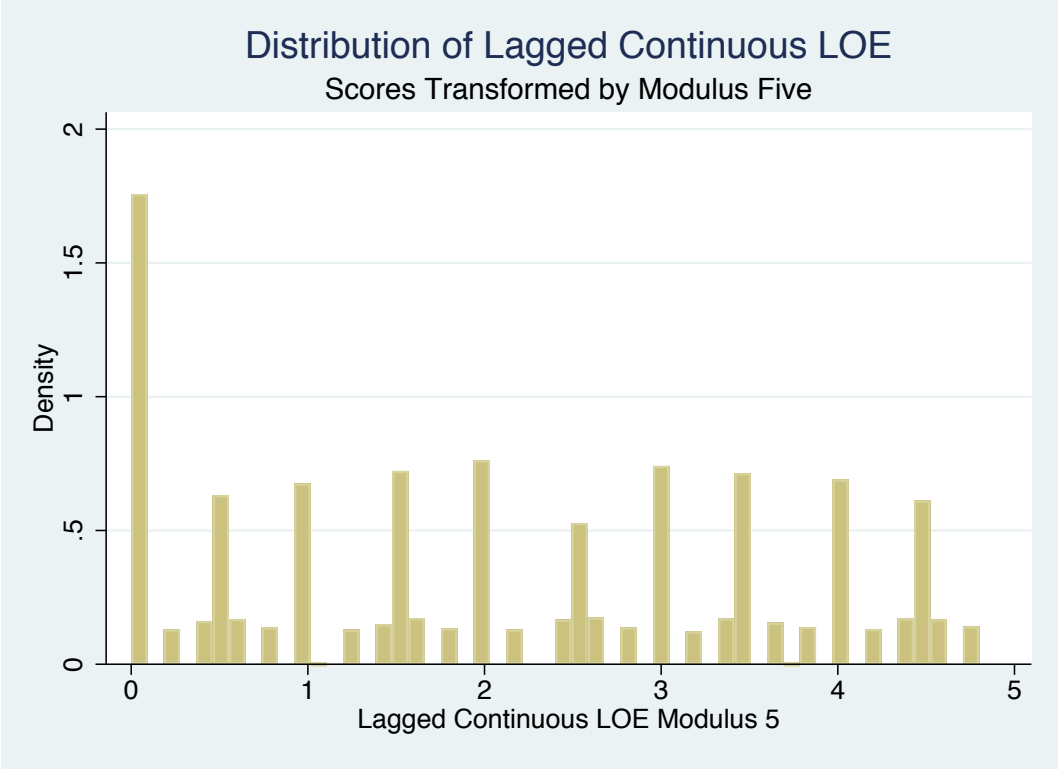


Figure 3: Distribution of Lagged Continuous LOE



E. Tennessee Educator Survey Items

Operationalized Variables	Survey Items	Scales	Years Administered	Operationalizations
Sum: Svy PD Hours (pdhrs)	Content: In-depth study of topics in my subjects	None; 1-5 Hours; 6-20 Hours; 21-40 Hours; More than 40 Hours	2013, 2014, 2015	Converted to hours by assigning each response to the lower bound of each item response interval (e.g. the response "1-5" assigned to 1, response "More than 40" assigned to 40). Add together all responses across these items within a year to produce a PD hours sum.
	Preparing students to take the TCAP			
	Preparing students to take the CRA and/or writing assessments (2014, 2015 only)			
	Analyzing and interpreting student assessment results			
	Classroom organization			
	Teaching special student populations (e.g., English language learners and students)			
	Student behavior management			
	Addressing students' socio-emotional development and/or student behavior			
	Reviewing standards and curriculum to determine learning outcomes for my students			
	Pedagogy: Strategies for teaching my subject(s)			

Sum: Svy Tch Collab (tchcollab)	Met with other teachers to discuss standards, instruction, and/or student learning Met with the whole faculty at my school Worked with other teachers to develop materials or activities for particular classes Observed another teacher's classroom Reviewed student assessment data with other teachers to make instructional decisions	Never; About once a semester; About once a month; Two or three times a month; About once a week; More than once a week	2015	Responses converted into collaborative meetings per year. Never = 0, semesterly = 2, monthly = 7, two or three times monthly = 14, weekly = 28, more than weekly = 56. Add all responses across items within a year to produce a sum.
Sum: Svy Exerted More Effort (effortsum)	Focusing on the content covered by TCAP Engaging in other self-selected professional development opportunities to improve my content knowledge and/or teaching skills Reflecting on and discussing teaching and learning with my inquiry team or other teachers, coaches, etc. Tutoring individuals or small groups of students outside of class time Engaging in informal self-directed learning (e.g., reading a mathematics	Less time and effort than last year; The same time and effort as last year; More time and effort than last year; Not applicable	2013, 2014	Assigned "Less time" and "The same time" to 0, "More time" to 1. Add together all responses across these items within a year to produce a sum.

	education journal, using the Internet to enrich knowledge and skills)			
	Re-teaching topics or skills based on students' performance on classroom tests			
	Assigning or reassigning students to groups within my class			
	Preparing lessons			
	Differentiating instruction to address individual student needs			
	Communicating with parents orally or in writing			
	Attending district- or school-sponsored workshops			
	Integrating material from multiple subjects into lessons I teach (e.g., incorporating mathematics content into science or social studies classes)			
	Completing tasks required for teaching observations and evaluation activities			
	Disciplining students			
Svy Hrs Improved Instruction (insthrs)	Approximately how much time have you invested so far during the 2013-2014 school year in efforts to improve your instructional practices?	1-10 hours; 11-20 hours; 21-40 hours; 41-60 hours; 61-	2014	Converted to hours by assigning each response to the lower bound of each response interval

		80 hours; 81-100 hours; More than 100 hours		(e.g. the response "1-10" assigned to 1, response "More than 100" assigned to 100)
Sum: Svy Hrs Prepped for Obs (obshrs)	How much TOTAL TIME have you spent on the following activities related to observations of your teaching during this school year?	None; Less than 1 hour; 1 to 2 hours; 2 to 3 hours; 3 to 5 hours; Over 5 hours	2013, 2014	Converted to hours by assigning each response to the lower bound of each response interval (e.g. the response "Less than 1 hour" assigned to 0.5, response "1 to 2" assigned to 1, "Over 5 hours" assigned to 5)
Evaluations Will Improve Teaching (imprvtch)	In general, the teacher evaluation process used in my school has led to improvements in my teaching. (2015 only)	Strongly Disagree; Disagree; Agree: Strongly Agree	2013, 2014, 2015	Strongly Disagree, Disagree assigned to 0; Agree, Strongly Agree assigned to 1.
	The teacher evaluation process used in my school will improve my teaching. (2013, 2014 only)			
Post-Observation Feedback is Useful (fbuseful)	School leadership provides useful feedback about my instructional practices. (2015 only)	Strongly Disagree;	2014, 2015	For the 2014 responses, if a teacher Agreed or

	Respondents were first asked if they were observed by any of the following: Principal, Assistant/ Vice Principal, Department Head, Senior Teacher/ Mentor/ Lead Teacher, Instructional Coach, Observer not working at their school, Other. For each role group selected, respondents were then provided this item: "This observer provided useful feedback about my teaching." (2014 only)	Disagree; Agree; Strongly Agree		Strongly Agreed that any one of the role groups provided useful feedback the variable was coded as 1, the variable was coded as 0 if the teacher selected only Disagree or Strongly Disagree. For the 2015 responses Strongly Disagree and Disagree were assigned to 0, Agree and Strongly Agree assigned to 1.
Changed Teaching due to Evaluations (chngtch)	I made changes to my teaching based on my evaluation results	Strongly Disagree; Disagree; Agree: Strongly Agree	2014	Strongly Disagree, Disagree assigned to 0; Agree, Strongly Agree assigned to 1.
Observers are Qualified (obsqual)	My observers are qualified to evaluate my teaching		2013, 2014	
Evaluations are Fair (faireval)	The processes used to conduct my teacher evaluation are fair to me		2013, 2014	

F. Sensitivity of Instrument Validity to Treatment of Survey Measures

I used measures based on survey items to check for the presence of two threats to the internal validity of my instruments: an impetus to improve and psychological boost. I concluded the instruments are not threatened by either psychological effect. However, findings in Tables 8 – 10 may be sensitive to my treatment of survey items. In this section I explore the sensitivity of findings from the impetus to improve and psychological boost tests to different treatments of survey items. These sensitivity tests yield qualitatively similar results to those discussed in Chapter 4.

Sensitivity of Tests to an Impetus to Improve

The original tests for an impetus to improve used five different measures based on survey items: hours spent in PD (*PDhrs*), frequency of teacher collaboration (*tchcollab*), hours spent improving instruction (*insthrs*), hours spent preparing for observations (*obshrs*), and effort invested in different activities (*effortsum*). The first four measures are based on survey items asking teachers about the amount of time spent on various activities. Each response to the original survey items[28] is an interval/ frequency. For example, in response to items contributing to *PDhrs* teachers could choose: None, 1-5 Hours, 6-20 Hours, etc. The original operationalization of these items converted these ordinal responses onto a continuous scale by assigning each response the lower boundary of the chosen interval/ frequency. Thus, if a teacher chose the "1-5 Hours" response this was converted to a value of one. The converted responses of items associated with *PDhrs* and *tchcollab* were then collapsed into single measures by adding

[28] Despite the ordinal scale of these outcomes there is no evidence supporting the parallel regressions assumption.

all responses together, respectively. The *insthrs* and *obshrs* were each measured by a single item. Responses to original *effortsum* items were not intervals/ frequencies. These responses asked if a teacher exerted less, the same, or more effort on a particular activity. I originally operationalized the lowest and middle categories to zero, and highest category to one, before collapsing all *effortsum* responses into a sum.

In this section I check the sensitivity of tests regarding an impetus to improve to different operationalizations of individual survey items and collapsing functions. A new version of *PDhrs*, *tchcollab, insthrs,* and *obshrs* items assigns original responses to the higher boundary of the chosen interval/ frequency. I refer to this as the *MAX* conversion. A new version of *effortsum* items assigns the middle category to one instead of zero, which I refer to as the *HI* conversion. I also collapse multiple items measuring the same impetus to improve by taking means and sums. These new operationalizations yield eleven new impetus to improve items: *PDhrsMAX, PDhrsmn, PDhrsMAXmn, tchcollabMAX, tchcollabmn, tchcollabMAXmn, effortsumHI, effortmn, effortHImn, insthrsMAX,* and *obshrsMAX.* All *MAX* (*HI*) measures are based on the *MAX* (*HI*) conversion and *mn* items are based on means instead of sums. Collapsed measures without "*mn"* in the label are based on sums.

Results from these sensitivity tests are in Tables 29 and 30. Results in Table 29 are based on the pooled sample and use all four instruments whereas findings in Table 30 only use 425-threshold samples and instruments. Results in both tables are qualitatively similar to those in Tables 8 and 9. Moreover, relationships between individual instruments and impetus to improve measures resemble those discussed in Chapter 4. In a bandwidth of 20 the 200-Apprentice and 200-Professional instruments positively predict all versions of *tchcollab*. The 200-Apprentice instrument negatively predicts all versions of *effortsum* in a bandwidth of 20. The 200-

Professional and 425-Professional instruments positively predict *obshrsMAX*. And again, the 425-threshold instruments are unrelated to impetus to improve measures in the 425-threshold sample (see Table 30).

Sensitivity of Tests to a Psychological Boost

In Chapter 4 I tested for the presence of a threatening psychological performance boost using survey items I characterized as measuring "reinforcing perceptions." I argued that if evaluation scores (i.e. LOE) produced by the teacher evaluation system improved teacher performance via a psychological boost, psychologically boosted teachers should hold more positive views of the evaluation system. There was no evidence of such reinforcing perceptions. In this section I test the sensitivity of those findings to my treatment of the survey items.

Original tests for reinforcing perceptions used five survey items, each measuring teacher perceptions about the legitimacy or usefulness of the evaluation system. None of these items were collapsed into single measures, unlike some of the impetus to improve measures. Each version of *PDhrs, tchcollab,* and *effortsum* were based on multiple items administered as a cluster of items on the same survey (see Appendix E). Items measuring reinforcing perceptions were not administered as a cluster nor were all these items administered on the same survey (see Appendix E). In this section I collapse all reinforcing perceptions items into a sum of all binary responses. This new operationalization only uses responses from the 2014 TES because that was the only year all reinforcing perception items were included on the same survey. Table 31 shows results from this new operationalization. Again, there is no evidence of reinforcing perceptions.

Table 29: Alternative Operationalizations of Impetus to Improve Survey Outcomes

	$w = 20$	$w = 30$	$w = 40$
Sum: MAX Hrs in PD (*PDhrsMAX*)	0.85 [0.494]	1.00 [0.406]	1.20 [0.310]
Mean: Hrs in PD (*PDhrsmn*)	1.43 [0.223]	1.31 [0.265]	1.28 [0.274]
Mean: MAX Hrs in PD (*PDhrsMAXmn*)	0.67 [0.609]	0.76 [0.554]	0.93 [0.445]
Sum: MAX Svy Tch Collab (*tchcollabMAX*)	5.63*** [< 0.001]	1.11 [0.348]	1.74 [0.138]
Mean: Svy Tch Collab (*tchcollabmn*)	5.58*** [< 0.001]	1.02 [0.395]	1.79 [0.129]
Mean: MAX Svy Tch Collab (*tchcollabMAXmn*)	5.44*** [< 0.001]	1.01 [0.400]	1.68 [0.151]
Sum: HI Svy Exerted More Effort (*effortsumHI*)	3.57** [0.007]	1.87 [0.112]	0.59 [0.671]
Mean: Svy Exerted More Effort (*effortmn*)	4.13** [0.003]	1.69 [0.150]	0.60 [0.661]
Mean: HI Svy Exerted More Effort (*effortHImn*)	4.13** [0.003]	1.69 [0.150]	0.60 [0.661]
Sum: MAX Svy Hrs Improved Instruction (*insthrsMAX*)	0.14 [0.967]	0.34 [0.852]	0.41 [0.803]
Sum: MAX Svy Hrs Prepped for Obs (*obshrsMAX*)	0.68 [0.606]	2.54* [0.038]	2.09 [0.079]

Note: *p*-values in brackets, number of teacher-year records in sample in parentheses. All models include teacher demographics, certification status, controls for the distribution of teacher effectiveness at the school level, a second order polynomial of LOE interacted with teacher certification status, and year fixed effects. Samples sizes are the same as corresponding samples in Table 8. * ($p < 0.05$), ** ($p < 0.01$), *** ($p < 0.001$)

Table 30: Alternative Operationalizations of Impetus to Improve Outcomes: 425-Threshold

	$w = 20$	$w = 30$	$w = 40$
Sum: MAX Hrs in PD (*PDhrsMAX*)	0.37 [0.693]	1.11 [0.329]	0.56 [0.572]
Mean: Hrs in PD (*PDhrsmn*)	0.15 [0.860]	0.95 [0.385]	0.45 [0.638]
Mean: MAX Hrs in PD (*PDhrsMAXmn*)	0.46 [0.629]	0.95 [0.386]	0.31 [0.732]
Sum: MAX Svy Tch Collab (*tchcollabMAX*)	1.58 [0.208]	0.89 [0.412]	0.94 [0.392]
Mean: Svy Tch Collab (*tchcollabmn*)	1.61 [0.201]	0.81 [0.443]	1.27 [0.280]
Mean: MAX Svy Tch Collab (*tchcollabMAXmn*)	1.61 [0.201]	0.85 [0.429]	1.04 [0.352]
Sum: HI Svy Exerted More Effort (*effortsumHI*)	0.92 [0.399]	0.12 [0.886]	< 0.01 [0.999]
Mean: Svy Exerted More Effort (*effortmn*)	0.81 [0.444]	0.05 [0.950]	< 0.01 [0.996]
Mean: HI Svy Exerted More Effort (*effortHImn*)	0.81 [0.444]	0.05 [0.950]	< 0.01 [0.996]
Sum: MAX Svy Hrs Improved Instruction (*insthrsMAX*)	0.06 [0.937]	0.15 [0.857]	0.40 [0.670]
Sum: MAX Svy Hrs Prepped for Obs (*obshrsMAX*)	0.87 [0.420]	0.69 [0.500]	0.15 [0.863]

Note: Ibid.

Table 31: Alternative Operationalization of Reinforcing Perceptions Outcomes

	$w = 20$	$w = 30$	$w = 40$
Sum: Reinforcing Perceptions (*reinforcesum*)	1.18	0.99	0.78
	[0.306]	[0.372]	[0.460]
	(2775)	(4252)	(5547)

Note: p-values in brackets, number of teacher-year records in sample in parentheses. All models include teacher demographics, certification status, controls for the distribution of teacher effectiveness at the school level, and second order polynomial of LOE interacted with teacher certification status.

G. Non-linear Effects

Models referenced in the text assume a linear relationship between observations received and teacher performance. However, this assumption is invalid if the effect of an additional observation depends on the number of observations received. For example, diminishing marginal effects are plausible. After receiving some number of observations per year, subsequent post-observation feedback may overwhelm the teacher, stymieing productive responses. It may also be that feedback provided by observers becomes incoherent over time.

I explore non-linear effects in two ways. In full-sample RDDs I modify the functional form of the instruments and endogenous predictors. This approach will not work in local RDDs since the instruments in these models are binary. To explore non-linearities in local RDDs I allow first- and second-stage estimates to vary with each threshold. There is little reason to believe non-linearities exist in any model.

To estimate non-linearities in the full-sample RDDs I model non-linear forms of the instrument and endogenous predictor: this is the only difference between these new models and equations 9 and 10. In the first set of models estimating non-linear effects I model the instrument and endogenous predictors as second order polynomials[29]. The results of these models are graphed in Figures F1 – F3, which graph marginal effects. These figures suggest there is little, if any, curvature in these new relationships. Moreover, the difference in predicted outcomes from one observation to the next remains almost perfectly constant (see Figures 4 – 6). I also explore non-linearities by taking the natural log of the instrument and endogenous predictor. Figures 7 – 9 graph the marginal effects of this non-linear modeling. Again, there is little reason to think the

[29] Third and fourth order polynomial versions could not estimate standard errors.

relationship between observations and teacher performance is non-linear in full-sample RDD models.

To estimate non-linearities in the relationship between observations and teacher performance using local RDDs I exploit the fact these models use two discontinuities. I create three endogenous predictors by interacting the number of observations received with a new three-category variable. I assign teachers with an LOE1 to one group, those with an LOE2 - LOE4 to a second, and LOE5 teachers to the final group. The policy-assigned number of observations in each category does not vary for teachers holding the same certification status. Replacing the single endogenous predictor in equation 2 with the interacted three endogenous predictors is the only difference between these new models and equations 1 and 2. Results from these new models are listed in Table 32.

There is no evidence the relationship between observations and the outcomes of interest varies by LOE. While point estimates differ based on LOE, these differences are not statistically different. For example, in a bandwidth of 20 the relationship between observations and changes in TLM RLA scores for teachers with a LOE1 and LOE5 are 0.08 and 0.23. But, the corresponding 95% confidence intervals are approximately [-0.567, 0.732] and [-0.917, 1.222], which overlap with one another.

Figure 4: Quadratic IV and Endogenous Predictors

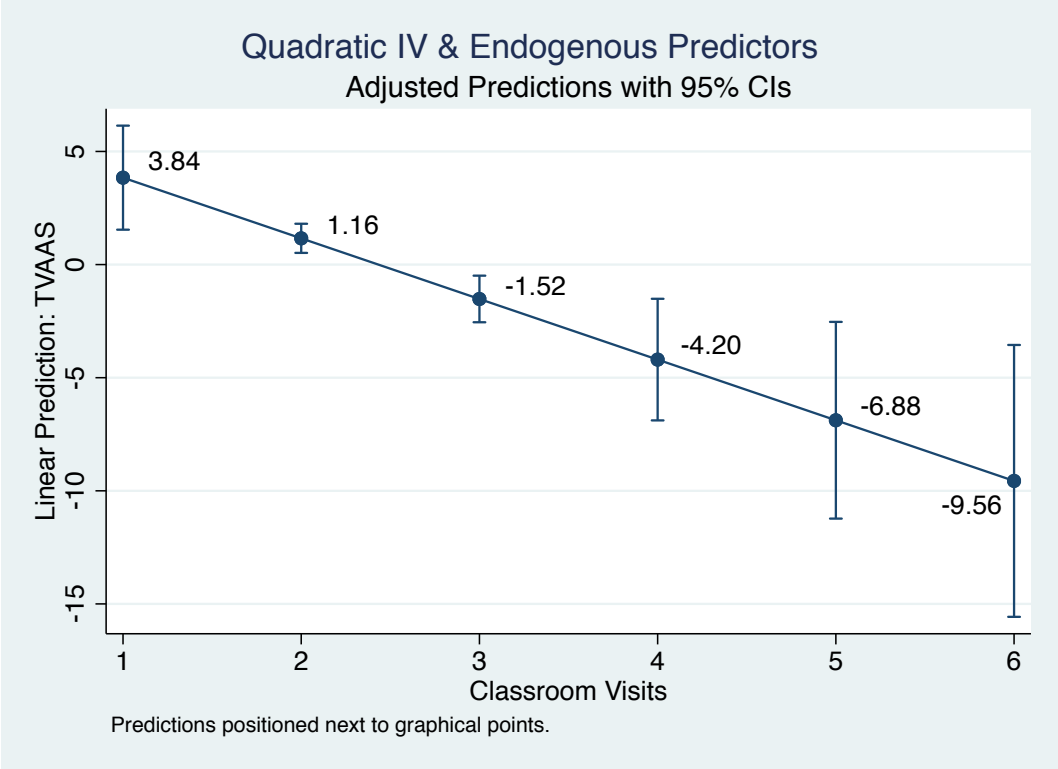


Figure 5: Quadratic IV and Endogenous Predictors

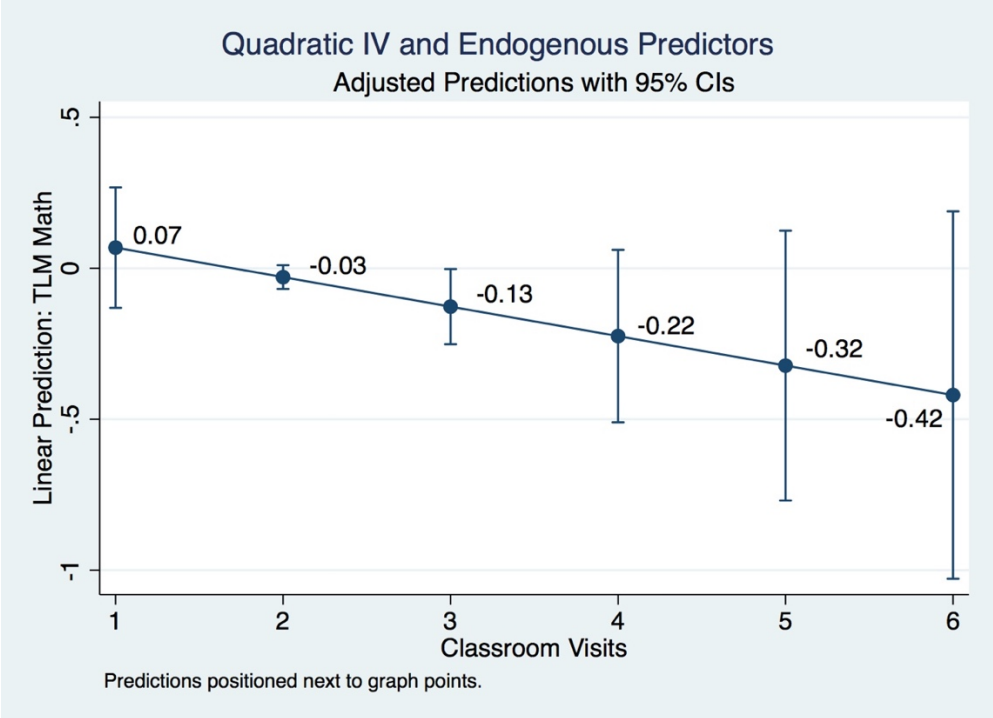


Figure 6: Quadratic IV and Endogenous Predictors

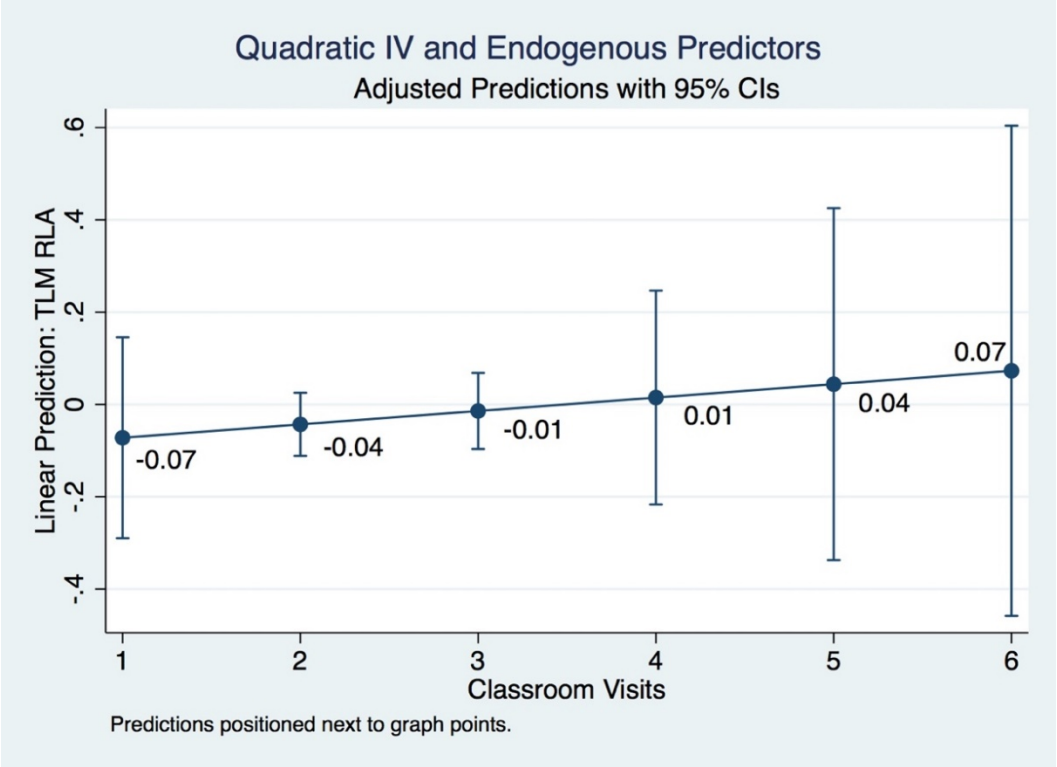


Figure 7: Natural Log of IV and Endogenous Predictors

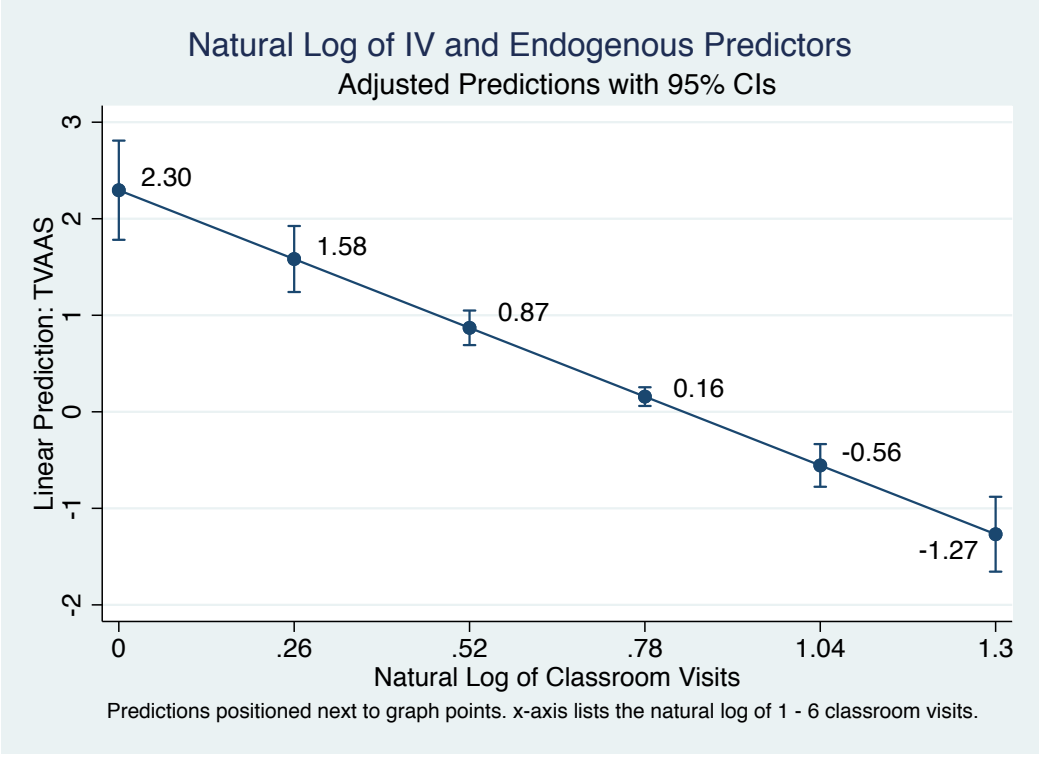


Figure 8: Natural Log of IV and Endogenous Predictors

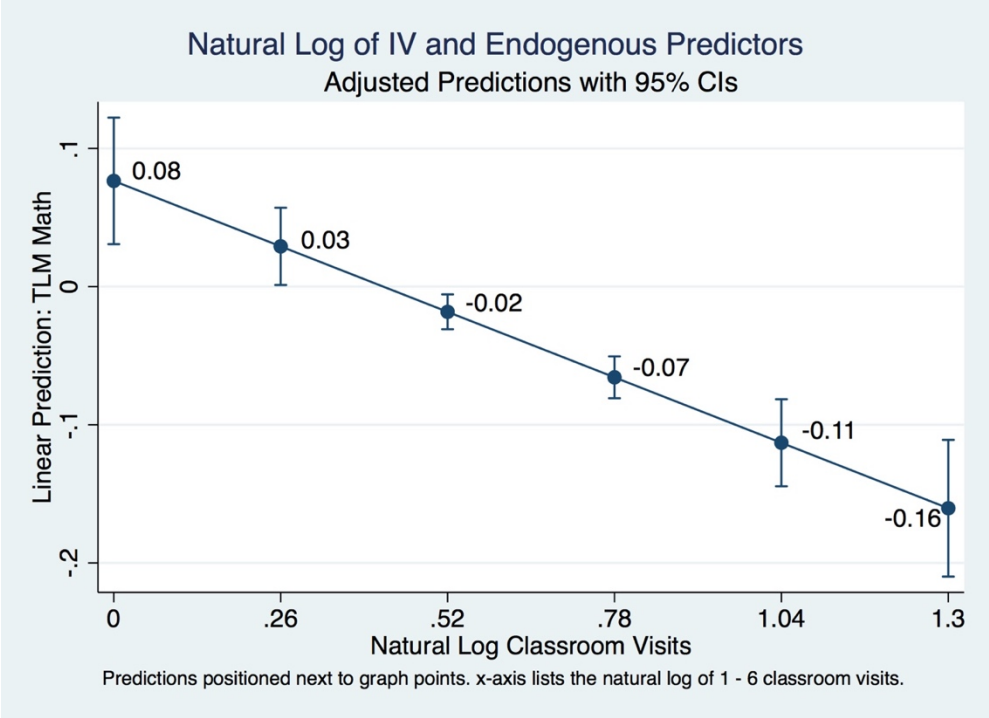


Figure 9: Natural Log of IV and Endogenous Predictors

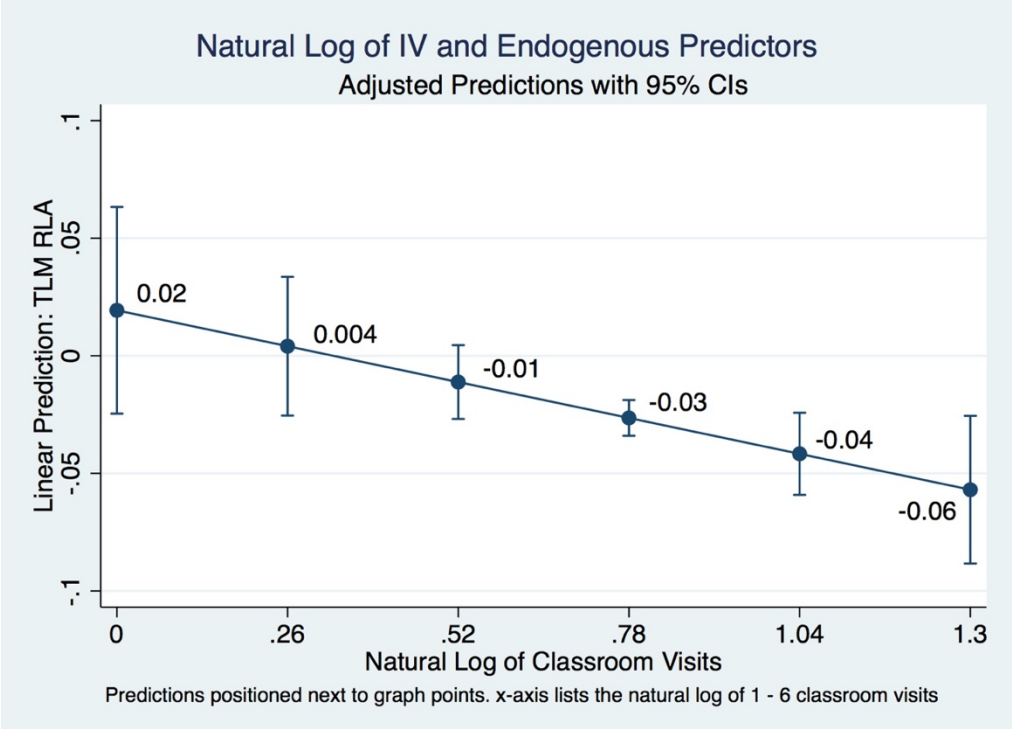


Table 32: Pooled Local RDD Estimated Non-Linear Relationships

	DV=TVAAS			DV=TLM Math			DV=TLM RLA		
2nd Stage Non-Linear Observations	$w = 20$	$w = 30$	$w = 40$	$w = 20$	$w = 30$	$w = 40$	$w = 20$	$w = 30$	$w = 40$
LOE1	-1.21	-1.86	-2.39	0.01	-0.02	-0.09	0.08	-0.06	-0.13
	[2.558]	[1.833]	[1.517]	[0.113]	[0.089]	[0.108]	[0.331]	[0.111]	[0.119]
LOE2 – 4	1.27	-0.03	-0.91	0.04	0.03	-0.11	0.15	-0.08	-0.21
	[4.248]	[3.173]	[2.570]	[0.192]	[0.154]	[0.191]	[0.546]	[0.193]	[0.210]
LOE5	2.14	0.21	-1.25	0.12	0.09	-0.12	0.23	-0.11	-0.29
	[6.040]	[4.567]	[3.777]	[0.268]	[0.217]	[0.277]	[0.760]	[0.280]	[0.302]
N(Tch-Yrs)	10014	152507	20731	3348	5205	7197	3143	4906	6783

Note: Teacher-clustered standard errors in brackets. Estimated relationships are main effects, not interactions. Effects separated by teacher LOE, which are the same LOE used in the discontinuous policy assignment of observations. Models include previously discussed controls.

H. Heterogeneous Effects by Teacher and School Characteristics

There is no evidence an additional observation improves teacher performance. However, heterogeneous effects may exist. I interact potential moderators with the instrumental and endogenous variables in the local RDD described by equations 1 and 2 to test for heterogeneous effects.

Operationalization of Moderators

I explored heterogeneous effects with respect to teacher years of experience by categorizing the continuous years of experience variable. I assigned teachers with [0, 5), [5, 10), and [10, 70) years of experience to three different categories, then replaced the continuous years of experience control with the categorical experience variable. I discuss estimates produced by this moderation analysis elsewhere (see Table 19).

Prior research implied teacher perceptions about the usefulness/ credibility of the observation or evaluation system may moderate the effectiveness of observations. I conducted moderation analyses using survey measures of these teacher perceptions. The Tennessee Educator Survey (TES) asked teachers to report whether they dis/ agreed with the following statements: evaluations will improve teaching (*imprvtch*); evaluations are conducted fairly (*faireval*); my observer is qualified (*obsqual*); and, post-observation feedback is useful (*fbuseful*) (see Appendix E). I used these four measures to create school level moderators. I first dichotomized each of the four survey measures, then calculated the proportion of survey respondents in a school-year cell reporting that they agreed with each statement, yielding four school-year variables. Then, I rank ordered schools within each study year and identified the

quartiles of these distributions. This produced four time-variant, school level moderators, with each variable having four categories. Importantly, the survey items underlying these four school level moderators were measured near the end of the academic year during which teachers received their observations. For this reason, I assume teachers with good observation-related experiences will report higher opinions of the observation and evaluation system. If observations have a positive impact on performance for any of these teacher subgroups, research reviewed in Chapter 2 suggests it would be the subgroup of teachers reporting that those observations were useful/ acceptable.

The last set of moderators concern observer effectiveness. I estimate LATEs moderated by school-year level administrator: skills as a teacher evaluator, skills as a supporter of teacher professional learning, and general effectiveness as measured by the summative administrator TEAM (*admTEAM*) score and LOE (*admLOE*). TDOE administrative data include each of these measures. I constructed the other two observer effectiveness moderators.

Over the study period, the administrator TEAM rubric underwent multiple revisions. However, each version of the rubric included indicators describing administrator behaviors related to their skills as a teacher evaluator and supporter of teacher professional learning. After reviewing administrator rubric used during the study period, I identified indicators describing behaviors related to teacher evaluation and professional learning. Table 33 presents these indicators, the domains they belonged to, and years during which they were on the rubric. I find the mean score across administrator TEAM indicators measuring their skill as a teacher evaluator (indicators QTE1 – QTE5 and PLG1, see Table 33). This becomes the *admTE* moderator (see the *Heterogeneous Effects by Teacher and School Characteristics* section in Chapter 5). I created the *admPL* moderator (see Chapter 5) by taking the mean score of items related to the

administrator's skill as a supporter of teacher professional learning (CI1, ILCI1, PLG2 and PLG3 in Table 33). There is one caveat to the operationalization of the four observer effectiveness moderators: not all observers receive administrator TEAM or LOE scores because not all observers are administrators. School administrators conducted over 80% of observations during the study period. Some observers are district employees or teachers themselves.

Findings

I discuss most moderation analyses in Chapter 5 (see Tables 18 – 21). Findings discussed here pertain to moderators concerning teacher perceptions. Table 34 displays results from moderation by *evalfair* (top panel) and *imprvtch* (bottom), and Table 35 includes results from moderation by *obsqual* (top panel) and *fbuseful* (bottom panel). Similar to results in Tables 18 – 21, almost all point estimates in Tables 34 and 35 are negative or statistically insignificant. However, almost none of the results exhibit hypothesized patterns across subgroups. The only findings providing very weak support for the hypothesized heterogeneous patterns are effects on TVAAS moderated by perceptions about the usefulness of feedback (*fbuseful*). Point estimates increase as the proportion survey respondents in a school report finding feedback useful. Yet, none of these are statistically distinguishable from one another.

Table 33: Selected Administrator TEAM Rubric Indicators

Academic Years	TEAM Domains	TEAM Indicators
	Skill as a Teacher Evaluator	
2012-13, 2013-14	Quality of Teacher Evaluation	QTE1 - Accurately calibrates evidence to the rubric
		QTE2 - Effectively communicates the importance, intent and process of evaluation to educators
		QTE3 - Provides accurate, high quality feedback to teachers and about instructional practices
		QTE4 - Uses data to reflect on evaluation trends
		QTE5 - Performs the process of teacher evaluation with fidelity
2014-15	Professional Learning and Growth	PLG1 - Evaluation
	Skill as a Supporter of Teacher Professional Learning	
2012-13, 2013-14	Continuous Improvement	CI1 - Professional Learning Support
	Instructional Leadership for Continuous Improvement	ILCI1 - Capacity Building
2014-15	Professional Learning and Growth	PLG2 - Differentiated Professional Learning
		PLG3 - Induction, Support, Retention, and Growth

Table 34: Heterogenous Effects by Perceptions About Evaluation System

	TVAAS			TLM Math			TLM RLA		
	$w = 20$	$w = 30$	$w = 40$	$w = 20$	$w = 30$	$w = 40$	$w = 20$	$w = 30$	$w = 40$
	School-Level Proportion of TES Respondents Agreeing Evals Are Fair								
1st Qrt	1.58	0.82	1.10	-0.15	-0.08	-0.09	-0.13	-0.08	-0.08
	[-4.35,7.51]	[-2.54,4.19]	[-2.36,4.56]	[-0.30,0.01]	[-0.20,0.04]	[-0.20,0.02]	[-0.31,0.05]	[-0.21,0.05]	[-0.20,0.04]
2nd Qrt	2.20	0.43	0.25	0.05	0.02	0.01	-0.10	-0.07	-0.09*
	[-2.68,7.07]	[-2.59,3.44]	[-2.88,3.38]	[-0.13,0.24]	[-0.15,0.19]	[-0.14,0.16]	[-0.24,0.04]	[-0.19,0.06]	[-0.19,-0.00]
3rd Qrt	2.68	1.46	1.32	> -0.01	-0.02	0.03	-0.19*	-0.13	-0.11
	[-3.10,8.45]	[-2.24,5.17]	[-2.35,5.00]	[-0.17,0.17]	[-0.19,0.15]	[-0.11,0.17]	[-0.37,-0.01]	[-0.32,0.06]	[-0.26,0.04]
4th Qrt	2.72	0.26	0.88	0.03	-0.02	< 0.01	-0.02	-0.04	-0.03
	[-3.77,9.22]	[-2.94,3.47]	[-2.28,4.04]	[-0.18,0.24]	[-0.17,0.13]	[-0.13,0.13]	[-0.31,0.27]	[-0.21,0.14]	[-0.16,0.09]
	1382	2136	3013	637	998	1373	569	890	1228
	School-Level Proportion of TES Respondents Agreeing Evals Will Improve Teaching								
1st Qrt	-1.53	-0.92	-0.92	-0.08	-0.06	-0.04	-0.03	-0.04	-0.06
	[-3.37,0.31]	[-2.20,0.35]	[-2.22,0.37]	[-0.19,0.03]	[-0.16,0.04]	[-0.13,0.05]	[-0.13,0.07]	[-0.12,0.04]	[-0.13,0.02]
2nd Qrt	-1.28	-1.67*	-1.58*	-0.01	-0.02	0.02	-0.10	-0.12	-0.15*
	[-3.09,0.53]	[-3.07,-0.27]	[-2.87,-0.28]	[-0.14,0.12]	[-0.14,0.09]	[-0.09,0.12]	[-0.25,0.05]	[-0.23,0.00]	[-0.26,-0.03]
3rd Qrt	-0.69	-0.61	-0.33	-0.03	-0.03	< 0.01	0.01	-0.03	-0.06
	[-2.56,1.19]	[-2.11,0.90]	[-1.70,1.04]	[-0.19,0.12]	[-0.16,0.11]	[-0.12,0.12]	[-0.09,0.11]	[-0.10,0.05]	[-0.14,0.02]
4th Qrt	-0.32	-0.33	-0.08	-0.06	-0.03	0.01	-0.05	-0.09	-0.12*
	[-2.91,2.25]	[-1.97,1.31]	[-1.51,1.35]	[-0.20,0.08]	[-0.15,0.09]	[-0.09,0.12]	[-0.16,0.05]	[-0.19,0.00]	[-0.22,-0.02]
	2297	3621	5151	1041	1658	2320	1002	1594	2208

Note: Teacher-clustered standard errors in brackets. Models include previously discussed controls. * ($p < 0.05$), ** ($p < 0.01$), *** ($p < 0.001$)

Table 35: Heterogenous Effects by Perceptions About Observation System

	TVAAS			TLM Math			TLM RLA		
	w = 20	*w* = 30	*w* = 40	*w* = 20	*w* = 30	*w* = 40	*w* = 20	*w* = 30	*w* = 40
	School-Level Proportion of TES Respondents Agreeing Observers Qualified								
1st Qrt	-0.73	-0.81	-0.73	-0.06	-0.01	-0.02	-0.07	-0.04	-0.04
	[-4.82,3.36]	[-3.01,1.40]	[-2.83,1.37]	[-0.16,0.05]	[-0.12,0.09]	[-0.11,0.08]	[-0.28,0.15]	[-0.21,0.12]	[-0.17,0.10]
2nd Qrt	0.24	-0.06	-0.71	0.01	0.03	0.03	0.02	-0.04	-0.03
	[-4.77,5.25]	[-2.80,2.68]	[-3.35,1.93]	[-0.11,0.12]	[-0.10,0.17]	[-0.09,0.15]	[-0.17,0.20]	[-0.22,0.14]	[-0.17,0.10]
3rd Qrt	0.31	0.45	0.17	-0.01	0.07	0.05	-0.08	-0.07	-0.12**
	[-2.79,3.39]	[-1.65,2.55]	[-1.74,2.07]	[-0.19,0.17]	[-0.10,0.23]	[-0.09,0.18]	[-0.20,0.04]	[-0.20,0.05]	[-0.22,-0.03]
4th Qrt	0.45	0.07	0.52	-0.14	-0.08	-0.02	-0.17	-0.17	-0.18
	[-6.46,7.36]	[-3.43,3.57]	[-2.42,3.45]	[-0.45,0.17]	[-0.31,0.14]	[-0.20,0.16]	[-0.42,0.09]	[-0.40,0.07]	[-0.36,0.01]
	1397	2154	3036	643	1010	1387	564	881	1223
	School-Level Proportion of TES Respondents Agreeing Feedback is Useful								
1st Qrt	-1.11	-1.21	-1.01	-0.04	-0.02	-0.02	-0.05	-0.09	-0.10
	[-2.97,0.76]	[-2.75,0.33]	[-2.46,0.43]	[-0.19,0.11]	[-0.15,0.10]	[-0.13,0.09]	[-0.18,0.07]	[-0.20,0.02]	[-0.21,0.02]
2nd Qrt	> -0.01	-0.68	-0.61	0.01	-0.04	-0.01	0.03	-0.02	-0.02
	[-1.90,1.89]	[-2.23,0.88]	[-2.05,0.83]	[-0.12,0.14]	[-0.16,0.08]	[-0.12,0.10]	[-0.10,0.16]	[-0.12,0.08]	[-0.14,0.09]
3rd Qrt	-0.41	-0.47	-0.48	-0.16	-0.14*	-0.10	-0.03	-0.09	-0.07
	[-2.22,1.40]	[-2.07,1.12]	[-1.99,1.03]	[-0.33,0.00]	[-0.28,-0.01]	[-0.22,0.03]	[-0.18,0.12]	[-0.22,0.03]	[-0.19,0.04]
4th Qrt	0.17	-0.19	-0.06	-0.06	-0.04	-0.02	-0.03	-0.09	-0.11*
	[-1.52,1.85]	[-1.56,1.17]	[-1.36,1.24]	[-0.20,0.08]	[-0.16,0.07]	[-0.12,0.09]	[-0.16,0.11]	[-0.21,0.02]	[-0.22,-0.00]
	1585	2533	3664	704	1140	1626	680	1117	1576

Note: *Ibid*

REFERENCES

Aaronson, D., Reserve, F., Barrow, L., Sander, W., Altonji, J., Butcher, K., … Diciccio, T. (2007). Teachers and Student Achievement in the Chicago Public High Schools, *25*(1).

Alexander, K. (2016). *TEAM Evaluator Training*.

Baumeister, R. F. (1998). The Self. In D. Gilbert, S. Fiske, & G. Lindzey (Eds.), *The Handbook of Social Psychology* (1st ed., pp. 680–740). Boston, MA: McGraw Hill.

Brophy, J., & Good, T. L. (1986). Teacher Behavior and Student Achievement. In M. C. Wittrock (Ed.), *Handbook of Research on Teaching* (3rd ed., pp. 328–375). New York, NY: Macmillan.

Cattaneo, M., Jannson, M., & Ma, X. (2016). Simple Local Regression Distribution Estimators with an Application to Manipulation Testing.

Cherasaro, T. L., Brodersen, R. M., Reale, M. L., & Yanoski, D. C. (2016). *Teachers' responses to feedback from evaluators: What feedback characteristics matter? Making Connections*. Washington, D.C.

Cohen, J., & Goldhaber, D. (2016). Building a More Complete Understanding of Teacher Evaluation Using Classroom Observations. *Educational Researcher*, *45*(6), 0013189X16659442. https://doi.org/10.3102/0013189X16659442

Curtis, R., & Wiener, R. (2012). *Means to an End: A Guide to Developing Teacher Evaluation Systems that Support Growth and Development*. Washington, D.C.

Daley, G., & Kim, L. (2010). National Institute for Excellence in Teaching A Teacher Evaluation System That Works. *Working Paper*.

Danielson Group. (n.d.). The Framework for Teaching. Retrieved from http://www.danielsongroup.org/article.aspx?page=frameworkforteaching

Georgia Department of Education. (2012). *Teacher Keys and Leader Keys Effective Systems*.

Guerin, B. (1993). *Social Facilitation* (1st ed.). Cambridge, UK: Cambridge University Press.

Halverson, R., Kelley, C., & Kimball, S. M. (2004). Implementing Teacher Evaluation Systems: How Principals Make Sense of Complex Artifacts to Shape Local Instructional Practice. In W. K. Hoy & C. G. Miskel (Eds.), *Educational Administration, Policy, and Reform: Research and Measurement*. Greenwich, CT: Information Age Publishing.

Harris, D. N., & Sass, T. R. (2011). Teacher training, teacher quality and student achievement. *Journal of Public Economics*, *95*(7–8), 798–812. https://doi.org/10.1016/j.jpubeco.2010.11.009

Hill, H. C., & Grossman, P. (2013). Learning from Teacher Observations: Challenges and Opportunities Posed by New Teacher Evaluation Systems. *Harvard Educational Review*, *83*(2), 371–385. https://doi.org/10.1017/CBO9781107415324.004

Hoy, W. K., & Woolfolk, A. E. (1993). Teachers' sense of efficacy and the organizational health of schools. *Elementary School Journal*, *93*, 356–372.

Ilgen, D. R., Fisher, C. D., & Taylor, M. S. (1979). Consequences of Individual Feedback on Behavior in Organizations. *Journal of Applied Psychology*, *64*(4), 349–371. https://doi.org/10.1037/0021-9010.64.4.349

Jackson, C. K., Rockoff, J. E., & Staiger, D. O. (2014). Teacher Effects and Teacher-Related Policies. *Annual Review of Economics*, *6*(1), 801–825. https://doi.org/10.1146/annurev-economics-080213-040845

Jawahar, I. M. (2010). The Mediating Role of Appraisal Feedback Reactions on the Relationship Between Rater Feedback-Related Behaviors and Ratee Performance. *Group & Organization Management*, *35*(4), 494–526. https://doi.org/10.1177/1059601110378294

Kimball, S. M. (2003). Analysis of Feedback, Enabling Conditions and Fairness Perceptions of Teachers in Three School Districts with New Standards-Based Evaluation Systems. *Journal of Personnel Evaluation in Education*, *16*(4), 241–268. https://doi.org/10.1023/A:1021787806189

Kinicki, A. J., Prussia, G. E., Wu, B. J., & McKee-Ryan, F. M. (2004). A Covariance Structure Analysis of Employees' Response to Performance Feedback. *The Journal of Applied Psychology*, *89*(6), 1057–1069. https://doi.org/10.1037/0021-9010.89.6.1057

Kluger, A. N., & DeNisi, A. (1996). The Effects Of Feedback Interventions On Performance: A Historical Review, A Meta-Analysis, And A Preliminary Feedback Intervention Theory. *Psychological Bulletin*, *119*(2), 254–284. https://doi.org/10.1037/0033-2909.119.2.254

Kraft, M. A., & Gilmour, A. F. (2016a). Can Principals Promote Teacher Development as Evaluators? A Case Study of Principals' Views and Experiences. *Educational Administration Quarterly*, *52*(5), 711–753. https://doi.org/10.1177/0013161X16653445

Kraft, M. A., & Gilmour, A. F. (2016b). Revisiting the Widget Effect: Teacher Evaluation Reforms and the Distribution of Teacher Effectiveness. In *Association of Education Finance and Policy* (pp. 1–31). Washington, D.C.

Levy, P. E., & Williams, J. R. (2004). The Social Context of Performance Appraisal: A Review and Framework for the Future. *Journal of Management*, *30*(6), 881–905. https://doi.org/10.1016/j.jm.2004.06.005

London, M., & Smither, J. W. (2002). Feedback orientation, feedback culture, and the longitudinal performance management process. *Human Resource Management Review*, *12*(1), 81–100. https://doi.org/10.1016/S1053-4822(01)00043-2

Manna, P. (2011). *Collision Course: Federal Education Policy Meets State and Local Realities*. CQ Press.

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating Models for Teacher Accountability* (No. 0833035428) (pp. 1–191). Washington, D.C. Retrieved from http://www.questia.com/PM.qst?a=o&se=gglsc&d=102693148

Mehta, J. (2013). *The Allure of Order*. Oxford University Press.

Morgan, S. L., & Winship, C. (2007). *Counterfactuals and Causal Inference: Methods and Principles for Social Research* (1st ed.). New York, NY: Cambridge University Press.

Murnane, R., & Willett, J. B. (2011). *Methods Matter: Improving Causal Inference in Educational and Social Science*. Oxford, UK: Oxford University Press.

Murphy, K. R., & Cleveland, J. N. (1995). *Understanding Performance Appraisal: Social, Organizational, and Goal-Based Perspectives*. Thousand Oaks, CA: Sage Publications.

Neumerski, C. M., Grissom, J. A., Goldring, E., Cannata, M., Drake, T. A., Rubin, M., & Schuermann, P. (2014). Inside Teacher Evaluation Systems: Shifting the Role of Principal as Instructional Leader. In *Association of Education Finance and Policy* (pp. 1–32). San Antonio, TX.

Papay, J. P., & Kraft, M. A. (2013). 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. https://doi.org/10.1016/j.jpubeco.2015.02.008

Raudenbush, S. W., Rowan, B., & Cheong, Y. F. (1992). Contextual effects on the self-perceived efficacy of high school teachers. *Sociology of Education*, *65*(2).

Rigby, J. G. (2015). Principals' Sensemaking and Enactment of Teacher Evaluation. *Journal of Educational Administration*, *53*(3), 374–392. https://doi.org/10.1108/JEA-04-2014-0051

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).

Sartain, L., Stoelinga, S. R., Brown, E. R., Luppescu, S., Matsko, K. K., Miller, F. K., … Glazer, D. (2011). *Rethinking teacher evaluation: Lessons learned from observations, principal-teacher conferences, and district implementation*. Chicago, IL.

SAS. (2015). *Technical Documentation for 2015 TVAAS Analyses 1.1*.

SAS. (2016). *Technical Documentation of 2016 TVAAS Analyses*.

Stecher, B. M., Garet, M. S., Hamilton, L. S., Steiner, E. D., Robyn, A., Poirier, J., … Brodziak de los Reyes, I. (2016). *Improving Teaching Effectiveness*. Santa Monica, CA.

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). https://doi.org/10.1162/EDFP_a_00186

Steinberg, M. P., & Sartain, L. (2015). Does teacher evaluation improve school performance? Experimental evidence from Chicago's Excellence in Teaching Project. *Education Finance and Policy*, *10*(4), 535–572. https://doi.org/10.1162/EDFP_a_00173

Sun, M., Mutcheson, R. B., & Kim, J. (2016). Teachers' Use of Evaluation for Instructional Improvement and School Supports for Such Use. In J. Grissom & P. Youngs (Eds.), *Improving Teacher Evaluation Systems: Making the Most of Multiple Measures* (1st ed., pp. 102–116). New York, NY: Teachers College Press.

Taylor, E. S., & Tyler, J. H. (2012). The Effect of Evaluation on Teacher Performance. *American Economic Review*, *102*(7), 3628–3651. https://doi.org/10.1257/aer.102.7.3628

Tennessee Board of Education. Teacher and Principal Evaluation Policy (2013). Nashville, TN.

Tennessee Department of Education. (2015). Guide to 2015-16 Level of Overall Effectiveness Data in TNCompass. Nashville: Tennessee Department of Education.

Tennessee Department of Education. (2016). Evaluation | TEAM-TN. Retrieved September 22, 2016, from http://team-tn.org/evaluation/

Tennessee Department of Education. (2018). IPI | TEAM-TN. Retrieved January 25, 2018, from http://team-tn.org/ipi/

Tversky, A., & Kahneman, D. (1991). Loss Aversion in Riskless Choice: A Reference-Dependent Model. *The Quarterly Journal of Economics*, *106*(4), 1039–1061.

US Department of Education. (2009). *Race to the Top Program Executive Summary*.

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The Widget Effect*.