Teacher Evaluation, Instructional Practice and Student Achievement: Evidence from the District of Columbia Public Schools and the Measures of Effective Teaching Project

A Dissertation

Presented to

The Faculty of the Curry School of Education

University of Virginia

In Partial Fulfillment

Of the Requirements for the Degree

Doctor of Philosophy

by

Melinda Adnot

August 2016

August 2016

Education Policy Studies
Curry School of Education
University of Virginia
Charlottesville, Virginia

APPROVAL OF THE DISSERTATION

This dissertation, *Teacher Evaluation, Instructional Practice and Student Achievement: Evidence from the District of Columbia Public Schools and the Measures of Effective Teaching Project*, has been approved by the Graduate Faculty of the Curry School of Education in partial fulfillment of the requirements for the degree of Doctor of Philosophy

____
James H. Wyckoff (Chair)

____
Benjamin L. Castleman

____
Julia J. Cohen

____
Thomas S. Dee (Stanford)

____
Bridget K. Hamre

TABLE OF CONTENTS

Page

ELEMENTS

DEDICATION

I dedicate this dissertation to my many families who have offered me so much support through this process—my biological family, my academic family, and my friends. I especially thank Charles for his love and patience this past year. I could not have done this without you.

ACKNOWLEDGEMENTS

I am immensely grateful to my advisor, Jim Wyckoff, for the excellent advising and endless patience he provided while I wrote this dissertation and over the past five years. For co-leading our DCPS research project and his sage advice, I am also very grateful to Tom Dee. I also thank my dissertation committee, Ben Castleman, Julie Cohen, and Bridget Hamre for their feedback and support during the dissertation process, and Allison Atteberry, Matt Kraft, and Roddy Theobald for helpful comments and suggestions on previous drafts of this work.

I would also like to thank the District of Columbia Public Schools for providing data used in this dissertation, and Scott Thompson, Kim Levengood, Luke Hostetter, Alden Wells and Austin Zentz of DCPS for answering questions about the data and about IMPACT. I also use data from the Measures of Effective Teaching Project, and I appreciate the work done by the original MET researchers and staff at ICPSR as well.

Funding for this work was provided by grants R305H140002 and R305B090002 from the Institute of Education Sciences to the University of Virginia and by the American Education Research Association Measures of Effective Teaching Dissertation Fellowship. The views expressed in this paper are mine, and may not reflect the views of DCPS, IES, or AERA. All errors or omissions are my own.

LIST OF TABLES

LIST OF FIGURES

DISSERTATION OVERVIEW

In the United States, we expect our schools to help students achieve a multitude of goals, from the development of academic, social and emotional skills to preparation for college and workforce participation. Recent research has made clear the important role that teachers play in the pursuit of these goals. Teachers have large effects on students' test outcomes (Aaronson, Barrow, & Sanders, 2007; Chetty, Friedman, & Rockoff, 2014; Kane & Staiger, 2008; Rivkin, Hanushek, & Kain, 2005; Rockoff, 2004) and other, non-test goals such as GPA, on-time grade progression, and motivation and mindset regarding academic work (Blazar, 2015; Jackson, 2012; Jennings & DiPrete, 2010; Kraft & Grace, 2016; Ruzek, Domina, Conley, Duncan, & Karabenick, 2014). Moreover, these effects persist much later into students' lives, and impact future measures of well-being such as college attendance and future income as well (Chetty, Friedman, & Rockoff, 2014; Jackson, 2012).

While it is clear that there are meaningful differences between teachers in their ability to help students realize educational goals, the factors that cause large differences in individual teacher performance are not well understood. For instance, we know that teachers' performance improves on average over time (Atteberry, Loeb, & Wyckoff, 2013; Papay & Kraft, 2015; Rockoff, 2004), but gaining more experience does not invariably lead all teachers to become high-performing. Most other observable credentials also have limited power to predict future job performance (Goldhaber, 2015; Wayne & Youngs, 2003).

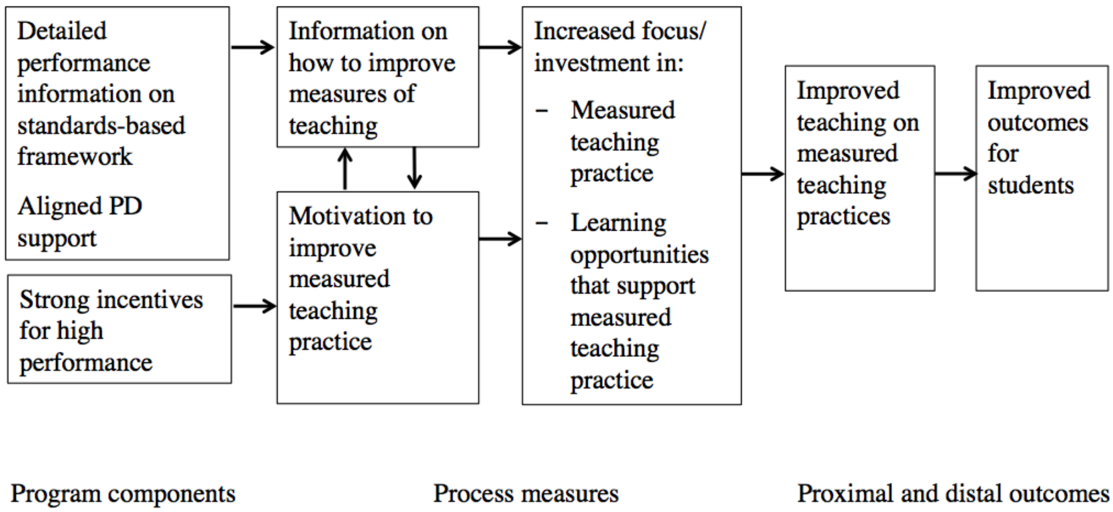
If we can’t identify good teachers at the time of hire and the development of expert practice doesn’t happen automatically for all teachers, then there are two options for improving the performance of the teacher workforce: selectively retain teachers once their performance has been demonstrated, or put practices in place that help people improve. Both of these options rely on performance information that (1) differentiates between teachers in their performance level, and (2) provides them with formative guidance on improving their practice.

In the past, evaluations of teachers’ job performance failed to offer meaningful information, resulting in ratings that labeled every teacher good or great, and provided teachers little formative feedback about their teaching (Weisberg, Sexton, Mulhern, & Keeting, 2009). However, we’ve seen rapid policy reform in the area of teacher performance evaluation in the past eight years (Steinberg & Donaldson, 2016), much of it encouraged by the policy priorities of the Obama administration through programs such as Race to the Top (RTTT) and the Teacher Incentive Fund (TIF) competitive grant programs. While one goal of these systems is to provide an overall measure of teacher performance that can be used for accountability purposes, another, perhaps more important goal, is to use the individualized feedback on specific aspects of instruction to help teachers improve (Hill & Grossman, 2013; Papay 2012). The introduction of standardized observation frameworks and their use for performance evaluation has been an important innovation in this pursuit. These tools provide detailed descriptions of the teacher practices that are hypothesized to help realize educational goals for students, and provide a common understanding of the elements of effective teaching (Pianta & Hamre, 2009).

There is emerging evidence that evaluation that features observation can improve overall teacher performance and student achievement (Dee & Wyckoff, 2015; Steinberg & Sartain, 2015; Taylor & Tyler, 2012). However, we still know too little about how best to use the elements of the evaluation process and the information arising from evaluation to encourage improvement in teaching practice and outcomes for students. In addition, the recent passage of the Every Student Succeeds Act (ESSA) is likely to increase state-level differentiation of performance evaluation policies as states revisit the systems adopted over the last eight years. This coming policy shift makes clear the need both for impact studies that investigate the overall effects of evaluation programs, and for research which explores how evaluation programs are functioning at a more granular level. The goal of this dissertation is to provide information on these important questions.

Figure 1 presents a conceptual model for how improvement in teaching may occur in the context of high-stakes, observation-focused performance evaluation. This model is not intended to illustrate every force that could be at play in every evaluation system, but rather to give a broad overview of the theorized path demonstrating how program components may lead to improvements in teaching and student achievement.

Figure 1. Improving teaching in the context of high-stakes, observation-focused performance evaluation


Detailed performance information on standards-based framework
Aligned PD support
Strong incentives for high performance
Information on how to improve measures of teaching
Motivation to improve measured teaching practice
Increased focus/ investment in:
– Measured teaching practice
– Learning opportunities that support measured teaching practice
Improved teaching on measured teaching practices
Improved outcomes for students
Program components
Process measures
Proximal and distal outcomes

Many performance evaluation systems, such as the IMPACT evaluation system in the District of Columbia Public Schools (DCPS) that is the focus of my dissertation, combine the elements of detailed performance information on standards-based observation frameworks, aligned professional support, and strong incentives. These program components theoretically provide teachers with information on their teaching performance, and motivation to improve on measured aspects of teaching. The formatively useful information and motivational aspects of evaluation systems are distinct but influence one another, as indicated by the arrows in the diagram above. For instance, better information on how to improve measured teaching practice is likely motivating to teachers, as it leads them to have a stronger expectation that effort they put toward improving will have successful results. This leads teachers to increase the investment they make toward improving measures of teaching, both by investing in the tasks themselves and in development opportunities which may help them improve. In turn, this theoretically leads to improvements in measured teaching practice and improved student outcomes.

The three chapters of my dissertation examine various elements and paths of the conceptual model presented above. Chapters 1 and 2 both use a natural experiment created by the design of IMPACT to examine the effects of strong incentives on specific measures of teaching practice and student achievement. While the causal identification of these chapters focuses on the incentives, these incentives are part of a system where performance information and professional supports are critical components that also influence the ultimate outcomes.

Chapter 1 examines the effects of IMPACT incentives on specific aspects of classroom practice using a regression discontinuity (RD) design that is created by the stark incentive contrasts at two performance thresholds: one that is associated with the potential for dismissal if performance does not improve, and the other that implies a potential base salary increase. In this chapter, I develop a conceptual model based on the expectancy-value framework (Atkinson, 1957; Eccles et al., 1983) that examines how teachers may be differentially motivated to invest in improving specific aspects of their instructional practice based on their expectation of success in each area. I hypothesize that expectancy—and thus observed improvement—will be critically influenced by the specificity of the performance measures themselves, and the difficulty of the instructional tasks they describe. I find that low-performing teachers facing a dismissal threat do, in fact, improve most consistently on instructional standards that detail more specific strategies for success, following the second and third years of the program. High-performing teachers who are eligible for a permanent base salary increase also improve select aspects of teaching following the first year of the program, though these results are less robust to a range of alternative specifications. The finding that low-performing teachers improve where they are offered specific descriptions of effective teaching raises the question of whether these improvements in measured practice lead to authentic and ongoing changes in teaching that have benefits for students. These standards may not be the highest leverage areas for improving student outcomes, but this study suggests that it is where low-performing teachers will focus given the design of the current system.

In the Chapter 2, which is joint with Tom Dee, Veronica Katz, and Jim Wyckoff, we provide important evidence on the question posed above: are the incentive-induced improvements we observe in teaching associated with improved student achievement in DCPS? We employ a more economically-oriented conceptual framing and an RD design to examine the effects of the same IMPACT incentives for low-performing teachers on student achievement outcomes. Consistent with Chapter 1 and earlier work by Dee and Wyckoff (2015), we find no student achievement effects for low-performing teachers after the first year of the program, but we do observe positive effects of roughly seven percent of a standard deviation of student achievement in each of the next two years. These average effects are driven by improvements in math, especially following the third year of IMPACT. Results in English Language Arts are mostly null, though some specifications suggest an effect following the second year. The magnitude of the average effects we observe is educationally significant: it equates to between 12 and 20 percent of a year of learning, depending on the grade, or approximately seven percent of the black-white achievement gap.[1] Moreover, an analysis of treatment effect heterogeneity suggests

[1] Empirical benchmarks for student achievement effect sizes drawn from Hill, Bloom, Black & Lipsey, 2007.

that traditionally underserved student subgroups (i.e., students receiving special education and limited English proficiency services, and students receiving free and reduced price lunch) benefit as much or more from the presence of the incentive compared with their general education and more advantaged peers.

Finally, in the third chapter of my dissertation, I explore whether information on teaching practice captured by classroom observation can be used to identify "profiles of instruction" or patterns in teachers' instructional practice. If there are groups of teachers who share similar characteristics in their instructional practice (e.g., weak ability to provide strong content explanations, but strong classroom management), it may be helpful to explicitly identify these profiles in order to coordinate professional support. In contrast to the first chapter, which considers teachers' responses on each aspect of teaching practice a distinct outcome, this chapter explores the extent to which these practices are interrelated within a teachers' overall professional practice. I employ latent profile analysis to identify profiles of instruction using data from two contexts: DCPS IMPACT and the Measures of Effective Teaching (MET) research project. While the profiles of instruction in MET provide some information on teachers' relative strengths and weaknesses, the defining characteristic of the profiles in DCPS is simply being more or less effective across all aspects of practice. This difference between the findings in MET and DCPS is driven by lower dimensionality in the information captured through classroom observation in DCPS. This limited dimensionality has implications not only for the construction of instructional profiles, but for the goal of providing useful formative feedback to teachers more broadly.

These chapters provide novel evidence in an area of education research where policy and practice are evolving rapidly and the evidence base is thin. Chapter 1 is the first paper to provide rigorous and detailed evidence on teachers' responses to strong incentives embedded in an information-rich, high-stakes evaluation. Chapter 2 explores whether these incentives also causally effect student achievement. Finally, Chapter 3 uses a novel analytic technique to identify different typologies of teaching using multiple sources of classroom observation data. These chapters were written with the objective of providing policymakers with information on both *whether* evaluation impacts teacher and student performance, and *how* this process may occur. As such, they are well-positioned to inform future evaluation policy design decisions, which is likely be an active policy area as states revisit their evaluation systems under ESSA in the coming years. It is my hope that this research helps policymakers improve these systems over time in ways that support teaching and learning in our schools.

CHAPTER 1

Effects of Evaluation and Incentives on Instructional Practice: Evidence from the District of Columbia Public Schools' IMPACT Teacher Evaluation System

Abstract – In recent years, many states and districts have introduced new teacher evaluation policies aimed at improving the overall performance of the teacher workforce. However, while we have extensive evidence that teachers meaningfully impact students' learning, we still know little about how teachers respond to performance evaluation and incentives by changing their instructional practice. Do some aspects of instruction improve more than others, and why might these patterns of improvement occur? Using a unique four-year panel of data from the District of Columbia Public Schools' IMPACT teacher evaluation system, I find that teachers do improve their classroom practice in response to IMPACT's strongest incentives, conditional on the decision to remain in the district, following the second and third years of the program. While I detect broadly positive effects of the incentives on instruction, results also suggest that teachers facing the strongest incentives may focus on specific areas of instruction where they have the greatest expectation of improvement. In particular, I find that low-performing teachers facing a dismissal threat experienced the most robust and consistent improvements on instructional standards that detailed specific strategies for success following the second and third years of the program. This work underscores the importance of design considerations in evaluation systems that employ multiple measures of teacher performance, and begins to build our understanding of how teachers respond to these systems.

1. INTRODUCTION

In recent years, a majority of states and districts have made dramatic changes to teacher evaluation policy, with the goal of driving improvements to the overall performance of the teacher workforce. By the start of the 2014-15 school year, 78 percent of states and 85 percent of the 25 largest districts had revised and implemented new evaluation systems that incorporate multiple measures of teachers' job performance (Steinberg & Donaldson, 2015). A smaller number of states also require that the information from performance evaluation be used to inform personnel decisions such as retention, compensation, and professional development (Doherty & Jacobs, 2015).[2] This reform has been strongly encouraged by the policy priorities of the Obama administration through programs such as Race to the Top (RTTT) and Teacher Incentive Fund (TIF) competitive grant programs, and has been buoyed by philanthropic interest and investment (e.g., the Gates Foundation's Measure of Effective Teaching project). It has also been highly controversial: more than a dozen lawsuits have been filed contesting the legality of high-stakes evaluation (Sawchuk, 2015) and the proverbial jury in the court of public opinion remains out as well.

Much of this newly-directed attention to teacher performance is motivated by an increasingly large body of research demonstrating that teachers are a critically important determinant of students' educational and economic outcomes, and that there are large differences between teachers in their ability to help students succeed (Aaronson, Barrow, & Sander, 2007; Chetty, Friedman, & Rockoff, 2014; Jackson, 2012; Kane & Staiger,

[2] As of November 2015, the National Council on Teacher Quality reports that 23 states require that performance be considered in tenure decisions and 28 articulate that poor performance evaluations can result in dismissal. Only seven states require evaluation results be considered in the determination of base pay (NCTQ, 2015).

2012; Rivkin, Hanushek, & Kain, 2005; Rockoff, 2004). The importance of, and heterogeneity in, teacher performance underlies the theory of action in evaluation reform: if districts are able to measure performance accurately, they can use this information to differentially retain more effective teachers and to help the majority of teachers improve.

While we have extensive evidence that teachers meaningfully impact student outcomes, we know much less about how they respond to performance evaluation and incentives by changing their instructional practice. This is not surprising given the novelty of teacher workforce policies that incorporate these elements, but given the rapidity with which these changes are rolling out, better understanding whether and how evaluation improves teacher performance has vast implications for education policy. Earlier work on IMPACT, the teacher evaluation system in the District of Columbia Public Schools (DCPS) on which this study is also based, shows that teachers do respond to incentives, both through their retention decisions and by improving their overall effectiveness conditional on their decision to remain in teaching (Dee & Wyckoff, 2015). There is also emerging evidence that evaluation systems which feature classroom observations and feedback on performance have positive effects on student achievement as well (Steinberg & Sartain, 2015; Taylor & Tyler, 2012).

These studies provide important first-order evidence on the effects of evaluation on overall teacher performance and student achievement, but shed no light on how teachers alter their teaching in response to evaluation. For instance, do teachers improve all elements of their instructional practice equally when they face strong incentives and receive frequent instructional feedback? Or do they selectively focus on some aspects of teaching more than others? What are the characteristics of those areas that experience the greatest growth, and the least? The answers to these questions have important implications for districts' hiring, evaluation, and professional development (PD) efforts: if teachers appear unable to change certain elements of their practice even in the presence of frequent feedback and strong incentives, this suggests that districts should prioritize these competencies during hiring, or explore new strategies to help teachers develop these skills.

Using a unique four-year panel of data from DCPS's IMPACT teacher evaluation system, I examine the extent and type of improvements that teachers make to instructional practices when they experience strong incentives as part of an observation-focused evaluation system. Under IMPACT, teachers are evaluated using multiple measures of performance including observations of classroom practice, contribution to student test outcomes, and assessments of professionalism and contribution to the school community. The classroom observation component of the evaluation process occupies a central role: for teachers in non-tested grades and subjects (i.e., 79 percent of teachers in DCPS) it comprises 75 percent of the overall evaluation score.[3] It is also intended to offer teachers formative information on how to improve their instructional practice: teachers are observed up to five times a year and meet afterwards with evaluators to discuss results and receive individualized feedback. The overall evaluation score is then associated with large rewards and consequences. Teachers are eligible for large financial bonuses and permanent salary increases for "Highly Effective" (HE) performance, and can also be dismissed for "Ineffective" or twice-"Minimally Effective" (ME) performance.

[3] For teachers of subjects and grades with an end-of-year standardized assessment, classroom observation is weighted at 35 percent.

Overall, instructional practice in DCPS has improved since the beginning of IMPACT, shifting the performance distribution of overall classroom observation scores to the right in each successive year (Figure 1.1). This is encouraging, but the rising scores could be influenced by a number of factors, including compositional changes to the workforce and score inflation. To better understand incentive-induced growth in the district, I leveraging a natural experiment created by the design of IMPACT and use a Regression Discontinuity (RD) design to provide an internally-valid estimate of the effects of strong incentives on specific components of teaching practice. I study the effects of IMPACT incentives at two performance thresholds: the ME threshold, which implies a threat of dismissal for low-performing teachers if their performance does not improve to Effective in the next year, and the HE threshold, at which high-performing teachers are eligible for a permanent base salary increase if they attain a second HE rating.

To preview my results, I find that low-performing teachers do improve their classroom practice in response to the dismissal threat, conditional on their decision to remain in the district, following the second and third years of the program.[4] While I detect broadly positive effects of the incentives on instruction in these years, results also suggest that these teachers may focus on specific areas of instruction where they have the greatest expectation of improvement. In particular, improvements are most consistent and robust for instructional standards that detail more specific strategies for success in the years where we observe positive incentive effects. High-performing teachers who are eligible for a permanent base salary increase also improve select aspects of teaching

[4] Consistent with earlier work by Dee & Wyckoff (2015) and Chapter 2 of this dissertation, I observe null effects following the first year of the program. As I'll discuss later, political instability in District politics following the first year of the program may have damaged IMPACT's credibility and contributed to these null effects.

following the first year of the program, though these results are less robust to a range of alternative specifications. Interestingly, the improvements I detect for high-performing teachers are concentrated in areas of practice where the rubric is much less prescriptive in nature, suggesting that there may be differences in how low- and high-performers respond to incentives by pursuing improvement in various aspects of their professional practice.

This is the first paper to rigorously examine teachers' behavioral responses to an at-scale, long-standing evaluation system. It makes a unique contribution to the literature on teacher evaluation by opening the "black box" of teacher improvement in the context of high-stakes evaluation to understand the overall positive effects we observe in other studies with unprecedented granularity. This work has important implications for design considerations in evaluation systems that employ multiple measures of teacher performance, and begins to build our understanding of how teachers improve their practice on the job in the context of consequential performance evaluation.

2. CONCEPTUAL MODEL: INCENTIVES, INFORMATION AND IMPROVEMENT

2.1 Information and Accountability Goals in Teacher Evaluation. There are two hypothesized paths through which teacher evaluation has the potential to improve teaching. The first, or the "information hypothesis," holds that evaluation is expected to provide teachers with formatively useful information on their practice that can help them become better practitioners. In this line of thinking, if teachers are motivated to develop their practice but lack information on how to do so, the feedback provided through the classroom observation process could provide guidance. Proponents of evaluation-based development efforts suggest that the targeted feedback from standards-based classroom observation may offer a more promising route to improvement than traditional seat time PD models (e.g., Allen et al., 2011; Cohen, Schuldt, Brown & Grossman, 2015; MET 2010; Pianta, 2012; Taylor & Tyler, 2012). For instance, an evaluation of MyTeachingPartner, which offers teachers video-based coaching and feedback using the Classroom Assessment Scoring System (CLASS) found that teachers who participated in the program improved student achievement 22 percent of a standard deviation (SD) in the next year (Allen et al., 2011).

The second path through which evaluation programs may improve teaching, or the "accountability hypothesis," holds that if performance is measured accurately and high performers are retained at higher rates than low performers, then overall teacher quality in the district will improve. Additionally, the presence of rewards and consequences (e.g., financial bonuses or termination) may also motivate teachers to increase or redirect effort toward evaluation-relevant tasks. The field of personnel economics has a well-established theoretical basis for the use of incentives to improve performance (see Lazear & Oyer, 2009 for a review), and empirical studies of piece-rate contracts bear this out (Bandiera, Barankay, & Rasul, 2007; Lazear, 2000, Shearer, 2004). However, incentives in professions that require the use of subjective performance measures and incorporate multiple goals have been shown to be relatively more susceptible to distortions and unintended consequences (Baker, 1992; Baker, Gibbons, & Murphy, 1994; Holmstrom & Milgrom, 1991). For instance, Holstrom and Milgrom develop a "multi-tasking" model indicating that, while incentives may lead employees to work harder, they may also cause them to under-invest in tasks that are less easily measured in the short-term (1991).

In education, the most germane application of incentive theory has been performance pay programs that offer teachers financial bonuses for improving student test outcomes. However, a number of recent experimental evaluations of performance pay programs have found them largely ineffective (Glazerman & Seifullah, 2012; Fryer et al., 2012; Springer, et al., 2010; Springer, et al., 2012).[5] This has led policymakers and researchers to revisit the theory of action by which these programs operate. Performance incentives are thought to work best in contexts where the measured outcomes can be reliably increased with increases in effort, or where the expected marginal benefit of effort is high (Lazear & Oyer, 2009; Milgrom & Roberts, 1992). Newly introduced evaluation and compensation systems such as IMPACT recognize the limitations of test-based incentive systems that only reward distal outcomes. Theoretically, these new systems may be more likely to motivate teachers than a test-based incentive system, since the proximal measures of classroom practice are apparent to teachers and they better understand how their efforts will lead to improvements in the measured outcome.

Healthy skepticism exists regarding the extent to which accountability and formative feedback goals can be met simultaneously within a single system (Donaldson & Papay, 2015; Hill and Grossman, 2013; Papay 2012) and there has been considerable debate over which avenue has the greater potential to drive overall improvements to teaching and student learning. Some scholars have also suggested that the accountability and development goals within evaluation systems may complement one another and encourage improvement more effectively than a system set up to address either policy goal alone (Goldhaber, 2015; Rice, 2009).

[5] One exception is a "loss-aversion" style incentive where teachers were awarded a bonus at the beginning of the year and had to return the money if test scores did not improve (Fryer et al., 2012).

2.2 Evaluation and Instructional Improvement. How do teachers respond to evaluation systems like IMPACT that employ both accountability and performance feedback as mechanisms to improve teacher quality? A simple behavioral model suggests that, given limited time and resources, teachers invest in improving practices and skills that have the greatest expected marginal benefit. This is the logic behind expectancy models of motivation (Atkinson, 1957; Barron & Hulleman, 2014; Eccles et al., 1983; Heneman, 1998; King-Rice et al., 2015; Vroom, 1964; Wigfield & Eccles, 1992).[6] Put simply, expectancy models suggest that individuals are motivated to spend time and effort on the pursuit of activities on which they expect to do well and value a successful outcome. [7] There are many factors that influence a teachers' expectancy and value of success across different areas of instructional practice, even within the context of performance evaluation. Some factors may be relatively idiosyncratic such as individual predilection towards specific instructional tasks, while others may be influenced by systemic causes such as the quality of support that is available to teachers in particular areas. In this study, I focus in on the policy-relevant question of how expectancy and value may be impacted by the attributes of the classroom observation itself.[8]

2.2.2 Expectancy. Expectancy is an individual's expectation of success in undertaking a given task, or in this situation, a teacher's answer to the question: "How readily can I improve this measured component of practice?" Two factors that are

[6] The expectancy-value model of motivation is most often used to study student motivation in educational settings, but is related to Valence-Instrumentality-Expectancy frameworks that have been employed to study response to performance pay policies (King-Rice, Malen, Jackson, & Hoyer 2015), and it's simplified exposition is helpful here.

[7] An alternative, more economically-oriented conceptual framing is that teachers seek to optimize their response to the evaluation system by maximizing rewards (e.g., gaining higher scores) while minimizing effort costs. I focus on expectancy-value theory here as it is helpful to draw on the psychological literature to understand teachers' motivational responses. However, both conceptualizations lead to similar hypotheses about behavior in this context.

particularly salient with regard to performance evaluation are: (1) the specificity with which the instructional standard is articulated, and (2) the inherent difficulty of the instructional task itself.

It is relatively well-established that the provision of specific performance feedback that offers individuals direction on how to improve increases the likelihood of behavior change (Cannon & Witherspoon, 2005; Locke & Latham, 2002; Kluger & DeNisi, 1996). For instance, Locke and Latham write in their theory of goal-setting that, for complex tasks (e.g., the tasks outlined in a classroom observation rubric), "goal effects are dependent on the ability to discover the appropriate task strategies" (2002). Thus, provision of appropriate task strategies increases expectancy of success for a particular instructional standard (or goal) by decreasing the ambiguity. Classroom observation rubrics vary in terms of the scope and level of detail that is used in describing instructional practice, and researchers have cautioned that observation instruments with relatively broad descriptions of effective practice may be less formatively useful (Hill & Grossman, 2013; West et al., 2015). Further, "decomposition" of teaching practices into more specific actions that teachers can attend to is theorized to improve the quality of feedback and increase expectation of improvement (Grossman et al., 2009). To the extent that standards within an observation instrument differ in terms of how explicitly they describe effective practice, teachers may be inclined to focus on those standards which more clearly outline strategies for success.

It is also intuitive that some aspects of teaching are more difficult than others to master, and that expectancy of success would be higher for less difficult practices. Atkinson (1957) placed objective task difficulty, which he defined as the proportion of people to succeed in a given task, at the center of his classic introduction to expectancy-value theory. Other motivation theorists have appropriately pointed out that it is not *objective* task difficulty but *perceived* task difficulty which influences the expectation of success (e.g., Eccles & Wigfield, 1995). However, prior work examining differences in teaching practice find that the highest and lowest scoring areas of instruction are remarkably consistent across settings, with items relating to questioning technique often receiving the lowest scores, and items related to the classroom environment often receiving the highest (Garrett & Steinberg, 2015; Hamre et al., 2013; Kane, Taylor, Tyler & Wooten, 2010; Kane & Staiger, 2012; Sartain, Stoelinga & Brown, 2011). As such, it seems a reasonable expectation that the relative absolute difficulty of standards may be something that teachers are aware of and that this may also influence where they are most likely to invest. Alternatively, it is also possible that the relevant construct here is not absolute task difficulty, but the difficulty of improving particular aspects of teaching. It is unclear a priori whether or not these are the same; for instance, teachers could perceive that they have little room to improve standards on which they already excel (i.e., they face "ceiling effects"). On the other hand, while there is considerable room for improvement on the standards where teachers score the lowest, this may mean that these areas are simply challenging practices to master, and expectancy of success may remain low for these despite ample room to grow on the scoring scale.

2.2.1 Value. The value component of expectancy-value theory relates to the value an individual places on successfully completing a task, or in the context of high-stakes evaluation, a teacher's answer to the question: "What is the benefit of improving this measured aspect of teaching?" Expectancy-value theorists identify utility value (i.e., usefulness), interest value (i.e., inherent enjoyment), and attainment value (i.e., importance to identity) as dimensions of the value construct (Wigfield & Eccles, 2000). Strong incentives embedded in evaluation systems (e.g., potential for dismissal or for permanent salary increase) are expected to increase the utility value of improving measured aspects of instruction under the theory of change implied by high-stakes accountability programs. That is, teachers value improving their evaluation ratings because it is necessary to achieve other goals like continued employment and higher pay that are not related to the task itself. The information supplied by performance ratings also likely affects individuals' need for professional competence (Eraut, 1994; Ryan et al., 1983). Teachers with a low prior rating may experience particularly high attainment value of improvement. Thus, evaluation may increase both the utility and attainment value of improving measures of instructional practice.

Conversely, scholars in the field of self-determination theory maintain that the presence of extrinsic rewards has the potential to dampen intrinsic interest in a task. This line of thinking holds that, while teaching itself is inherently an interesting task, utilizing external motivators may decrease internal drive and negatively impact performance. Lab-based experimental evidence supports this view (see Deci, Koestner, & Ryan, 1999 for a review), but the results of field-based examinations are more mixed (Fang & Gerhart, 2005; Hulleman & Barron, 2010; Rynes, Gerhart, & Parks, 2005). More work is needed to draw firm conclusions regarding the effect of evaluation and incentives on intrinsic motivation, but it represents a potential competing force that could diminish positive utility and attainment effects.

2.3 Summary and Research Questions. Expectancy-value theory suggests that the incentives embedded in IMPACT will motivate teachers to improve measured aspects of their instructional practice because they increase the *value* of a successful outcome, and that these improvements will be concentrated on aspects of teaching where teachers have the strongest *expectation* that their effort will lead to improvements. I hypothesize that expectation of success is influenced in part by the specificity of strategies outlined by the instructional rubric and the difficulty of the instructional tasks. Specifically, the research questions I address in this paper are:

1. Do teachers respond to the incentives embedded in a consequential, observation-based evaluation system by improving some aspects of instructional practice more than others?
2. Are the greatest improvements concentrated on instructional standards that are (a) more specific and (b) less difficult?

The next section connects the expectancy-value model of teacher motivation to the evaluation process employed in IMPACT.

3. DCPS IMPACT

3.1 Structure of IMPACT. The District of Columbia Public Schools introduced the IMPACT teacher evaluation system in 2009. Under IMPACT, teachers are evaluated using multiple measures of performance and are eligible to earn large financial bonuses and permanent salary increases for excellent performance and can also be terminated for poor performance. From 2009-10 to 2011-12, teachers received overall scores ranging from 100-400 that translate into performance ratings of Ineffective (I; 175 and below), Minimally Effective (ME; 176-249), Effective (E; 250-349) or Highly Effective (HE; 350

and above).[9] Teachers rated **I** are dismissed immediately, **ME** teachers must improve to an **E** rating in the next year or be dismissed, and **HE** teachers receive a one-time bonus and are eligible for a permanent salary increase if they receive a second consecutive HE rating in the next year. Effective teachers experience no additional incentives or consequences.

The overall IMPACT score comprises multiple measures of teacher performance, including: classroom observation ratings on a district-created protocol called the Teaching and Learning Framework (TLF), teacher individual value-added (IVA) in tested grades and subjects[10], teacher-assessed student achievement (TAS), and measures of teacher core professionalism (CP) and contribution to school community (CSC) and school value-added (SVA). For teachers in tested grades and subjects for whom IVA can be calculated (called IMPACT Group 1 in DCPS), this measure comprises 50% of the overall score, with TLF (35%), CSC (10%) and school value-added (5%) making up the rest.[11] For general education teachers without value-added scores (IMPACT Group 2), 75% of their overall IMPACT score is based primarily on TLF with minor weight given to TAS (10%), CSC (10%) and SVA (5%). Poor performance on tenets of professional practice such as excessive tardiness can result in a subtraction of points from the total score, but there are no positive points added if these core professional expectations are met.

[9] In 2012-13, DCPS made two changes to the way they assign overall performance ratings: they raised the threshold for Minimally Effective from 175 to 200, and they created a Developing rating from 250-299 to differentiate teachers within the Effective rating category. Since the treatment years of our study are 2009-10 through 2011-12, this is only relevant to the extent that all teachers faced somewhat more rigorous overall scoring in our last year of outcome data.

[10] Value-added models seek to isolate a teacher's contribution to student test score growth by controlling for observable characteristics of students and schools.

[11] The weight of IVA was reduced from 50% to 35% in 2012-13 for Group 1 teachers, and TAS was added as 15% of the overall score. Teachers for whom value-added is calculated comprise about 17 percent of the teaching workforce.

In addition to being among the earliest multi-faceted evaluation programs introduced, IMPACT is also unique in a number of other ways. First, IMPACT was implemented district-wide from the beginning, rather than as a pilot program or for only a portion of teachers in the district. As such, teachers may have internalized that the policy would persist over the course of the study period.[12] IMPACT is also unique in the strength of its incentives and the multiple performance measures upon which those incentives are based. The financial incentives in DCPS are meaningfully larger than other recent performance pay experiments, and include permanent increases to base salary as well as the unique threat of dismissal in the case of poor performance. Moreover, IMPACT's strong incentives are based primarily on a standards-based observation rubric of professional practice which is more transparent and actionable to teachers than measures based directly on student achievement.

3.2 The Teaching and Learning Framework. Particularly important to this study is the classroom observation component of IMPACT: the district-created rubric known as the Teaching and Learning Framework (TLF). TLF draws on instructional research including Charlotte Danielson's Framework for Teaching (FFT; 2007), the University of Virginia's Classroom Assessment Scoring System (CLASS; LaParo, Hamre & Pianta, 2008) and Wiggins and McTighe's Understanding by Design (2005), with goals of (1) creating a common language to discuss teaching and learning, and (2) providing clear expectations for teacher performance in DCPS (DCPS IMPACT

[12] A notable exception to this was a brief period of uncertainty regarding the persistence of IMPACT following the first year of implementation when Mayor Adrian Fenty lost the Democratic primary election to challenger Vincent Gray in September 2010. Fenty had championed Chancellor Michelle Rhee and her policies, and it was unclear to what extent the program would persist under the new mayoral administration. In October of 2010, however, Gray installed Deputy Chancellor Kaya Henderson as interim Chancellor, and made her appointment permanent in June, effectively and confirming the continuance of IMPACT under his mayoral administration.

Guidebook, 2010). The framework was originally designed by teachers, school leaders and central office staff in 2008-09 and was streamlined after the first year of implementation, yielding the nine teaching standards (called “Teach” standards in DCPS) in Table 1. The “Teach” standards are not explicitly classified into sub-categories, but the first seven standards relate mostly to instruction (content delivery, questioning technique, etc.), while Teach 8 and 9 relate more to the classroom environment.[13] The TLF rubric contains detailed descriptions of four levels of performance: Ineffective, Minimally Effective, Effective, or Highly Effective practice. Descriptions of Effective practice for each standard in the rubric are included in Appendix Table A1.1. TLF provides descriptions of both teacher and student actions that observers may use to anchor their judgments of performance. For instance, one descriptor of effective practice for Teach 1: Lead well-organized, objective-driven lessons states: “The objective of the lesson is clear to students. For example, the teacher might clearly state and explain the objective, or students might demonstrate through their actions that they understand what they will be learning and doing” (DCPS Guidebook, 2010). Teach 2: Explain content clearly, indicates that effective practice is characterized by “Explanations of content [that] are clear and coherent, and… build student understanding of content” (DCPS Guidebook, 2010). Some standards, Teach 5, 6 and 7, which will be discussed later, even include a list of suggested strategies for demonstrating effective practice, in addition to the rubric performance-level descriptors. Teachers are observed using TLF up to five times each year: three times by a school administrator, and twice by a Master Educator—a content area expert who is employed by the district expressly for the purpose of conducting

[13] Two other domains of TLF called “Plan” and “Increase Effectiveness” are described in the IMPACT guidebook, but are not scored.

evaluations.

Like other classroom observation protocols such as FFT and CLASS, TLF represents a set of hypotheses about the teacher practices that help students learn. One way to assess the validity of the information captured by standardized protocols is to examine their relationship to other measures of effectiveness, such as value-added. There have been a number of researcher-designed observation protocols developed over the last decade, some of them content-specific such as the Mathematical Quality of Instruction (MQI; Hill et al., 2005) and the Protocol for Language Arts Teaching Observation (PLATO; Grossman et al., 2013), and others content-agnostic such as CLASS and FFT. CLASS, FFT, MQI and PLATO were all included in the recent large-scale Measures of Effective Teaching (MET) project, which found meaningful relationships between these observation instruments and other measures of teacher effectiveness including value-added and student surveys (MET 2012). The correlation between each instrument and implied value-added[14] in the same year for the five protocols ranged from 0.12 to 0.34 (MET, 2012). This aligns with other work in the field that has also identified low to moderate correlations between classroom observation scores and contributions to student test outcomes (Grossman et al., 2013; Kane, Taylor, Tyler & Wooten, 2011). In DCPS, we find that TLF is also moderately correlated to individual value-added measures (IVA) during the study period. Over the three years from 2010-11 to 2012-13, we observe a 0.33 same-year correlation between TLF and overall IVA, with a 0.35 correlation for IVA in math and 0.30 in reading. This provides suggestive evidence that TLF captures some aspects of teaching performance that are also reflected in a teachers' ability to

[14] These correlations between observational measures and implied value-added in the same year are based on the relationship between value-added and observational measures from different sets of students to avoid spurious correlation due to unmeasured student traits.

increase student achievement on standardized tests, and that the strength of this relationship for TLF may be similar to other observation protocols in use in the field.[15]

3.3 Expectancy across Teach Standards. The conceptual model in this paper suggests that teachers' expectancy of success will vary across instructional standards if there are meaningful differences in (1) the specificity of the description of effective practice provided by the TLF rubric, and (2) the difficulty of the instructional task. I identify standards that are distinctly high in terms of their specificity based on a review of the rubric, as well as the most and least difficult areas of instruction as suggested by outside research and the first year of IMPACT data. While true differences in how specificity and difficult impact teachers' expectancy would best be informed by a survey that captures teachers perceptions of these constructs, identifying those standards that are exceptionally low and high on these constructs provides guidance in forming hypotheses regarding the standards where teachers are most likely to focus when faced with strong incentives.

Three standards on the rubric stand out as being more specific and educative in their descriptions of effective practice. Teach 5: Check for understanding, Teach 6: Respond to misunderstanding, and Teach 7: Develop higher-order thinking through effective questioning, all include a list of concrete strategies for effective practice in the rubric itself. For instance, Teach 5 includes suggestions such as "Use exit slips," "Have students respond on white boards," and "Use think-pair-share."[16] Teach 6 suggests that teachers might "Use cue cards," or "Use think-alouds." These lists represent concrete,

[15] More formal investigation into the psychometric properties of TLF is currently being conducted as part of a UVa-Stanford IES practitioner/researcher partnership in DCPS.

[16] An example strategy list for Teach 5 from the DCPS IMPACT Guidebook can be found in the appendix; Figures A1.1 and A1.2.

discrete actions that teachers can incorporate into their classroom routines to demonstrate effective practice, and that evaluators will recognize as effective following TLF. The remaining six standards all contain rubric descriptions of effective practice, but contain no comparable strategy list for teachers to consult. [17] As a result, based on the conceptual model and the strategies offered in the TLF rubric, we might expect teachers to have the highest expectancy of successful improvement on Teach 5, 6 and 7.[18]

The classification of standard difficulty is less straightforward. Since basing task difficulty on data from the present sample would be endogenous, I instead use outside research on classroom practice and baseline data from the first year of IMPACT[19] in DCPS to determine the instructional standards that may have the highest and lowest inherent level of difficulty. Both of these approaches have a similar drawback: neither the original IMPACT rubric nor the rubrics employed in other districts describe the exact instructional tasks used in DCPS during our study period.[20] As a result, I look across

[17] In my judgment, there are also some, more limited suggestions embedded within the rubric text for Teach 3 and 4. For example, Teach 3 suggests "…the teacher might differentiate content…(using strategies that might include…flexible grouping, leveled texts, or tiered assignments) in order to ensure that students are able to access the lesson" (DCPS IMPACT Guidebook, 2010). However, in the interest of transparency, I only classify standards as containing or not containing a strategy list. In future work, rubric reviews by independent experts or rubric text analysis may make a more granular ranking of standard specificity possible.

[18] In the spring of 2013, DCPS introduced a "suggestion document" *for Teach 2 only* which included concrete strategies for effective instruction. This could be argued to alter the specificity of Teach 2 in our study for 2012-13 only. A suggestion document for all standards was released the following year, in 2013-14 (Source: Meeting with DCPS IMPACT team, 04/28/16).

[19] The first year of IMPACT data is not outcome data in our study, as it is used to assign teachers to an initial rating category that has associated incentives in the next year.

[20] For instance, in DCPS, Teach 3: Engage students at all learning levels in rigorous work, was essentially newly introduced in 2010-11 following the districts' revisions to the rubric. Effective practice on this standard requires that the teacher makes the lesson accessible and challenging to almost all students, knows each student's level and differentiates content accordingly, ensuring the lesson is student-centered and students have ample time to practice and demonstrate learning (DCPS Guidebook, 2010). The previous version of Teach 3 called "Engage all students in learning," measured student engagement, asking evaluators to determine the proportion (less than half, about half, three-quarters, or nearly all) who were actively engaged throughout the lesson, while the previous Teach 7: Invest students in learning, combined some elements of knowing and engaging students at their level with the communication of high

studies with observation rubrics derived from the Danielson *Framework for Teaching*[21] and the first year of IMPACT not to rank the difficulty of all standards, but only to identify the consistently most and least difficult and examine patterns. This information is summarized in Table 1.2. Two clear conclusions emerge. First, across all studies, the lowest scoring item relates to the use of effective questioning and discussion techniques.[22] The second-lowest scoring standard varies across studies, including various elements of instruction such as using assessment, providing students with feedback, addressing misconceptions, and focusing students on the lesson objective. Second, all of the highest scoring items in the external studies relate to the classroom environment: these include creating a respectful environment, managing student behavior, and organizing the class time and space. These dimensions of instructional practice are helpfully clear in the current version of TLF, with Teach 7 describing the development of higher-level understanding through effective questioning, and Teach 8 (Maximize instructional time) and 9 (Build a supportive, learning focused classroom) describing effective classroom environments. As discussed previously, while a traditional interpretation of expectancy-value theory suggests that we should expect to see stronger effects of incentives on less difficult standards, i.e., Teach 8 and 9 (Atkinson, 1957) and weaker effects for more difficult standards (e.g., Teach 7), these may or may not be the most/least difficult areas of practice to improve. I next turn to describe the data and sample that will be used to explore how IMPACT's incentive effects vary across areas of

expectations to students. Similarly, observation instruments used in other districts also employ similar but not analogous instructional standards.

[21] While DCPS drew on many sources in the creation of TLF, it shares the most instructional competentices with FFT.

[22] This was called "probing" for higher-level understanding in DCPS in 2009-10, but the rubric language makes clear that this is accomplished through questioning and extending student discussion (DCPS Guidebook, 2009-10).

instructional practice in DCPS.

4. DATA AND SAMPLE

4.1 Data. This analysis employs administrative data on teachers in DCPS from 2009-10 to 2012-13. In each year I have detailed teacher evaluation records from IMPACT. IMPACT data include a teacher's overall IMPACT score and rating, and scores for each component, including ratings on each the nine TEACH standards for each observation. Two of these observations are conducted by Master Educators, and three are conducted by administrators. Human resources data files provide demographic information on teachers such as race, gender, and education, and allow the construction of a measure of teaching experience using salary information.

4.2 Sample. There are approximately 3,500 teachers in DCPS in each year of the study. I limit my analysis to general education teachers. This restriction eliminates very specialized teachers, such as special education or ESL teachers and insures that the classroom observation outcomes are from the TLF framework and not another instructional rubric.[23] This eliminates roughly 800 non-general education teacher observations in each year. I make two additional sample restrictions for the regression discontinuity (RD) analytic sample. Because RD designs capture a local estimated treatment effect *at a specific point in the performance distribution*, it is important that observations far from that point are not having a large influence on the estimates. To ensure this, I construct two separate analytic samples around the ME (i.e., Minimally Effective) and HE (i.e., Highly Effective) performance thresholds. The "ME" RD sample is comprised of teachers with initial scores in the ME or E performance bands (e.g.,

[23] Non-general-education teachers are often rated on non-TLF rubrics that contain special education or early childhood teaching competencies.

assigned an initial score from 176 to 349), and the "HE" RD sample is comprised of teachers in the E or HE performance bands (e.g., assigned an initial score from 250 to 400). By eliminating the non-relevant performance band, I ensure that our local estimates are not being driven by effects from the other performance threshold.[24] Second, I focus on teachers who have earned their first ME or HE rating in the prior year. At the ME threshold, I focus on the performance improvement of teachers who are experiencing the dismissal threat associated with being once-rated Minimally Effective, but who choose to return to the district.[25] At the HE threshold, teachers earning their second HE rating have already obtained the permanent salary increase and no longer experience the incentive.[26] It is important to note that there is only outcome data in our sample for teachers who return to an instructional position in the next year. This attrition is not an internal-validity threat but rather a component of the IMPACT theory of change, which incorporates both selection and incentive effects. Dee & Wyckoff also address this issue. They find that "an ad hoc empirical decomposition based on our RD design suggests incentive effects rather than selection effects. Using the sample of teachers who returned, we estimated an RD specification where IMPACT performance in the prior year is the dependent variable. We find small and statistically insignificant effects that are consistent with the hypothesis of behavioral change in response to the incentives" (2015. p. 21). One important implication of this is that my estimated effects should be interpreted as improvements in practice *conditional* on teachers' decisions to remain in DCPS.

I examine RD outcomes for three cross-sections: 2010-11 teacher outcomes as a

[24] Further explanation of RD designs follows in the next section on empirical strategy.

[25] Including teachers who earn their second ME rating and are mechanically separated results in larger estimated effects.

[26] These sample restrictions echo the technique of frontier RD which uses multiple variables to determine assignment to treatment (Reardon & Robinson, 2012; Wong, Steiner & Cook, 2013).

function of 2009-10 IMPACT ratings, 2011-12 teacher outcomes as a function of 2010-11 IMPACT ratings, and 2012-13 teacher outcomes as a function of 2011-12 IMPACT ratings. I refer to these cross-sections by their outcome year, t+1. Table 1.3 provides an overview of the treatment, outcome and covariate measures for the descriptive sample and the two RD analytic samples. We observe 8,063 teacher-year observations in the full sample; 2700 in 2010-11, 2,694 in 2011-12, and 2,669 in 2012-13. In the descriptive sample of all general education teachers, 80 percent of teachers are retained in year t+1. The retention rate is slightly higher in the ME sample (81 percent) and notably higher in the HE sample (85 percent). Intent-to-Treat (ITT) and treatment status and demographic statistics are reported for this ITT sample, though fewer teachers persist into year t+1, resulting in fewer observations for outcome variables. The proportion of teachers who receive an initial ME or HE rating is relatively low—only 16 percent of the ME sample and only 13 percent of the HE sample—indicating that a majority of the observations in both samples are teachers scoring in the Effective performance band. Not surprisingly, teachers in the HE sample have higher scores on IMPACT and TLF overall in year t+1, as well as each of the Teach standards. The lowest-rated standard (Teach 7) and highest-rated standard (Teach 9) are consistent across samples, and are also consistent with the hypothesized most and least difficult competencies as identified by external research. Teacher demographic traits are relatively similar across all three samples: teachers in the district are predominantly female (71 percent in the base sample), hold graduate degrees (67 percent), and over half are black (52 percent).

5. METHODOLOGY

5.1 Descriptive analysis. I begin the empirical work in this paper by providing a descriptive look at changes in instructional practice in DCPS as measured by TLF in the full sample. I provide cross-sectional estimates that include improvements that may result from compositional shifts in the workforce, and also examine within-teacher changes as well. I examine within-teacher change for only two cross-sections of data, from 2010-11 to 2011-12 and from 2011-12 to 2012-13, because the TLF rubric underwent a revision after the first year of implementation and thus we do not have consistent longitudinal data on teaching practice from the first to the second year.[27]

5.2 Regression discontinuity designs. In the main results for this paper, I use a regression discontinuity design to estimate the causal effect of receiving an ME or HE rating on teaching practice in the next year. A critical design feature in IMPACT that allows this type of analysis is the sharp incentive contrast that teachers experience based upon their overall IMPACT score in the prior year. As illustrated earlier, a teacher who scores 249 IMPACT points is assumed to be no different than a teacher scoring 250, except one teacher receives a low-performance signal and threat of dismissal in the next year if she does not improve, while the other receives the message that her performance meets the district's standard.

RD designs have a strong causal warrant due to the arguably random assignment of treatment (here, the incentives associated with an ME or HE rating) right around sharp cut points (Campbell, 1969; Lee & Lemieux 2009). An important concern with any RD design is that there may be systematic sorting of teachers across the performance

[27] This is not an issue in the main RD analysis because we only use TLF as outcome data beginning in 2010-11. However, it may be relevant to the extent that teachers were learning revised district expectations for teaching practice following the first year of IMPACT.

threshold. If teachers are able to manipulate the variable that "assigns" them to one side of the threshold or the other, this introduces bias into the estimates because there are likely other differences as well between teachers who are able to sort over the threshold and those who are not. The process of appealing low ratings and obtaining revised, final IMPACT scores could introduce this type of bias into our estimates. To avoid this, I use teachers' initial IMPACT scores to determine assignment to treatment, and the estimates we detect can thus be interpreted as Intent-to-Treat (ITT) estimates.

I present graphical and parametric evidence to illustrate the relationship of this assignment variable with both final ME/HE ratings in the same year (i.e., first stage estimates) and measures of their teaching practice in year t+1 (i.e., reduced-form estimates). The core estimating equation is as follows:

$$Y_{is(t+1)} = \alpha + \beta I(S_{it} \leq 0) + f(S_{it}) + \theta X_{it} + \delta_s + \varepsilon_{it} \qquad (1)$$

where $Y_{is(t+1)}$ is a measure of teaching practice for teacher i in school s in year t+1. The parameter β reports the intent-to-treat effect of receiving an initial ME or HE rating on $Y_{is(t+1)}$, or the "jump" in the outcome variable at the ME/HE performance threshold conditional on a smooth centered function of S_{it}, teacher covariates (θX_{it}), and school fixed effects (δ_s). The term ε_{it} is a mean-zero error term, and robust standard errors are reported to allow for heteroscedasticity. The inclusion of school fixed effects ensures that our estimates are not drawn from comparisons of classroom practice in t+1 for teachers across the threshold who face different school contexts. Differences in the improvement of classroom practices have been demonstrated to differ across school contexts (Sass, Hannaway, Xu, 2014; Xu, Ozek, Hansen 2015), and are controlled for here so that they are not reflected in estimated incentive effects.

5.3 RD First stage. First-stage estimates rely on this general specification, but instead report how strongly receiving an initial ME/HE rating affects receipt of a final ME/HE. In each year of this study, the jump in treatment across the performance threshold is either sharp or nearly sharp. When this is the case, ITT estimates are nearly equivalent to "treatment-on-the-treated" estimates. Figure 1.2 illustrates this: in each panel, the fitted line indicates the probability of receiving a final ME rating conditional on receiving an initial ME rating. In Panel A, a limited number of ME rating appeals (85) in 2009-10 result in a first stage estimate of 0.8. In this year, the reduced-form results are interpreted as applying to teachers who "complied" with their initial status in the first year. In Panel B the relationship between initial and final ME ratings is virtually sharp: there were only three successful appeals in 2010-11. In 2011-12, (Panel C) there were no successful appeals. At the HE threshold there are no successful appeals in any year—only teachers initially rated HE receive a final HE rating. At the ME threshold in 2010-11 and 2011-12 and at the HE threshold in all years, the ITT estimate is either nearly or exactly the same as the TOT estimate since all, or nearly all, teachers have complied with their initial status.

5.4 Checking RD Assumptions. The two primary threats to the internal validity of estimates in an RD design are manipulation of treatment status around the performance threshold and incorrect specification of the functional form of the assignment variable. Literature on RD designs recommends a number of analyses to provide a check on these assumptions (Imbens & Lemieux, 2008; Lee & Lemieux, 2009; McCrary, 2008).

Programmatic details from IMPACT suggest that it would be difficult to manipulate initial IMPACT score. As described earlier, IMPACT scores are comprised of a number of components, and are rated at different times by different sources. Some components (e.g., Individual Value Added) are not calculated until the summer, after all other data has been submitted. Thus, there is not a strong anecdotal basis for suspecting manipulation of assignment to treatment. An empirical examination of this assumption also suggests that it holds: the McCrary test, which examines the density of observations around the performance threshold, fails to reject the null hypothesis of a smooth distribution (Figure A1.3). Another way to examine the assumption of local quasi-random variation is to examine the balance of teacher covariates (Table A1.2). Here I do detect some imbalance: at the ME threshold, teachers just receiving an ME are more likely to be in Group 1 in 2011-12. This raises the question of whether, in this year, teachers in Group 2 may have been able to manipulate into the Effective performance band. However, since that this is the one result at the ME margin out of 36 separate regressions to attain a traditional (.05) level of significance, it does not represent a major threat. There is also some evidence of teacher racial imbalance around the HE performance threshold. This imbalance is only statistically significant at traditional levels for 2010-11, when teachers just on the HE side of the performance threshold are 10 percent more likely to be white and less likely to be black. [28] Although the estimates are positive for each of the next two years, they are not statistically significant. Further, as we will observe below, I do not find robust effects of the incentives on teaching practice at this margin.

The validity of RD designs also relies on correct specification of the functional form of the assignment variable. I examine robustness of my estimates to different

[28] Dee and Wyckoff (2015) also detect this imbalance and note that it could be an “artifact of multiple comparisons” (p. 32).

functional form assumptions. The primary specification conditions on a linear spline of the assignment variable with a slope that is allowed to vary on either side of the performance threshold. I also explored a specification that conditions on a quadratic form of the assignment variable allowed to vary across the threshold, and I report these results in the main tables because the estimates are sometimes sensitive to this change. In examining the model fit of the linear and quadratic specifications using the Akaike Information Criteria (AIC), it is not consistently clear which functional form to privilege: they are often close, and the better fit varies over years and instructional standards. Often my results are consistent despite varying functional form of the assignment variable, but when they are not I look to the AIC for guidance. Finally, I also report results based on local linear regressions (LLR) that restrict the bandwidth to an increasingly narrow range around the performance thresholds to decrease the model's reliance on functional form assumptions. These results do not always maintain statistical significance as the number of observations and statistical precision decrease, but we expect to see the magnitude of point estimates remain relatively stable in order to confirm that observations far from the performance threshold are not driving results.

This empirical strategy leverages cleanly identified quasi-random variation to determine whether highly incentivized teachers make changes to their classroom practice, conditional on their decision to remain in the district in year t+1. Like others who have sought to understand how treatment effects differ across more granular outcome measures (e.g., Cohodes, 2015) I look for differences in the magnitude and statistical significance of effects. Since best practice in RD designs also includes the triangulation of evidence across multiple specifications as well, interpretation of results is not without ambiguity. However, careful examination of estimates across a number of specifications allows me to highlight patterns of consistent behavior change and consider whether or not these are in line with the anticipated response outlined in the conceptual model.

6. RESULTS

6.1 Descriptive Changes in TLF Scores in the Full Sample. Examination of TLF outcomes over time in DCPS provides suggestive evidence that classroom practice is improving. As discussed previously, Figure 1.1 shows the performance distribution of average TLF scores for each of the outcome years in our data: 2010-11, 2011-12 and 2012-13.[29] Though these distributions largely overlap one another, each successive year sees the curve shift to the right, indicating an overall improvement in teaching practice in the district. Table 1.4 provides more detail on the average and standard deviations of individual TEACH standards that underlie this shift. The overall average improves by 0.05 TLF points in each year, or slightly more than 10 percent of a standard deviation (SD). This upward trend year-over-year is consistent across administrator and master educator ratings and most individual standards with the sole exception of Teach 4, which declines slightly from 2011-12 to 2012-13. It is also worth noting that the distribution of ratings compresses somewhat over time with the SD declining by 0.06 on overall TLF score from 2010-11 to 2012-13. For this reason and for ease of interpretation, I report effect sizes rather than TLF points in the main RD results.

While increasing average TLF scores are encouraging, there are a number of reasons that we would not want to interpret this as evidence that IMPACT is causing teachers to improve to their practice. One clear reason is because these overall

[29] Figure 1.1 and Table 1.4 both employ the full general education population (N=2,694 in 2010-11, N=2,669 in 2011-12, N=2,592 in 2012-13), which differs from the analytic sample in that it includes teachers who are new to the district.

improvements include both compositional changes to the workforce, and improvements within the current workforce. Table 1.5 examines within-teacher year-to-year change on TLF. The individual growth for teachers who remain in DCPS for two consecutive years is very similar to the overall improvement in Table 1.4, indicating that it is likely not compositional change that is driving the improvements in Figure 1.1. Even so, these descriptive improvements may be caused by factors other than IMPACT. They do provide suggestive evidence of an overall increase in TLF that is consistent with the conceptual model that IMPACT has induced teachers to improve their teaching practice. To explore this more rigorously we now turn to results from the regression discontinuity design.

6.2 RD Results – Minimally Effective Performance Threshold

6.2.1 Graphical RD Evidence. Graphical representation of RD analysis provides a clear illustration of how estimates are derived: the outcome variable is plotted as a function of the assignment variable on either side of the performance threshold. A "jump," or discontinuity, at the threshold is evidence of a treatment effect. Examination of average TLF outcomes assigned by administrators and master educators for all three cross-sections of data (Figure 1.3) illustrate important heterogeneity in incentive effects over time.

Panel A and B show no evidence that an ME rating affects teaching performance following the first year of IMPACT using either administrator or master educator scores. However, for 2011-12 and 2012-13 (Panels C through E), teachers rated as ME appear to improve their overall TLF ratings relative to otherwise similar teachers rated as E by 15 to 40 percent of an SD. As noted earlier, the end of 2010-11 was the first time that twice-

ME teachers were separated from the district. As such, 2011-12 is the first year that we observe performance effects once the dismissal threat associated with an ME rating has gained credibility. Thus, it may not be surprising to see larger effects of the dismissal policy in 2011-12 and beyond.

6.2.2 Parametric Results. The graphical evidence suggests that IMPACT incentives had a positive effect on teaching performance in 2011-12 and 2012-13 for teachers who were just rated ME. Table 1.6 presents estimates for the overall TLF average, master educator average and administrator average for the aggregate three-year sample, and separately for each cross-section. The two columns in each cross-section condition on different functional forms of initial IMPACT score: the first column uses a linear spline and is the preferred specification, while the second also includes a quadratic spline, as a robustness check. Column (1) provides some evidence of a positive effect of the ME rating on average TLF scores, though the combined TLF average that includes both master educator and administrator ratings is non-significant and the point estimate diminishes in the quadratic specification. However, the TLF average assigned by master educators (0.148) is moderately significant ($p<0.10$), and robust to decreases in bandwidth in the LLR specification (Table 1.8).

As the graphical evidence suggests, parametric estimates also report important heterogeneity in the estimated effect of an ME rating over the three years of the study. After observing overall null results in the first year of IMPACT, the effects of an ME rating on teaching in 2011-12 are positive and robust to decreases in bandwidth and flexible assumptions regarding functional form, though the estimates are sometimes imprecise and do not always attain statistical significance, varying between 0.185 SD in the linear and 0.363 ($p<0.05$) in the quadratic specification. Point estimates are also large and statistically significant in 2012-13 in the composite and master educator results (0.332 to 0.390 SD) but the results for administrator ratings are smaller do not hold up as well to inclusion of the quadratic term or in the LLR specification. These results indicate that teachers under the threat of dismissal can be responsive to this incentive—sometimes improving nearly a third of a standard deviation from one year to the next, conditional on their decision to remain in teaching in DCPS.

6.2.3 Effects for Individual Teach Standards. I next turn to the core analysis of this project: RD analysis of individual Teach standards at the ME threshold (Table 1.7). I focus this analysis on ratings assigned by master educators for two reasons: first, as district experts in evaluation, they are slightly more adept at differentiating performance between standards (the average inter-item correlation for master educators is 0.50 compared to 0.53 for administrators). Second, the aggregate results are less influenced by negative administrator ratings in 2010-11.[30] Unsurprisingly, individual standard results are also heterogeneous by year. In 2010-11, ME teachers did not differentially improve any of the standards relative to otherwise similar E teachers. However, patterns of improvement emerge for 2011-12 and 2012-13.

The conceptual model outlined in this paper predicts that we should see larger effects on Teach 5, Teach 6 and Teach 7, on which teachers might have a stronger expectation of successful improvement due to the inclusion of concrete strategies in the rubric. However, Teach 7, or the use of effective questioning and discussion, is also identified as the most difficult area of instructional practice to master, and thus this could

[30] Results that also incorporate administrator ratings are directionally similar but differ in the two ways described above; they can be found in Appendix Tables A1.3 and A1.4.

potentially dampen the expected effects of strategy inclusion. Looking at patterns across 2011-12 and 2012-13 in the main effects (Table 1.7), and considering the LLR specification (Table 1.8), we can first note the presence of broad positive effects in these years: point estimates are nearly always positive, and often attain statistical significance.[31]

However, the preponderance of evidence looking at the aggregate three-year results and the most consistent positive and significant results from 2011-12 and 2012-13 largely supports the hypotheses outlined in the conceptual model that teachers will experience the greatest improvement on more prescriptive and less difficult standards. The results from all years in Columns (1) and (2) provide some support for the predictions of the conceptual model: Teach 5 and 6 are the largest estimates (0.212, $p<0.01$ and 0.198, $p<0.05$) and retain their magnitude both when the quadratic assignment variable is included and when the bandwidth is reduced. However, Teach 4 (Provide students multiple ways to engage with content), which was not specifically predicted by our model as a standard where teachers would have a particularly strong expectation of success, also shows positive overall effects (0.173, $p<0.05$). Teach 7, a "high-specificity" standard which also represents the most challenging area of instruction captured by TLF, is not statistically significant in the aggregate results. None of the remaining standards (1, 2, 3, 8, 9) show compelling aggregate effects of the ME incentive.

In 2011-12 and 2012-13 (Table 1.7, Columns (5) to (8)), there are consistently large and positive effects across both years for all high-specificity standards (Teach 5, 6

[31] These broad positive effects that differ somewhat suggest that TLF may capture both general and specific factors related to teaching performance, and that a bi-factor analysis may yield interesting results. This is on my research agenda.

and 7) with the exception of Teach 7 in 2011-12, which is quite small (0.016), yet becomes large when the bandwidth is decreased to the tightest band around the threshold in the LLR specification (Table 1.8, Column (9)). Moreover, it is only on Teach 5, 6 and 7 in which we detect robustness in the LLR specification in both 2011-12 and 2012-13, though the magnitude and stability of the Teach 7 results are the least convincing. It is also worth noting the uniquely large and consistent effects we observe on Teach 5, which holds the distinction of being the highly-specific standard whose focus (checking for student understanding) never appeared among the most difficult areas of instruction in any of the outside research. We observe large (0.204-0.626 SD) effects on Teach 5 in 2012-12 and 2012-13, and these effects are robust to decreases in bandwidth and varying assumptions about functional form. Thus, it stands out as the standard that is most consistently responsive to the dismissal threat. Teach 3 (Engaging students at all levels in rigorous work) and Teach 4 are not identified as standards which our conceptual model specifically predicted effects through either high-specificity or low task difficulty, yet we do observe moderately-sized main effects in both years, though these effects are only robust in 2012-13 in the LLR specification.

We look to Teach 8 and 9, the standards related to classroom environment, for suggestive evidence of whether low-difficulty standards may be more likely to experience incentive effects. The results are inconclusive: there are large effects in the quadratic and LLR specifications in 2011-12 on Teach 8 and 9, but there are no statistically detectable effects in 2012-13. One potential explanation for this heterogeneity could be that the ITT population in the two years face somewhat different school contexts, which could be particularly relevant for Teach 8 and 9 since they describe effective classroom environments. I examine this in the data and find that this does not appear to be the explanation, as the school poverty rate for the ITT population is consistent across years. Another potential explanation for these heterogeneous effects could be that another driver of expectancy such as availability of high-quality professional learning opportunities, may differ over the years for these standards in important ways, but we unfortunately do not have access to this information.

However, one anecdotal difference that I am aware of that may explain the notably large and robust effects we observe on Teach 2 in 2012-13 is that a "suggestion document" was introduced for this standard only in early 2013. This document provided evaluators with concrete strategies for Teach 2 to share with teachers in their feedback. Thus, the concreteness of the feedback teachers received on Teach 2 may have been different in the second half of 2012-13 than in other years. Interestingly, we see a much larger incentive effect on Teach 2 in 2012-13 than we do in 2011-12.

Taken together, this analysis shows that there is meaningful variation in the size and consistency of incentive effects on different aspects of instructional practice for teacher facing the strong negative incentives associated with an ME rating. These differences are mostly consistent with the predictions of the conceptual model that teachers will focus on improving the aspects of instruction where they theoretically have the strongest expectation of success. However, we also observe positive effects on some standards—notably Teach 3 and 4—where the classification of specificity and difficulty did not predict the positive effects that we observe. Potential alternative hypotheses for these results will be revisited in the discussion.

6.3 RD Results- Highly Effective Threshold

6.3.1 Graphical Evidence. I now examine incentive effects for teachers who have just earned their first Highly Effective rating and are thus eligible for a permanent salary increase in the next year if they are again effective. Overall, these teachers appear to improve TLF scores in the first but not subsequent years (Figure 1.4). While we do observe a discontinuity in the 2010-11 graph, it is also true that in all years the observations in the bins just to the left of the performance threshold (i.e., teachers who just missed an HE rating) have a relatively high conditional average TLF score. One possible explanation for this is that the financial incentives available to HE teachers may also be motivating to teachers who are just below the HE threshold: they narrowly missed earning a large financial bonus and may internalize their proximity to the award and increase their effort. The visual evidence at the highly effective threshold does not anticipate especially robust effects in our parametric estimates.

6.3.2 Parametric Estimates. Table 1.9 provides point estimates that capture the effect of receiving one's first highly effective rating on TLF performance in the next year. A quick look across the columns indicates that the regression coefficients are consistent with the graphical results: there is some evidence of an effect in the 2010-11 school year though it is only robust to inclusion of the quadratic term in ratings assigned by administrators, and there are no statistically detectable effects in 2011-12 or 2012-13. It is worth noting that there are relatively fewer teachers who earn their first HE rating in 2011-12 (131) or 2012-13 (274) relative to 2010-11 (409). Thus, the effects in 2010-11 are driving the aggregate results in large part as well. While there are positive effects in the aggregate results in the specification conditioning on a linear spline, these results are not robust to inclusion of a higher order term which is the specification recommended by the AIC.

6.3.3. Effects for Individual TEACH Standards. Table 1.10 examines heterogeneity across Teach standards at the HE threshold. As with the ME threshold, I focus on ratings by master educators. In contrast to incentive effects for teachers at the ME threshold, we do not see broad positive effects for once-HE teachers who are eligible for a permanent salary increase in preferred specifications across years. In 2010-11, we detect positive and significant effects across many areas of practice in the linear specification, but these are varyingly robust to inclusion of the quadratic term and in many cases this is the specification that better fits the data. Only Teach 8 and 9, the low-difficulty standards related to classroom environment, are statistically significant in the preferred specification and somewhat robust to inclusion of the quadratic terms and decreases in bandwidth (Table 1.11), though the magnitude of effects becomes much smaller (0.095 and 0.073). There is also some supporting evidence for each of these standards in the following years which see no main effects: in 2011-12, Teach 9 does not reach statistical significance in the main results (0.202), but becomes larger and statistically detectable as the bandwidth is decreased. In 2012-13, Teach 8 is significant at the 10 percent level in the preferred specification, and is somewhat robust to decreases in bandwidth.

The incentive effects for HE teachers eligible for a salary increase if their performance remains highly effective are far smaller than those at the ME threshold: they are smaller in magnitude and less robust, and are concentrated on standards that relate to the classroom environment and do not contain explicit strategies for demonstrating effective practice. As mentioned earlier, this may have to do with the treatment contrast for teacher at this threshold: teachers on both sides of the HE threshold experience somewhat compelling reasons to improve (i.e., permanent salary increase or first-time financial bonus). It is possible that teachers on both sides of the HE threshold are responding to the specific information contained in TLF. This makes the effects that we do detect on Teach 8 and 9 even more thought-provoking: why do these particular aspects of practice experience effects of an HE rating, when others do not? One potential explanation could be that teachers who receive the designation of HE in one year experience positive effects in terms of professional identity and confidence that show up in more positive classroom environments in the next year. These finding at both the HE and ME performance thresholds have important implications for evaluation policy, which are further discussed below.

7. DISCUSSION

This study provides novel evidence on how teachers change their classroom practice in the face of high-stakes performance evaluation. Nearly all states and districts have rolled out new teacher evaluation systems over the past five years, and these systems require significant investments of time and effort by teachers, school leaders, and district officials. The theory of action behind teacher evaluation posits that pairing accountability pressure with performance feedback will improve teacher performance more effectively than either mechanism could alone. Yet we know virtually nothing about how the elements of these systems actually impact teaching. The goal of this analysis is to provide new insight on this important question.

Consistent with prior literature on the effects of observation-based evaluation and feedback (e.g., Steinberg & Sartain, 2015; Taylor & Tyler, 2012), I find that teachers do improve their practice when they face strong incentives in IMPACT—especially low-performing teachers who face a dismissal threat if they do not improve. This contrasts with most prior studies of teachers' responses to test-based incentives, which find that teachers do not change their practice when offered a financial incentive to improve student test scores but no information on how to do so. One clear takeaway from this work in light of prior findings is the importance of the information provided by evaluation for the improvement of performance. By articulating proximal teacher behaviors and holding teachers accountable for these rather than for test outcomes, teachers have a better sense of how to respond to the incentives.

Moreover, I find that, at least for low-performing teachers, incentives most strongly induce improvement for standards where the rubric is most educative in nature. In 2011-12 and 2012-13, we observe the most consistent, positive and significant effects across both years for Teach 5 and 6, the standards classified as having high-specificity, but not identified as the most challenging standard to demonstrate effective practice (Teach 7). These standards, and Teach 2, for which the district introduced additional prescriptive guidance in 2013, also represent the largest effects by far in 2012-13, which is consistent with a story of teachers who face accountability pressure learning to focus on these most prescriptive standards over time as the program persists. However, improvements do not always line up with the hypotheses generated by the conceptual model in all years: for instance, ME teachers experience positive effects on Teach 3 and 4, which do not offer suggested strategies for effective practice. This indicates that while the inclusion of prescriptive language in the rubric appears to augment the effects of incentives on many standards and in many years, it is not a prerequisite for changing practice.

Teachers who have received one HE rating and are eligible for a permanent salary increase if they receive another experience much smaller effects of this incentive than ME teachers who face a dismissal threat, and these effects are largely limited to Teach 8 and 9. The fact that we see no incentive effects on the most explicit standards for higher-performing teachers may have to do with the treatment contrast: teachers who just missed the HE rating may internalize their proximity to the large financial bonuses offered by IMPACT and also be experiencing an incentive effect in the following year. Alternately, concrete strategies for instructional practice may have less relevance for high-performing teachers than they do for low-performing teachers (e.g., these may be strategies they already employ). Or, consistent with our conceptual model, the value of the incentive for the HE teachers (i.e., salary increase) may have comparatively lower value than the incentive for ME teachers (i.e., job retention).

These findings have important implications for the design of evaluation policy. Perhaps most significantly, they offer an acute reminder of how the design—even the wording—of standardized observation instruments may critically impact the way that teachers respond to evaluation. Including concrete actions in the rubric that teachers can incorporate in their classrooms appears to often leads to better observation scores when teachers face strong incentives. This raises the question of whether these improvements in measured practice lead to authentic and ongoing changes in teaching that have benefits for students. High-specificity, less difficult standards may not be the highest leverage areas for improving student outcomes, but this study suggests that it is where low-performing teachers will focus given the design of the current system.

In the next chapter of this dissertation, my colleagues and I provide some evidence on this question. We examine incentive effects on student achievement at ME threshold and we do detect moderate improvements in student achievement in the same years where teaching improved, though the gains are concentrated in math (Adnot, Dee, Katz & Wyckoff, 2016). We cannot causally link the improvements we observe in teaching to the improvements in student achievement, but it is conceptually unlikely that these gains are unrelated. Future studies could experimentally vary the inclusion of specific/non-specific rubric language in a lower-stakes setting to examine how rubric specificity causally affects teachers' improvement on different areas of practice and student outcomes. It is more difficult to exogenously vary the difficulty-level of certain types of instructional tasks. However, a meta-analysis of classroom observation item-level results for quasi-experimental and experimental studies of teacher evaluation (e.g., Steinberg & Sartain, 2015; Taylor & Tyler, 2012) could be informative and will be increasingly feasible as this literature grows in the coming years.

The policy implications for these findings are clear: concrete suggestions offered in evaluation rubrics should either be distributed equally across all measured components of practice, or should be strategically employed to focus teachers toward areas of practice that research suggests are the highest leverage. If we know the most important areas of practice for teachers to improve, either overall or as individuals, then raising the specificity on these standards could focus teachers' attention here. Educational researchers have made strides in identifying high leverage instructional practices, however, there is still much more to be done in this area (Ball & Forzani, 2011; Cohen, 2015). For instance, there is evidence that mastering “foundational practices” such as time and behavior management and instructional planning can support teachers’ acquisition of more ambitious practices (Cohen, Chambers, Schuldt, Brown, & Grossman, 2015). Novice or low-performing teachers could receive more concrete descriptions of effective practice in these areas, which could motivate them to focus attention on these foundational practices first.

The finding that teachers’ ability to employ concrete suggestions could be limited by the inherent difficulty of particular instructional tasks also has implications for policy. This finding suggests that including prescriptive rubric language may not be sufficient for improvement for the most difficult teaching practices even in the context of strong incentives, and that greater support must be targeted toward these areas as well. However, this work also underscores how small adjustments in favor or against particular standards (e.g., adding a list of strategies to the instructional rubric) can have large effects on teacher behavior. Thus, districts should carefully consider how changes to evaluation programs or associated development opportunities may influence teachers’ expectancy and value across multiple goals, to minimize unintended consequences of well-meaning program revisions.

While this study makes an important contribution, it is not without limitations and more qualitative and quantitative work is needed to better understand the teacher responses introduced here. The empirical work in this paper provides internally-valid causal estimates of how instructional practice changes in the context of high-stakes evaluation, and the conceptual model provides one rationale for how we might understand these results. However, one important limitation associated with any study employing a regression discontinuity design is the local nature of the estimated effects. The estimates reported here are local average treatment effects that capture the contrast between teachers just earning the ME/HE rating, and cannot be generalized to teachers across the performance distribution.

There are also many alternate hypotheses for different factors that influence teachers' expectancy and value which, if tested, could change the implications of these findings for policy and practice. For instance, there is not sufficient data to understand how differences in professional learning opportunities may drive differences in how teachers focus their improvement efforts. If low-performing teachers are offered PD on the same skills that are articulated most specifically in TLF, then this could be an alternate explanation for why we see greater improvement in these areas. Additionally, variation in PD opportunities or systemic school- or district-wide focus areas over time could explain why results vary meaningfully across years. Determining how professional learning opportunities linked to evaluation results influence teachers' ability to respond to evaluation is an important area for future research, and one in which there is little work.[32] Finally, the expectancy-value model presented in this paper is based on teachers' *perceived* expectation and value of success, and additional survey work on teachers' perceptions of task specificity and difficulty, as well as the quality of professional learning opportunities that they receive across instructional standards would certainly provide additional insights in this area.

[32] One recent exception is Papay, Taylor, Tyler, & Laski (2016) which finds that matching teachers for peer-support based on classroom observation scores improves student test outcomes by 0.12 SD.

These findings have value beyond DCPS and beyond teacher evaluation in their generalizability to workplace situations where individuals face strong incentives and competing goals. My analysis provides empirical support for an application of expectancy-value theory in the area of workplace motivation in which individuals who experience strong incentives and face multi-faceted performance measures will be motivated to direct their attention toward performance measures where they expect they can be successful and on which they value a successful result. There is still much work to be done to understand teachers' responses to performance evaluation in the field of education, and how the elements of these programs support or hinder teachers' ability to help students learn. This paper provides a first-look at these important questions.

Table 1.1 Summary of the Teaching and Learning Framework

Standard	Description
Teach 1	Lead well-organized, objective–driven lessons
Teach 2	Explain content clearly
Teach 3	Engage students at all learning levels in rigorous work
Teach 4	Provide students multiple ways to engage with content
Teach 5	Check for student understanding
Teach 6	Respond to student misunderstandings
Teach 7	Develop higher-level understanding through effective questioning
Teach 8	Maximize instructional time
Teach 9	Build a supportive, learning-focused classroom

Source: DCPS IMPACT Guidebook, 2010-11.

Table 1.2 Identifying the Most/Least Difficult Instructional Competencies in External Research and IMPACT 2009-10

Study	Most Difficult		Least Difficult	
	Lowest Scoring	2nd Lowest Scoring	Highest Scoring	2nd Highest Scoring
Kane, Taylor, Tyler & Wooten, 2011 (Cincinnati Teacher Evaluation System)	The teacher engages students in discourse and uses thought-provoking questions aligned with the lesson objectives to explore and extend content knowledge.	(tied) The teacher provides timely, constructive feedback to students about their progress toward the learning objectives using a variety of methods, and corrects student errors/misconceptions.	The teacher creates an inclusive and caring environment in which each individual is respected and valued.	The teacher manages and monitors student behavior to maximize instructional time.
Sartain, Stoelinga & Brown, 2011(Chicago Excellence in Teaching Pilot)	Using Questioning and Discussion Techniques	Engaging Students in Learning	Organizing Physical Space	Creating an Environment of Respect and Rapport
Garrett & Steinberg, 2015 (The Measures of Effective Teaching Project)	Using Questioning and Discussion Techniques	Using Assessment in Instruction	Managing Student Behavior	Creating an Environment of Respect and Rapport
IMPACT, Baseline data from AY 2009-10 (author's own calculations)	Teach 5c: Probe for Higher-Level Understanding	Teach 1: Focus Students on Lesson Objectives	Teach 8: Interact Positively and Respectfully with Students	Teach 3: Engage All Students in Learning

Table 1.3. Descriptive Statistics, Base Sample and RD Analytic Samples

	Gen Ed Sample		ME RD Sample		HE RD Sample	
	N	Mean	N	Mean	N	Mean
Retained in DCPS, t+1	8063	0.80	6089	0.81	6630	0.85
Minimally Effective - ITT	8063	0.14	6089	0.16	-	-
Minimally Effective	8063	0.13	6089	0.15	-	-
Highly Effective - ITT	7555	0.11	-	-	6129	0.13
Highly Effective	7555	0.11	-	-	6129	0.13
Initial IMPACT Score, t	8063	299	6089	291	6630	315
IMPACT Score, t+1	6313	312	4958	302	5630	317
TLF Score, t+1	6312	3.15	4954	3.05	5629	3.20
TEACH 1 Score, t+1	6312	3.21	4954	3.11	5629	3.25
TEACH 2 Score, t+1	6312	3.20	4954	3.09	5629	3.24
TEACH 3 Score, t+1	6312	2.96	4954	2.85	5629	3.01
TEACH 4 Score, t+1	6312	3.25	4954	3.15	5629	3.30
TEACH 5 Score, t+1	6312	3.21	4954	3.11	5629	3.26
TEACH 6 Score, t+1	6090	3.09	4803	2.99	5430	3.14
TEACH 7 Score, t+1	6312	2.78	4954	2.67	5629	2.84
TEACH 8 Score, t+1	6312	3.25	4954	3.15	5629	3.30
TEACH 9 Score, t+1	6312	3.37	4954	3.28	5629	3.41
Female Teacher	8063	0.71	6089	0.71	6630	0.74
Teacher Gender Missing	8063	0.04	6089	0.04	6630	0.03
Black Teacher	8063	0.52	6089	0.54	6630	0.52
White Teacher	8063	0.31	6089	0.29	6630	0.33
Teacher Race Missing	8063	0.11	6089	0.11	6630	0.09
Graduate Degree	8063	0.61	6089	0.60	6630	0.68
Graduate Degree Missing	8063	0.09	6089	0.08	6630	0.05
Years of Experience	8063	10.5	6089	10.2	6630	10.5
Experience Missing	8063	0.02	6089	0.01	6630	0.01
Group 1 Teacher	8063	0.17	6089	0.18	6630	0.15

Table 1.4. TEACH Standard Scores, AY 2010-11 to 2012-13

	2010-11		2011-12		2012-13	
	Mean	*SD*	*Mean*	*SD*	*Mean*	*SD*
TLF Average	3.02	0.49	3.11	0.46	3.16	0.43
TLF Admin	3.07	0.55	3.18	0.51	3.24	0.48
TLF Master Ed	2.94	0.54	3.00	0.51	3.07	0.48
TLF 1	3.03	0.53	3.13	0.51	3.25	0.50
TLF 2	3.06	0.55	3.14	0.52	3.2	0.50
Teach 3	2.84	0.58	2.93	0.54	3.00	0.51
Teach 4	3.14	0.58	3.24	0.53	3.21	0.51
Teach 5	3.06	0.54	3.16	0.50	3.28	0.47
Teach 6	2.99	0.61	3.06	0.61	3.09	0.50
Teach 7	2.65	0.61	2.74	0.58	2.78	0.57
Teach 8	3.12	0.59	3.21	0.56	3.26	0.55
Teach 9	3.25	0.53	3.33	0.50	3.39	0.50
N	2693		2669		2582	

Notes: Sample includes all general teachers in each year. (which differs from the analytic sample that only includes teachers who return in t+1).

Table 1.5. Within-teacher Differences on TLF, *t* to *t+1*

	2010-11 to 2011-12		2011-12 to 2012-13	
	Mean	*SD*	*Mean*	*SD*
	(1)	(2)	(3)	(4)
TLF Average	0.054	0.35	0.051	0.34
TLF 1	0.075	0.48	0.106	0.47
TLF 2	0.045	0.48	0.067	0.47
TLF 3	0.045	0.50	0.057	0.50
TLF 4	0.041	0.48	-0.036	0.50
TLF 5	0.063	0.47	0.123	0.47
TLF 6	0.035	0.64	0.036	0.57
TLF 7	0.071	0.53	0.029	0.55
TLF 8	0.055	0.53	0.032	0.52
TLF 9	0.049	0.46	0.055	0.46
N	2070		2084	

Notes: Table is based on general education sample; teachers who are employed in teaching in year t and persist in year t+1.

Table 1.6. Reduced-form RD Estimates at Minimally Effective Performance Threshold, Overall TLF

	All years		2010-11		2011-12		2012-13	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
TLF	0.091	0.015	-0.097	-0.276**	0.210*	0.347**	0.332**	0.162
	(0.062)	(0.087)	(0.092)	(0.125)	(0.109)	(0.156)	(0.131)	(0.179)
TLF Mast. Ed.	0.025	-0.052	0.007	-0.117	0.185	0.363**	0.390***	0.293
	(0.066)	(0.093)	(0.102)	(0.143)	(0.119)	(0.173)	(0.132)	(0.186)
TLF Admin	0.025	-0.052	-0.149	-0.335**	0.194*	0.283*	0.193	0.007
	(0.066)	(0.093)	(0.093)	(0.131)	(0.117)	(0.159)	(0.139)	(0.187)
Linear spline	X		X		X		X	
Quadratic spl.		X		X		X		X

Notes: Each cell reports the results of a separate regression with the indicated dependent variable. Results condition on a smooth function of centered initial IMPACT score, specifications with a linear spline and quadratic spline are both shown above. Results also condition on teacher covariates and school fixed effects. Robust standard errors in parentheses. ***p<0.01, **p<0.05, *p<0.1

Table 1.7. Reduced-form RD Estimates at Minimally Effective Performance Threshold, Teach Standards

	All years		2010-11		2011-12		2012-13	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
TLF 1	0.054	-0.035	-0.064	-0.234	0.105	0.216	0.207	0.036
	(0.070)	(0.099)	(0.113)	(0.161)	(0.124)	(0.172)	(0.136)	(0.191)
TLF 2	0.128*	0.056	-0.032	-0.087	0.143	0.133	0.466***	0.344*
	(0.072)	(0.103)	(0.117)	(0.164)	(0.128)	(0.190)	(0.136)	(0.191)
TLF 3	0.130*	0.102	0.036	-0.031	0.260**	0.341**	0.218	0.215
	(0.068)	(0.097)	(0.109)	(0.157)	(0.122)	(0.172)	(0.136)	(0.193)
TLF 4	0.173**	0.175*	0.134	0.002	0.175	0.340*	0.289**	0.306*
	(0.068)	(0.097)	(0.109)	(0.153)	(0.125)	(0.183)	(0.126)	(0.177)
TLF 5	0.212***	0.194*	-0.005	-0.047	0.204*	0.424**	0.626***	0.470**
	(0.069)	(0.101)	(0.109)	(0.163)	(0.121)	(0.173)	(0.133)	(0.191)
TLF 6	0.198**	0.163	0.014	-0.041	0.206	0.322	0.439***	0.433*
	(0.084)	(0.124)	(0.142)	(0.214)	(0.159)	(0.227)	(0.160)	(0.233)
TLF 7	0.060	0.043	-0.053	-0.111	0.016	0.199	0.363***	0.219
	(0.067)	(0.096)	(0.108)	(0.157)	(0.119)	(0.168)	(0.138)	(0.194)
TLF 8	0.051	-0.000	0.033	-0.135	0.146	0.343**	-0.063	-0.074
	(0.070)	(0.100)	(0.110)	(0.153)	(0.124)	(0.173)	(0.149)	(0.213)
TLF 9	0.056	0.011	-0.008	-0.146	0.106	0.313*	0.130	0.058
	(0.070)	(0.097)	(0.112)	(0.160)	(0.119)	(0.160)	(0.145)	(0.202)
Linear spline	X		X		X		X	
Quadratic spl.		X		X		X		X

Notes: Each cell reports the results of a separate regression with the indicated dependent variable (ratings assigned by Master Educators only). Results condition on a smooth function of centered initial IMPACT score, specifications with a linear spline and quadratic spline are both shown. Results also condition on teacher covariates and school fixed effects. Robust standard errors in parentheses. ***p<0.01, **p<0.05, *p<0.1

Table 1.8 RD Regressions by Alternate Bandwidth, Minimally Effective Performance Threshold

		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
		n	TLF ME	TLF 1	TLF 2	TLF 3	TLF 4	TLF 5	TLF 6	TLF 7	TLF 8	TLF 9
All years	$\|s_i\|\le 60$	2,876	0.134*	0.014	0.144*	0.108	0.196**	0.227***	0.172*	0.059	0.017	0.030
			(0.076)	(0.080)	(0.083)	(0.078)	(0.079)	(0.081)	(0.098)	(0.076)	(0.082)	(0.080)
	$\|s_i\|\le 50$	2,294	0.078	-0.064	0.074	0.054	0.149*	0.171*	0.104	0.027	0.024	0.010
			(0.083)	(0.086)	(0.090)	(0.084)	(0.086)	(0.089)	(0.108)	(0.083)	(0.088)	(0.087)
	$\|s_i\|\le 40$	1,739	0.082	-0.051	0.032	0.043	0.115	0.155	0.163	0.036	0.079	0.015
			(0.092)	(0.095)	(0.099)	(0.093)	(0.095)	(0.098)	(0.121)	(0.091)	(0.098)	(0.096)
	$\|s_i\|\le 30$	1,235	0.100	-0.034	0.094	0.031	0.108	0.175	0.066	0.106	0.049	0.039
			(0.105)	(0.108)	(0.113)	(0.107)	(0.109)	(0.113)	(0.142)	(0.105)	(0.113)	(0.109)
2011-12	$\|s_i\|\le 60$	983	0.279**	0.220	0.174	0.256*	0.297**	0.293**	0.272	0.147	0.206	0.162
			(0.134)	(0.138)	(0.144)	(0.133)	(0.138)	(0.136)	(0.172)	(0.129)	(0.145)	(0.137)
	$\|s_i\|\le 50$	794	0.187	0.089	0.068	0.189	0.201	0.227	0.096	0.068	0.204	0.157
			(0.145)	(0.146)	(0.156)	(0.144)	(0.149)	(0.150)	(0.186)	(0.142)	(0.155)	(0.143)
	$\|s_i\|\le 40$	602	0.159	0.051	-0.057	0.122	0.096	0.221	0.136	0.056	0.324*	0.184
			(0.162)	(0.159)	(0.173)	(0.158)	(0.164)	(0.165)	(0.212)	(0.157)	(0.176)	(0.159)
	$\|s_i\|\le 30$	432	0.165	0.046	-0.047	0.035	-0.006	0.213	0.106	0.160	0.352*	0.259
			(0.193)	(0.188)	(0.208)	(0.183)	(0.193)	(0.192)	(0.256)	(0.183)	(0.206)	(0.182)
2012-13	$\|s_i\|\le 60$	815	0.296**	0.106	0.481***	0.158	0.317**	0.562***	0.326*	0.225	-0.188	0.020
			(0.144)	(0.146)	(0.150)	(0.145)	(0.144)	(0.159)	(0.171)	(0.145)	(0.158)	(0.157)
	$\|s_i\|\le 50$	625	0.201	0.013	0.288*	0.065	0.242	0.414**	0.201	0.178	-0.104	0.049
			(0.160)	(0.163)	(0.162)	(0.161)	(0.161)	(0.172)	(0.190)	(0.160)	(0.171)	(0.172)
	$\|s_i\|\le 40$	470	0.310*	0.065	0.374**	0.202	0.315*	0.447**	0.381*	0.217	0.022	0.058
			(0.183)	(0.184)	(0.183)	(0.182)	(0.182)	(0.199)	(0.218)	(0.179)	(0.193)	(0.195)
	$\|s_i\|\le 30$	326	0.207	-0.114	0.334*	0.168	0.238	0.417*	0.195	0.099	-0.019	0.007
			(0.203)	(0.206)	(0.200)	(0.204)	(0.211)	(0.222)	(0.246)	(0.197)	(0.223)	(0.216)

Notes: Each cell reports the results of a separate regression with the indicated dependent variable (ratings assigned by Master Educators only). Results condition on a linear spline of the assignment variable and teacher covariates. Robust standard errors in parentheses. ***p<0.01, **p<0.05, *p<0.1

Table 1.9. Reduced-form RD Estimates at Highly Effective Performance Threshold, Overall TLF

	All years		2010-11		2011-12		2012-13	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
TLF	0.118***	0.002	0.253***	0.099	-0.048	-0.056	0.103	-0.069
	(0.044)	(0.060)	(0.061)	(0.084)	(0.097)	(0.126)	(0.083)	(0.121)
TLF Mast. Ed.	0.084	-0.035	0.193**	0.028	-0.056	-0.047	0.111	-0.048
	(0.052)	(0.072)	(0.077)	(0.105)	(0.115)	(0.149)	(0.100)	(0.142)
TLF Admin	0.106**	0.024	0.250***	0.136*	-0.044	-0.059	0.051	-0.095
	(0.043)	(0.060)	(0.057)	(0.077)	(0.100)	(0.134)	(0.082)	(0.121)
Linear spline	X		X		X		X	
Quadratic spline		X		X		X		X

Notes: Each cell reports the results of a separate regression with the indicated dependent variable (ratings assigned by Master Educators only). Results condition on a smooth function of centered initial IMPACT score, specifications with a linear spline and quadratic spline are both shown above. Results also condition on teacher covariates and school fixed effects. Robust standard errors in parentheses. ***p<0.01, **p<0.05, *p<0.1

Table 1.10. Reduced-form RD Estimates at Highly Effective Performance Threshold, TEACH Standards

	All years		2010-11		2011-12		2012-13	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
TLF 1	0.147**	0.017	0.321***	0.085	0.003	-0.030	0.115	0.089
	(0.063)	(0.087)	(0.091)	(0.127)	(0.139)	(0.180)	(0.123)	(0.179)
TLF 2	0.108*	-0.046	0.093	-0.087	-0.002	-0.100	0.235*	0.052
	(0.062)	(0.086)	(0.088)	(0.120)	(0.143)	(0.186)	(0.125)	(0.183)
TLF 3	0.041	-0.050	0.160*	0.020	-0.074	-0.049	0.049	-0.033
	(0.063)	(0.087)	(0.094)	(0.129)	(0.133)	(0.187)	(0.121)	(0.172)
TLF 4	0.046	-0.015	0.188**	0.047	-0.259**	-0.257	0.108	0.062
	(0.058)	(0.079)	(0.083)	(0.113)	(0.120)	(0.157)	(0.123)	(0.174)
TLF 5	-0.014	-0.119	0.095	-0.001	-0.154	-0.174	-0.027	-0.191
	(0.063)	(0.088)	(0.090)	(0.125)	(0.140)	(0.182)	(0.120)	(0.171)
TLF 6	-0.003	-0.163*	-0.007	-0.193	-0.281	-0.500**	0.077	-0.095
	(0.071)	(0.097)	(0.107)	(0.146)	(0.199)	(0.251)	(0.124)	(0.170)
TLF 7	0.099	0.035	0.180*	0.052	0.139	0.315*	0.027	-0.193
	(0.062)	(0.087)	(0.095)	(0.133)	(0.134)	(0.175)	(0.114)	(0.158)
TLF 8	0.108*	0.038	0.215***	0.108	-0.064	-0.082	0.185*	0.132
	(0.055)	(0.076)	(0.081)	(0.110)	(0.132)	(0.172)	(0.106)	(0.150)
TLF 9	0.089	0.069	0.149*	0.155	0.202	0.263	0.026	-0.063
	(0.057)	(0.077)	(0.081)	(0.108)	(0.129)	(0.167)	(0.116)	(0.166)
Linear spline	X		X		X		X	
Quadratic spl.		X		X		X		X

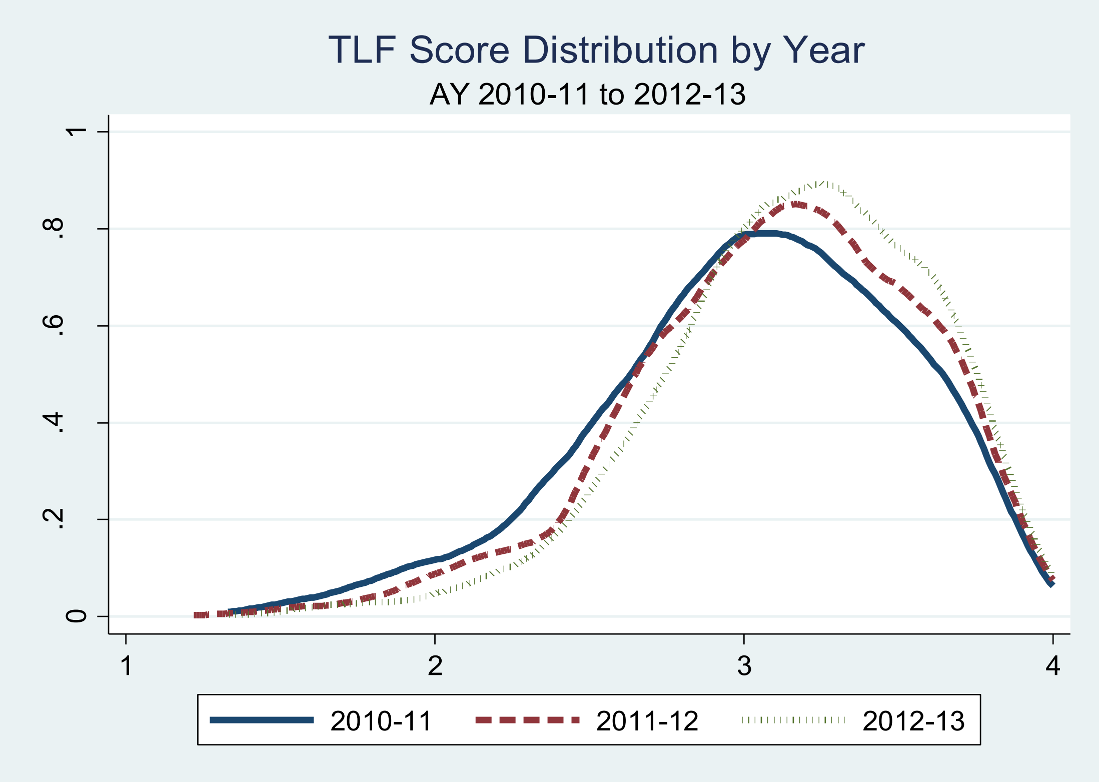
Notes: Each cell reports the results of a separate regression with the indicated dependent variable (ratings assigned by Master Educators only). Results condition on a smooth function of centered initial IMPACT score, specifications with a linear spline and quadratic spline are both shown above. Results also condition on teacher covariates and school fixed effects. Robust standard errors in parentheses. ***p<0.01, **p<0.05, *p<0.1

Table 1.11 RD Regressions by Alternate Bandwidth, Highly Effective Performance Threshold

		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
		n	TLF ME	TLF 1	TLF 2	TLF 3	TLF 4	TLF 5	TLF 6	TLF 7	TLF 8	TLF 9
All years	$\|s_i\|\le 60$	3,708	0.059	0.124*	0.085	-0.005	0.045	-0.016	-0.049	0.068	0.093	0.091
			(0.054)	(0.064)	(0.063)	(0.064)	(0.059)	(0.065)	(0.074)	(0.064)	(0.059)	(0.059)
	$\|s_i\|\le 50$	3,207	0.070	0.137**	0.090	0.021	0.061	-0.008	-0.030	0.076	0.097	0.075
			(0.055)	(0.066)	(0.065)	(0.066)	(0.061)	(0.067)	(0.076)	(0.066)	(0.061)	(0.061)
	$\|s_i\|\le 40$	2,635	0.050	0.118*	0.039	0.023	0.017	-0.046	-0.044	0.076	0.108*	0.076
			(0.058)	(0.068)	(0.068)	(0.069)	(0.064)	(0.069)	(0.080)	(0.069)	(0.064)	(0.063)
	$\|s_i\|\le 30$	2,067	-0.001	0.082	-0.019	0.009	-0.013	-0.111	-0.109	0.029	0.058	0.057
			(0.065)	(0.075)	(0.075)	(0.077)	(0.071)	(0.077)	(0.089)	(0.077)	(0.071)	(0.071)
	$\|s_i\|\le 20$	1,399	-0.052	0.022	-0.062	-0.065	-0.064	-0.170*	-0.150	-0.039	0.080	0.090
			(0.079)	(0.092)	(0.091)	(0.093)	(0.087)	(0.093)	(0.110)	(0.092)	(0.086)	(0.085)
2010-11	$\|s_i\|\le 60$	1,357	0.106	0.188**	0.076	0.061	0.090	0.097	-0.057	0.044	0.165*	0.119
			(0.079)	(0.092)	(0.089)	(0.095)	(0.082)	(0.093)	(0.109)	(0.096)	(0.084)	(0.084)
	$\|s_i\|\le 50$	1,187	0.103	0.183*	0.040	0.077	0.080	0.087	-0.071	0.051	0.182**	0.122
			(0.081)	(0.095)	(0.093)	(0.098)	(0.085)	(0.096)	(0.113)	(0.099)	(0.087)	(0.087)
	$\|s_i\|\le 40$	997	0.098	0.169*	0.024	0.105	0.045	0.045	-0.092	0.067	0.206**	0.112
			(0.085)	(0.098)	(0.098)	(0.103)	(0.089)	(0.100)	(0.119)	(0.104)	(0.092)	(0.093)
	$\|s_i\|\le 30$	795	0.003	0.086	-0.078	0.035	-0.028	-0.034	-0.171	-0.031	0.120	0.065
			(0.095)	(0.108)	(0.109)	(0.115)	(0.098)	(0.111)	(0.130)	(0.115)	(0.101)	(0.104)
	$\|s_i\|\le 20$	532	-0.062	0.027	-0.139	-0.012	-0.031	-0.192	-0.288*	-0.150	0.135	0.138
			(0.117)	(0.136)	(0.134)	(0.140)	(0.121)	(0.135)	(0.157)	(0.142)	(0.127)	(0.127)

Notes: Each cell reports the results of a separate regression with the indicated dependent variable (ratings assigned by Master Educators only). Results condition on a linear spline of the assignment variable and teacher covariates. Robust standard errors in parentheses. ***p<0.01, **p<0.05, *p<0.1

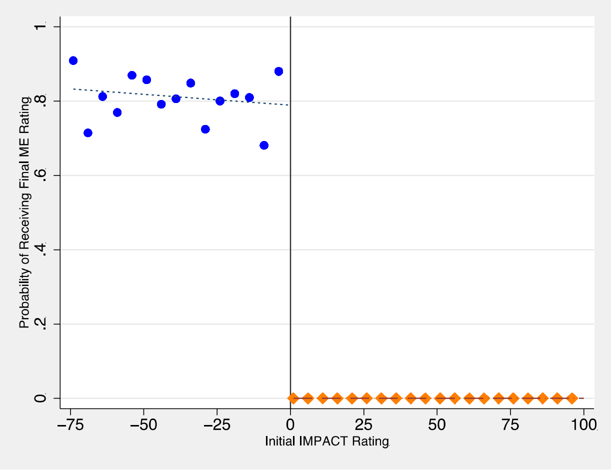
Figure 1.1. Overall TLF Score Distribution, AY 2010-11 to 2012-13



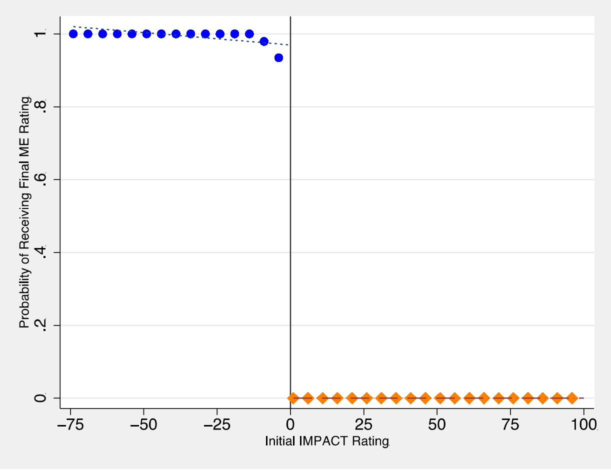
Notes: Sample includes all general teachers in each year. N=2,693 in 2010-11, N=2,669 in 2011-12, N=2,582 in 2012-13. This differs from the analytic sample that only includes teachers who return in t+1.

Figure 1.2. First stage, Intent-to-Treat at Minimally Effective Performance Threshold

Panel A. AY 2009-10



Panel B. AY 2010-11



Panel C. AY 2011-12

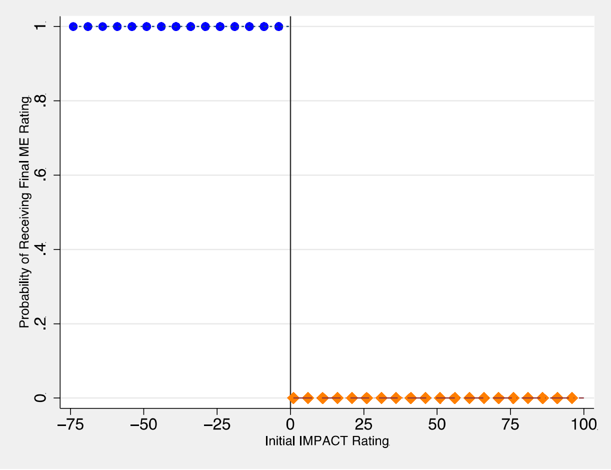
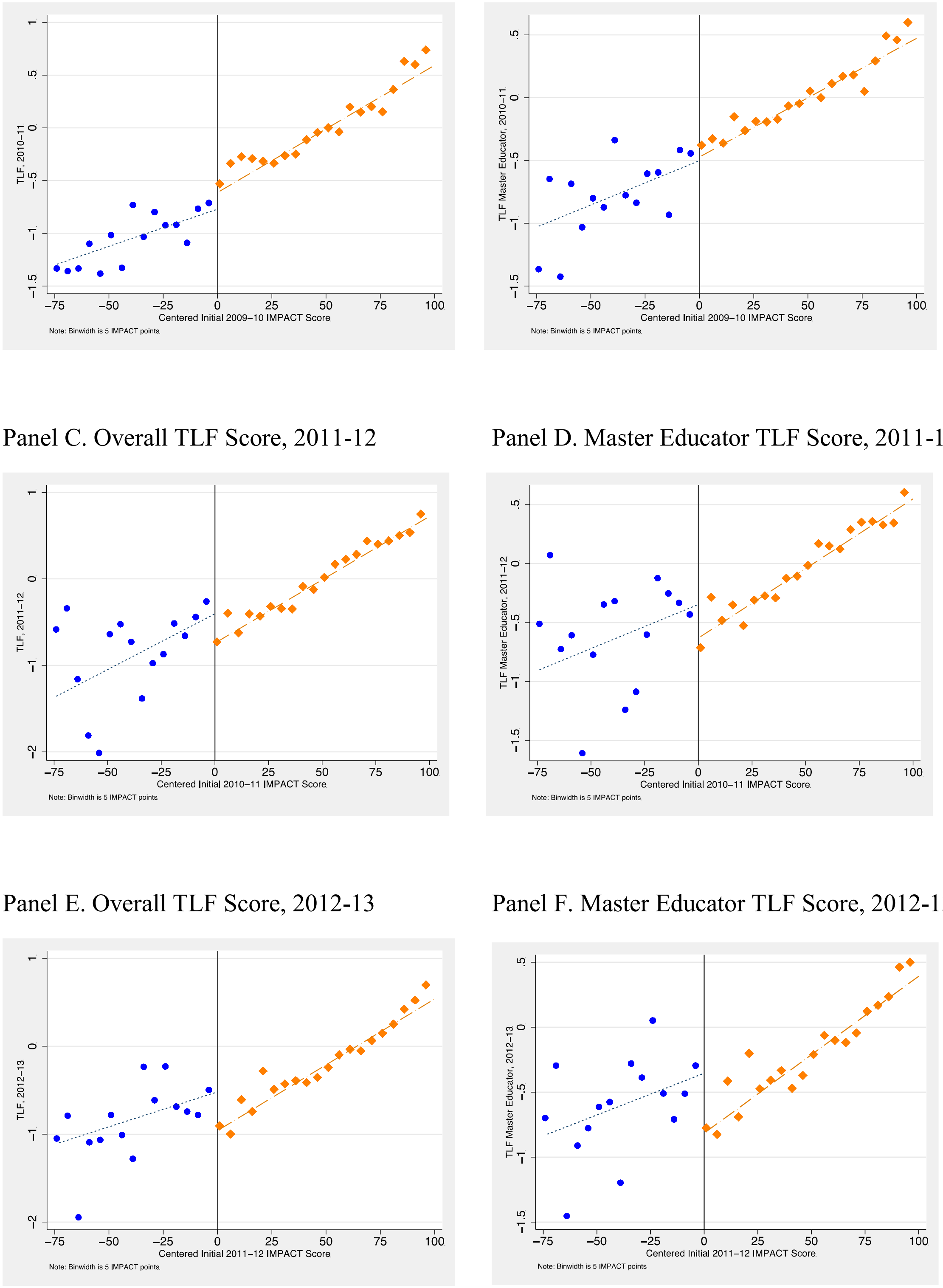


Figure 1.3. RD Graphs by Year, Minimally Effective Performance Threshold



Panel A. Overall TLF Score, 2010-11

Panel B. Master Educator TLF Score, 2010-11

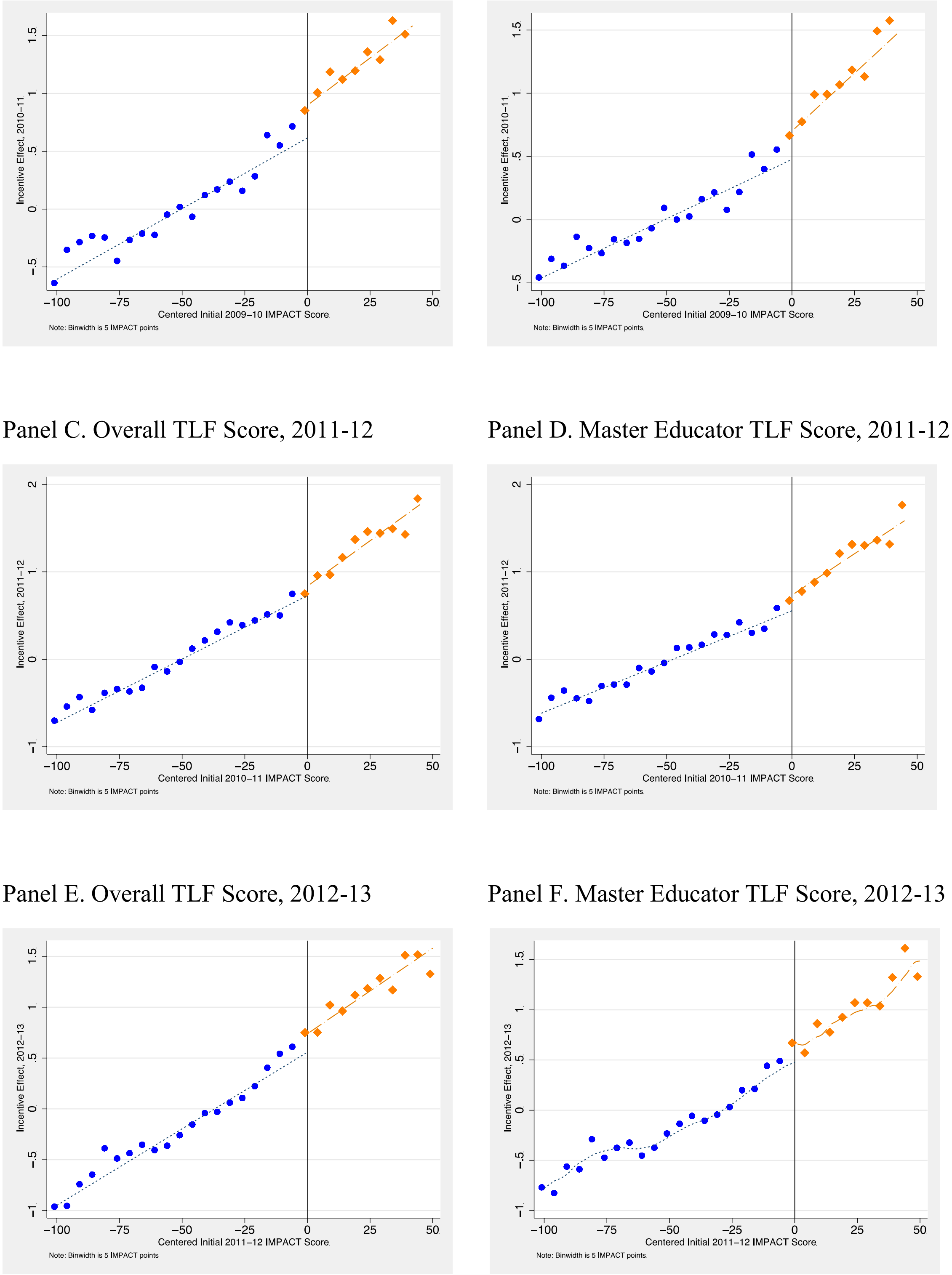
Panel C. Overall TLF Score, 2011-12

Panel D. Master Educator TLF Score, 2011-12

Panel E. Overall TLF Score, 2012-13

Panel F. Master Educator TLF Score, 2012-13

Figure 1.4. RD Graphs by Year, Highly Effective Performance Threshold



Panel A. Overall TLF Score, 2010-11

Panel B. Master Educator TLF Score, 2010-11

Panel C. Overall TLF Score, 2011-12

Panel D. Master Educator TLF Score, 2011-12

Panel E. Overall TLF Score, 2012-13

Panel F. Master Educator TLF Score, 2012-13

CHAPTER 2

Effects of Teacher Evaluation and Incentives on Student Achievement: Evidence from the District of Columbia

(with Thomas Dee, Veronica Katz, and James Wyckoff)

Abstract – The restructuring of teacher evaluation systems to include multiple measures of teacher performance, including teachers' contribution to student test score growth and classroom observation, has been one of the most important educational policy shifts this decade. Proponents suggest that these policies hold promise for improving teacher performance and student achievement because they incentivize both proximal measures of teaching and distal student achievement outcomes, and offer teachers formative feedback that may help them improve. This study presents causal evidence on the student achievement effects of the incentives embedded in IMPACT, the District of Columbia Public Schools' teacher evaluation system. In particular, we examine the effects of a strong incentive for low-performing teachers whose ratings place them near a threshold that implies a threat of dismissal if performance does not improve in the next year. Using a regression discontinuity design, we find no student achievement effects in 2010-11 for teachers rated Minimally Effective (ME) during 2009-10, the first year of the program. We do, however, observe positive effects of the ME rating of roughly seven percent of a standard deviation of student achievement in each of the next two years, driven by improvements in math, with a particularly large effect observed in math in 2012-13 (0.23 SD).

1. INTRODUCTION

Recent research has demonstrated what educators and parents have always known to be true: effective teaching is a critical determinant of student success. Teachers have a greater impact on student outcomes than any other school-based factor, and there are large differences among teachers in their ability to help students achieve (Aaronson, Barrow, & Sander, 2007; Kane & Staiger, 2008; Rivkin, Hanushek, & Kain, 2005; Rockoff, 2004). Moreover, assignment to a more effective teacher improves not only students' outcomes during school but also indicators of well-being in adulthood such as earning a higher salary, and being less likely to become pregnant as a teenager (Chetty, Friedman, & Rockoff, 2014). Despite this broad agreement regarding the importance of effective teaching, we have at best mixed evidence on how to design policies that drive improvements in teacher performance. More traditional forms of professional development (i.e., the "single-shot" workshop) often feel unhelpful to teachers (Darling-Hammond, et al. 2009), and largely do not lead to improvements in student achievement when rigorously evaluated at scale (Garet et al., 2008; Garet et al., 2011; Yoon, et al., 2007). While most performance improvement occurs during teachers' early career period (Atteberry, Loeb, & Wyckoff, 2013; Rockoff, 2004), systematic differences in teacher preparation (Boyd et al., 2009; Goldhaber, Liddle, & Theobald, 2012) and early career experiences such as mentoring (Glazerman, et al., 2010) explain little of the variation in individual performance improvement.

Within the last decade, a majority of states and districts have become interested in implementing complementary performance evaluation and incentive programs as a means of improving the overall performance of the teacher labor force. The proposed reforms, which rely on multi-faceted measures of job performance to determine teacher job security, pay and access leadership opportunities, have been highly controversial. The District of Columbia Public Schools (DCPS) was the first to implement such a system in the 2009-10 academic year. IMPACT links uniquely high-powered incentives, including the potential for performance-based dismissal, to multiple measures of teaching performance including frequent classroom observation, student test performance, and assessments of a teacher's contribution to the school community and professionalism. Because IMPACT was the earliest and is arguably the most comprehensive model of evaluation and compensation reform in the country, there has been a keen interest in evaluating its effectiveness, and particularly in determining the extent to which the system has had an impact on student academic outcomes.

A number of articles have described how student achievement has changed in DCPS since the introduction of IMPACT, but none have been able to provide causal evidence on this question. Recent analyses that track the academic progress of DC students on both the DC Comprehensive Assessment System (CAS) exams and the National Assessment of Educational Progress (NAEP) over the last ten to fifteen years find encouraging trends (Ozek, 2014; Washington Post, 2013), but it has been difficult to disentangle previous upward trajectories and concurrent policy shifts from the effects of IMPACT. In seminal work on IMPACT, Dee and Wyckoff leverage a natural experiment created by the design of IMPACT to determine the causal effect of the program's largest incentives on teacher retention and overall performance (2015). The regression discontinuity (RD) research design they use does not provide an overall estimate of IMPACT's effect, but rather delivers a highly internally-valid estimate of a specific program feature: the incentives associated with having just received a Minimally Effective (ME) or Highly Effective (HE) rating. These ratings are associated with particularly strong incentives in the next year: teachers rated ME must improve to an Effective rating or face dismissal, and teachers rated HE are eligible for a permanent base salary increase if they obtain a second consecutive HE rating. Dee and Wyckoff find that the dismissal threat associated with an ME rating increases voluntary attrition by 11 percentage points (an increase of more than 50 percent), and that teachers who remain in the district improve their overall performance by 0.24 of a teacher-level standard deviation (SD). Teachers who receive an HE rating also improve their performance, by 0.27 teacher-level SD though these results were not robust to alternative specifications.

While the earlier Dee and Wyckoff paper focused on changes in teacher retention and overall performance, this analysis explicitly addresses the effect of IMPACT's incentives on student achievement outcomes on the DC CAS exam. Conceptually, improvements in an overall measure of teacher performance that is based on multiple measures of performance should correspond with improvements in student outcomes as well, if the performance measures (e.g., classroom observation) capture aspects of teacher performance that relate to teachers' ability to improve student test outcomes.[33] However, if improvements largely occur on process measures of teaching that do not lead to improved achievement outcomes for students, student achievement gains could be small or none. We explore this empirically by employing individual student-level test

[33] The teacher performance measure that most narrowly captures a teacher's contributions to student test score gains is the "value-added" score, which statistically isolates the teacher's effect on the student test scores themselves. Even for this measure of teacher performance that aligns most closely to student test performance, a one SD in teacher value-added is associated with a 0.13 SD in student test performance in reading and 0.17 SD in math (Hanushek & Rivkin, 2010). For teacher performance measures that also capture other dimensions of performance that are not related to test performance, we would expect this relationship to be weaker. Thus, the student achievement effects of teacher policies are much smaller than the teacher performance effects.

performance data differentiated by subject to examine the effects of high-powered teacher incentives within IMPACT on student achievement over the first three years of the program. This granular student-level data also provides the opportunity to explore the potentially heterogeneous effects that IMPACT incentives have on specific student subgroups. In this paper, we focus on the incentives associated with an ME rating because we unfortunately have insufficient sample size and statistical power to examine effects at the HE threshold.

Consistent with the previous study, we find no student achievement effects following the first year. As we discuss later, political instability in District politics following the first year of the program may have damaged IMPACT's credibility and contributed to these null effects. However, we do observe positive effects of approximately seven percent of a SD of student achievement in each of the next two years, though the positive result in 2011-12 is less robust in specification checks. Results in both years are driven largely by math, with a particularly large effect (0.23 SD) in 2012-13. This suggests, on average, a moderate positive effect of IMPACT's highest-powered incentive on students who are assigned to teachers who received a Minimally Effective rating. Moreover, exploring the heterogeneity in these effects by student subgroup suggests that traditionally underserved students—in particular, students who receive special education services—experience gains equal to or larger than their peers in most years. These improvements in student achievement in 2011-12 and 2012-13 coincide with improvements in teaching practice in the same years (Adnot, Chapter 1), which provides some support for a the theory that it is improvements in teaching that are leading to the concurrent achievement gains. However, we cannot determine this with certainty and we discuss this and other limitations of the study, as well as its implications for policy and practice, in the discussion.

2. THEORY AND BACKGROUND LITERATURE

Proponents of teacher evaluation propose that measuring teacher performance and tying personnel decisions to the results has the potential to improve the performance of the current teacher workforce in two primary ways: (1) by motivating current teachers to expend additional effort on incentivized tasks and on developing new evaluation-relevant skills, and (2) by changing the composition of the teaching workforce by encouraging the retention of high-performers and the attrition of low-performers (Donaldson & Papay, 2015; Hanushek & Lindseth, 2009). This analysis captures elements of both compositional and improvement effects. Dee and Wyckoff (2015) document meaningful compositional effects of IMPACT incentives for low-performing teachers who face the threat of dismissal, and explore these effects at length in their original paper on IMPACT. We also examine how compositional changes to the workforce influence student achievement during the IMPACT era in a companion paper (Adnot, Dee, Katz & Wyckoff, 2016). As such, while we acknowledge that compositional effects influence results in the current study as well, we focus here on the theory and background literature on teachers' response to incentives and skill development in the context of evaluation.

2.1 Teacher response to incentives. Personnel economics provides a well-developed theory describing the effects of incentives on individuals' motivation to invest in incentivized tasks. This theory holds that incentives work best in contexts where (1) measured performance can be reliably increased when employees increase or redirect effort, and (2) performance is measured with little noise (Lazear & Oyer, 2009). If there are discrepancies between either the investment of effort and measured performance or between measured performance and desired outcomes, the theorized effects of the incentives weaken as the expected marginal return to effort falls.

The most salient example of incentives in education has been the use of "pay-for-performance" programs that offer teachers financial bonuses for improving student test scores. While merit-based compensations schemes are certainly not new (e.g., Murnane & Cohen, 1986), there has recently been a resurgent interest in understanding the potential of these programs to motivate teachers to direct effort towards the improvement of student test outcomes. The results of recent experimental evaluations of pay-for-performance programs have been mixed, but a substantial number find largely null effects (Fryer, 2013; Springer et al., 2010; Springer et al., 2012). For example, The Project on Incentives in Teaching (POINT) was a three-year study that provided randomly assigned middle-school mathematics teachers in Nashville individual bonuses of as much as $15,000 if their students met ambitious performance thresholds (Springer et al., 2010). The availability of these incentives led to no detectable effects on student performance. Lower incentives amounts in New York City (up to $3,000) and Texas (up to $6,000) also had no detectable effects on student performance (Fryer, 2013; Marsh et al., 2011; Springer et al., 2012). Moreover, in follow-up surveys teachers report that that they did not find the incentive pay programs motivating, and did not change their instructional practice or hours worked in response to the opportunity for additional pay (Yuan et al., 2013).

These results raise considerable doubt regarding the efficacy of incentives for motivating teachers to focus on improving student achievement scores. The study authors examine a number of reasons that these bonus programs may have failed, which, in line with incentive theory, involve either a failure to motivate the expenditure of additional effort, or a disconnect between increased effort and increases in student achievement. For instance, teachers may not increase the effort they expend on improving student achievement outcomes because they are already be putting forth maximum effort so there is no marginal benefit to offering additional pay. Or they may perceive value-added measures to be so noisy that they can't be improved with increased effort.[34] Alternatively, teachers may put forth more effort, but this effort may not be well-targeted if it is not clear how best to help students achieve higher test scores. Study authors also note important limitations in program design that could have diminished effects including the short-term nature of the pilot programs and small incentive amounts.

Two experiments that employ different incentive designs and find positive effects provide insight into program features that may make incentive policies more effective. In a 9-school experiment in Chicago, researchers tested the effectiveness of three alternately-framed incentive schemes: one treatment arm offered a conventional bonus (up to $8,000) to individuals, one arm offered the bonuses to groups, and one arm awarded the bonus to teachers in advance and required them to return the money if students did not improve (Fryer et al., 2012). Only the "loss-oriented" treatment resulted in gains, which were quite large (i.e., 0.201 to 0.398 student achievement SD) and stand in contrast to small and statistically insignificant results for other treatment arms. The possibility of dismissal following a Minimally Effective rating in IMPACT shares this

[34] A quasi-experimental study of incentives in Texas, which also finds no effect of a bonus program, suggests this may be the case: they find that teachers who are close to the bonus threshold in one year are do not produce higher student achievement gains in the next year (Brehm, Imberman, Lovenheim 2015).

"loss aversion" feature, but could be considered an even stronger incentive as the value of job loss is significantly greater than the bonus amount that could potentially be "lost" in the Chicago experiment.

A second experiment evaluating the federally-funded Teacher Incentive Fund (TIF) grants provides the opportunity to investigate the potential of financial incentives in the context of a broader human capital system that prioritized the development and retention of effective of teachers. There have been relatively few opportunities to observe incentives within a more comprehensive system, though study authors noted that achieving full program implementation (which entails multiple measure evaluation, additional pay opportunities and professional development) was a challenge with only one-half of grantee districts implementing all four program components (Chiang et al., 2015). Despite incomplete implementation and challenges communicating with teachers, researchers found that offering pay-for-performance bonuses improved student reading achievement by one percentile point, with a similar but statistically undetectable effect in math. In contrast to the previous null results, this study offers some indication that more robust incentive programs that pair incentives with formative feedback may be better situated to improve student achievement outcomes, though the effects detected in this case are quite small.

2.2 Teacher evaluation. Multi-faceted teacher evaluation systems are increasingly thought of as a critically important component of human capital reform in education. While financial incentives can theoretically increase motivation to perform well on measured tasks, they provide no information on how to do so, and evaluation systems with a robust observation and feedback component have the potential to provide this formatively useful information. Research has not coalesced around an ideal set of teacher performance measures, but a growing body of evidence points to the importance of using multiple measures of teaching performance such as classroom observations, measures of teacher contribution to student test score gains (e.g., value added measures), and input from students and peers to provide a more balanced and stable assessment of performance (Donaldson & Papay, 2015; Goe & Croft, 2009; Mihaly et al., 2013).

Classroom observations have been nearly universally included in revised evaluation programs, and provide an important opportunity for the provision of performance feedback (Goldring et al, 2015; Papay, 2012; Pianta, 2012). They are generally well-accepted by teachers as a valid tool for evaluation, and the clearly articulated descriptions of practice offer practical formative value. Recent studies of the reliability and validity of these measures recommend multiple observations (at least four) and multiple observers (at least two) to insure acceptable levels of reliability (Ho & Kane, 2013). There is still much to be learned about statistical properties of classroom observation measures (Gill, Bruch, & Booker, 2013; Goldhaber 2015), though recent work cautions that they can be significantly correlated to student characteristics (Garrett & Steinberg, 2015).

“Value added” measures (VAMs) employ statistical modeling to separate a teachers’ effect on student test scores from other observable student characteristics that also affect achievement. VAMs have attracted considerable attention, and are something of a lightning rod in discussions of evaluation, particularly in light of the recent introduction of new Common Core State Standard aligned exams. Critics of VAMs argue that they are too biased (due to student sorting among teachers, and test ceiling and floor effects,) and noisy (due to small sample sizes) to be useful in an individual evaluation (Rothstein, 2008; 2009; Baker et al. 2011; Darling-Hammond, 2015). Others contend that in well-specified models there is little evidence of bias (Chetty, Friedman, & Rockoff, 2014; Kane and Staiger, 2008) and that in practice these issues are small and are outweighed by other desirable features of VAMs (Glazerman et al., 2011; Goldhaber, 2015; Koedel & Betts, 2011; Whitehurst, Chingos, Lindquist, 2014).

Rigorous evaluations of observation-focused evaluation programs have found encouraging effects of these programs on student achievement, but the evidence base is still limited given the novelty of these reforms. Experimental and quasi-experimental work in Chicago and Cincinnati, respectively, find that being observed and receiving feedback improves student achievement in 10 percent in math in Cincinnati and 9.9 percent in reading in Chicago, even in the absence of strong performance incentives (Taylor & Tyler, 2012; Sartain & Steinberg, 2015). In Cincinnati, the authors suggest that the gains they observe represent an investment in the teacher's human capital (i.e., skill building) due to the continued effects they observe in the years after the teacher was evaluated and the low-stakes nature of the evaluation. But, they note that motivational effects of incentives are also at play in most evaluations, and that "these mechanisms are not mutually exclusive, which complicates any investigation of the effect of evaluation on performance" (p. 1, Taylor & Tyler, 2012). As described below, IMPACT includes both of these mechanisms, with incentives that are designed to focus teachers on evaluation-relevant tasks and formative feedback from the classroom observation process that may help to build new skills.

3. DESCRIPTION OF IMPACT

The performance assessment of DCPS teachers under IMPACT began in the 2009-10 school year. The structure of these assessments is fairly straightforward: at the conclusion of each academic year, teachers are provided with a single score that summarizes their performance across multiple measures for the academic year. For teachers of math and reading in grades 4 through 8 (i.e., the focus of this study), a major component of this overall score is “Individual Value Added” (IVA). The IVA measure captures a teacher’s estimated contribution to the achievement growth of their students as measured on the DC Comprehensive Assessment System (CAS) tests and conditional on student and peer traits. These so-called “Group 1” teachers are those for whom the available CAS data allow for the estimation of value added (i.e., only reading and math teachers in grades 4 through 8).

A second major component of the overall IMPACT score for these teachers reflects rigorously scored classroom observations tied to the district’s Teaching and Learning Framework (TLF). The TLF specifies the criteria by which DCPS defines effective instruction and structures a scoring rubric. TLF includes multiple domains such as leading well-organized, objective-driven lessons, checking for student understanding, explaining content clearly, and maximizing instructional time. A teacher’s TLF score is typically based on five formal observations with only the first observation announced in advance. Teachers for whom IVA cannot be calculated instead receive a Teacher-Assessed Student Achievement (TAS) score. At the beginning of each academic year, teachers choose (and administrators approve) learning goals based on non-CAS assessments. At the end of the year, administrators rate the teacher’s success in meeting these goals using a rubric that emphasizes student learning or content mastery. All teachers are also assessed by their administrators on a rubric that measures their support of school initiatives, efforts to promote high expectations, and partnerships with students' families and school colleagues: the Commitment to the School Community (CSC) measure. Teachers also received a score based on their school's estimated value added on the CAS tests (SVA). Finally, principals assess each teacher on their "Core Professionalism" (CP). The rubric for CP rates teachers on the basis of attendance, punctuality, policies and procedures and respect. Teachers are assumed to be professionals, and, therefore, CP scores can only reduce a teacher's overall IMPACT score.

During our study period, a Group 1 teacher's IVA score constituted 50% of their overall IMPACT score. Their TLF score constituted 35% of their total scores with the remainder consisting of CSC (10%) and SVA (5%) scores.[35] These summative IMPACT scores determine high-stakes outcomes for teachers. IMPACT scores allocated teachers to four performance categories during this period: Highly Effective (HE) teachers (scores of 350 or higher), Effective (E) teachers (scores from 250 to 349), Minimally Effective (ME) teachers (scores from 175 to 249) and Ineffective (I) teachers (scores below 175). Those teachers who were rated Ineffective were immediately dismissed. In contrast, under "IMPACT*plus*", DCPS also provided substantial financial rewards to teachers with HE ratings. Teachers who earn E ratings experience no unique consequences. However, teachers with an ME rating are subject to a clearly articulated dismissal threat: they are forcibly separated if their next rating is not E or HE. The dismissal threat implied by an ME rating is a uniquely high-powered incentive. We use the data and the RD designs we

[35] Under revisions that took effect for 2012-13, the weight placed on IVA scores has fallen from 50% to 35%.

describe below to identify the effect of this dismissal threat on teacher performance as measured by the test-based achievement of their students.

4. STUDENT AND TEACHER DATA

We base the analytical samples for our regression-discontinuity (RD) analyses on linked student-teacher administrative data from DCPS. Specifically, our analytical samples consist of three annual cross-sections of teachers who were rated under IMPACT in year t (i.e., the 2009-10, 2010-11, and 2011-12 academic years) and the students these teachers taught in the subsequent year t+1 (i.e., the 2010-11, 2011-12, and 2012-13 academic years[36]). We link the administrative data on these teachers and the students they taught by relying on the DCPS "rosters" created to construct value-added estimates.[37] Because our focus is on student-level test scores on the DC Comprehensive Assessment System (CAS) and on teachers whose evaluation includes those scores, our analytical sample reflects several other specific considerations.

First, we limit our sample to students in grades 4 through 8 so that we have students who participated in the DC CAS and for whom it is possible to condition on their prior year's performance. A second and closely related restriction is to include teachers in traditional K-12 public schools and who are in Group 1 in the baseline year. This insures consistency with our "intent to treat" (ITT) population of teachers who initially received an ME rating and were evaluated on their students' value-added performance on the CAS. Third, we also limit the sample to teachers whose initial IMPACT ratings in the baseline year fell in the ME or E performance bands. This allows

[36] Moving forward, we refer to these cross-sections by the outcome year, t+1.

[37] In these rosters, students can be linked to more than one teacher within a subject if multiple teachers have a role in the instruction that student receives in a given year. Prevalence of students linked to multiple teachers is 12% in math, 23% in reading in our data. When we limit the sample to students linked to only one teacher, our results are substantively unchanged.

our RD design to avoid the possible biases created by including observations around a distal threshold (e.g., the "highly effective" band) that may have relevance. However, we find that including these observations does not meaningfully change our results. Fourth, our sample only includes teachers who can be linked to the CAS scores of the students they teach in the next year. This implies that teachers in our intent-to-treat population are not in our analytical sample (i) if they no longer teach in DCPS classrooms or (ii) if they change assignments to teach in non-tested grades or subjects. The first source of attrition is not an internal-validity threat but rather an acknowledged component of IMPACT's theory of change (i.e., both selection and incentive effects are relevant). The performance of ME-rated teachers may improve through both the attrition of low-performing teachers and the improvement of those ME-rated teachers who choose to remain.[38] However, there would be an implied internal-validity threat in this setting if ME-rated teachers were differentially likely to switch to non-tested teaching assignments. We have examined auxiliary RD regressions to test this hypothesis and report these results along with a series of other robustness checks in the next section.

Our final linked samples yield 553 teacher-year observations linked to 17,959 student-year observations in math, and 579 teacher-year observations linked to 19,348 student-year observations in reading. Our teacher-level variables include their *initial* IMPACT score in each baseline year (i.e., the assignment variable in our RD design) as well as an indicator for whether their final IMPACT rating for that year was ME. Across

[38] Dee and Wyckoff (2015) also confront this interpretative issue. They find that "an ad hoc empirical decomposition based on our RD design also suggests incentive effects rather than selection effects. Using the sample of teachers who returned, we estimated an RD specification where IMPACT performance in the prior year is the dependent variable. We find small and statistically insignificant effects that are consistent with the hypothesis of behavioral change in response to the incentives." (2015. p. 21)

each subject and year, the share of students observed with a teacher who was initially rated ME ranges from 17 to 24 percent (Table 2.1). Linked human-resources data from DCPS also provide us with information on each teacher's race and gender as well as on whether they have a graduate degree. These data also allowed us to construct a measure of teacher experience (Table 2.1).

Our student-level CAS data files provide scale scores on math and reading achievement over our three study years, and also include information on each student's status with respect to free or reduced-price lunch eligibility, special education, English-language proficiency, race, and gender. To construct our math and reading outcome measures, we first standardized the student scale scores within subject-grade-year cells. For each subject, we then regressed these student-level scores on the student's prior year score in the same subject and binary indicators for race, gender, lunch status, special-education status, and English-proficiency status. The residuals from these regressions constitute our outcome measures. Thus, these residualized achievement outcomes are the difference between a student's actual performance and his predicted achievement based on prior achievement and the available, observed characteristics.

In Table 2.1, we show descriptive statistics for these data, separately by subject and year. Students in these samples are split evenly between elementary and middle-school grades. A substantial majority of these students are either black or Hispanic, eligible for free or reduced-price lunches, and attend high-poverty schools. Interestingly, roughly 2 to 5 percent of students are linked to teachers who are not officially classified as "Group 1" teachers though they are linked to students in tested subjects and grades in year t+1. This designation reflects a reporting requirement that, for IVA to be calculated (and for a teacher to be in Group 1), they must be linked to current and prior test-score data for 15 more students in the subject. For teachers in our ITT population whose students did not meet this requirement, their official designation was Group 2. We include these observations in our analysis because these teachers were likely to view themselves as Group 1 teachers. However, excluding them does not meaningfully change our results. It is also interesting to note that some of the observed traits of students change noticeably across subjects and years. For example, the percentage of students eligible for free/reduced-price lunches fell appreciably over the sample years. These year-to-year changes could reflect changes in the underlying population as well as in program participation or data-reporting practices. We do notice that there was also a 2011-12 increase in the number of students rostered to teachers on average (i.e., 6 to 9 students). In various checks associated with our RD design, we see some suggestion of imbalance in the number of rostered students that operates in opposing directions over the years of our study. This is explored in the next section in a series of robustness checks.

5. METHODOLOGY

We use a regression discontinuity (RD) design to estimate the causal effect of receiving a Minimally Effective (ME) rating on residualized achievement of a teacher's students in the next year. In general, well-functioning RD designs have a strong causal warrant because they leverage the credibly random assignment to "treatment" (here, an ME rating) of observations around an arbitrary threshold value (Campbell, 1969; Lee & Lemieux 2009). An important concern with any RD design is that there may be systematic sorting of teachers to either side of the threshold. For example, if teachers are able to manipulate the variable that "assigns" them to one side of the performance threshold or the other that this may bias results as there may be differences in performance between teachers who are able to sort over the threshold and those who are not. The process of appealing low ratings and obtaining revised, final IMPACT scores, for instance, would certainly introduce this type of bias into our estimates. To avoid this, we use teachers' *initial* IMPACT scores to determine assignment to treatment, and the estimates we detect can thus be interpreted as ITT estimates.

We present graphical evidence along with parametric and non-parametric estimates to illustrate the relationship of this assignment variable with both ME status and the test-performance of the teacher's students in the next year. Our core analyses are based on reduced-form specifications of the following general form:

$$Y_{ijsg(t+1)} = \alpha + \beta I\left(S_{jt} \leq 0\right) + f\left(S_{jt}\right) + \pi_{g(t+1)} + \delta_{s(t+1)} + \varepsilon_{ijsg(t+1)} \qquad (1)$$

where $Y_{ijsg(t+1)}$ represents residualized student achievement for student i with teacher j in school s and grade g in year t+1. The term $\varepsilon_{ijsg(t+1)}$, is a mean-zero error term. Because our student-level observations are nested within specific teachers, we construct our standard errors allowing for heteroscedasticity clustered at the teacher level. The variable, S_{jt}, is teacher j's initial IMPACT score from year t, centered on the value 249. When a teacher's value for this assignment variable is less than or equal to zero, they have an "intent to treat" as a ME teacher. The parameter, β, identifies the "jump" in the outcome variable at this threshold conditional on a smooth function of S_{jt} and fixed effects unique to each school and each grade (i.e., $\delta_{s(t+1)}$ and $\pi_{g(t+1)}$).

5.1 First Stage. We rely on the same basic RD specifications to model the "first stage" relationship between our intent-to-treat measure (i.e., $I(S_{jt}) \leq 0$) and whether a teacher's final status was ME or not. Across our studies, the jump in this treatment status at the threshold is either sharp or nearly so due to a modest number of successful teacher appeals. This implies that our estimates of β are either exactly or nearly equal to the effect of *having* ME status (i.e., the "treatment on treated" estimand). Figure 2.1 examines the strength of the relationship between receiving an initial ME rating and a final ME rating in each treatment year.[39] In 2009-10, four teachers in our sample successfully appealed their ME rating and ultimately received an Effective designation, which can be seen in Panel A. As the fitted line through the binned data on the ME side of the threshold implies, the probability of being assigned to a teacher who received a final ME rating, conditional on being assigned to a teacher who received an initial ME rating, is 81% in the reading sample and 95% in the math sample. The "fuzziness" of assignment to treatment in the first year does not represent an internal-validity threat, but does imply that the estimates from our main RD relate to the teachers who "complied" with their initial ME status in this year. In the second and third treatment years that we study, 2010-11 and 2011-12 (Panels B and C), there are no successful appeals out of the ME category in our sample. In these years, assignment to initial IMPACT rating perfectly predicts final IMPACT rating, yielding a "sharp" regression discontinuity.

5.2 Internal Validity Checks.[40] A key feature of this RD design is the promise of providing credible causal inference by leveraging the sharp incentive contrasts (i.e., a dismissal threat versus business as usual) for teachers whose initial IMPACT scores placed them on either side of the ME/E threshold. However, the causal warrant of RD designs like this turn on a number of important assumptions that need to be examined

[39] Corresponding parametric estimates of the first stage can be found in Appendix Table A2.1.

[40] The majority of this section will likely appear in an appendix when this paper is prepared for publication.

explicitly. The recent methodological literature has provided increasingly standardized guidance to checking the key threats to RD designs (Lee & Lemieux, 2009; Schochet et al., 2010). Our analysis tracks that guidance closely. For example, RD designs rely, sometimes critically, on the proper functional-form specification for the assignment variable in Equation (1). We examine the robustness of our main findings to such functional-form concerns in several ways. Our specification conditions on linear terms for the assignment variable but uses splines so that this slope can vary on either side of the threshold. However, we also explored controlling for a higher-order polynomial of the assignment variable (i.e., a quadratic) and report the Akaike Information Criteria (AIC) as one means of assessing the model fit. We also examine the robustness of our results across specifications that introduce school and grade fixed effects. Finally, we also report the results of straightforward local linear regressions that restrict our observations to those in an increasingly tight bandwidth around the threshold.

In addition to functional form considerations, an RD design would also be invalid in this context if teachers could systematically manipulate their initial IMPACT rating to place themselves on a particular side of the ME threshold. The institutional details surrounding the determination of IMPACT scores suggest that they are highly unlikely to be subject to such precise manipulations.[41] The overall IMPACT score consists of multiple components rated at different times by different sources (e.g., school administrators, master educators, and an independent research firm contracted to construct the value-added estimates). We also examined this question empirically through a density test (McCrary 2008) that tests the null hypothesis that the distribution of

[41] To be clear, the fact that teachers may generally endeavor to raise their scores does not constitute an internal-validity threat unless they can systematically manipulate their position relative to the threshold (Lee and Lemieux 2009).

observations in our ITT population is smooth around the threshold. We fail to reject this null hypothesis for each of our study years (i.e., the absolute values of test statistics range from 0.43 to 1.35). We also provide visual evidence of these densities in Figure A2.1. Another heuristic way to interrogate the validity of the RD design is to compare the balance of the *baseline* teacher covariates around the threshold. If our RD design leverages credibly quasi-random variation, we should find that the traits of teachers observed at baseline are balanced around the ME/E threshold. Table A2.2 reports the results of this auxiliary RD regression; there is no evidence of differences in observable teacher covariates across the ME/E threshold in year t. Taken together, the institutional details of IMPACT score assignment, and evidence of a smooth density function of initial IMPACT score and balanced teacher traits provide suggestive evidence that teachers were not able to manipulate their initial IMPACT score around the ME/E rating threshold.

Another type of internal-validity threat that is unique to our setting involves possible non-random sorting that occurs *after* the intent to treat. To be clear, the differential attrition of teachers from DCPS does *not* constitute such a threat. To the extent that teachers on the ME threshold perform better in period t+1, our theory of change accommodates this occurring through differential retention rather than only through the direct effect of the stark incentive contrasts. However, our design would have an internal-validity threat if teachers receiving the dismissal threat are more likely to seek different teaching assignments or different types of students (e.g., those with a higher propensity to achieve). We can examine the empirical relevance of such threats through auxiliary RD regressions of the post-treatment balance of teacher-student assignment and student traits around the ME/E threshold (Tables A2.3 to A2.5). While these checks often indicate a balanced sample around the ME threshold in t+1, particularly in checks that aggregate results for math and reading, we do see a few instances of imbalance, which we investigate further below.

First, we examine the probability that teachers in our analytic sample (i.e., Group 1 teachers in t linked to students in t+1) remain in Group 1 in year t+1 (Table A2.3). This check examines the likelihood that teachers who are in Group 1 in year t and who are linked to at least one student in t+1, are not linked to at least 15 students in t+1. We conduct this check to explore whether teachers who just receive an ME are able to manipulate the number of assigned students such that IVA is not ultimately included in their IMPACT score. An imbalance here could also signal a student assignment process by administrators that differentiates between teachers who just receive an ME from those who just receive an E, based on the quality signal. We detect no imbalance in 2010-11, but in 2011-12, teachers just rated ME are, counterintuitively, more likely to remain in Group 1 in specifications that condition on a linear form of the assignment variable (columns 1-3). However, this imbalance becomes statistically indistinguishable from zero in specification 4, which includes a quadratic form of the assignment variable (column 4), and this model appears to be the best fit according to the AIC. In 2012-13, propensity to remain in Group 1 in 2012-13 is negative, and is statistically significant at the 10 percent level in specification 4 (the model recommended by the AIC) in the stacked estimates, though non-significant in specifications 1-3. This constitutes a potential threat to internal validity, though visual inspection of the graphical plots does not strongly recommend inclusion of the quadratic functional form. [42] Furthermore, when we exclude Group 2 teachers in t+1, our results are largely unchanged.

We also examine the balance of rostered students at the threshold as a related check on differential student assignment (Table A2.4). In this robustness check as in the last, our findings are sensitive to inclusion of the higher-order polynomial form of initial IMPACT score in specification 4. We detect no significant imbalances in linear specifications in the stacked results, and significant imbalances when the quadratic assignment variable is included. In this case the best fitting model according to the AIC is the quadratic specification, which indicates some significant imbalance. However, these imbalances are concentrated in ELA where we do not detect robust effects, with no significant imbalances in the roster in math. There is also not a consistent pattern in the results (e.g., more assigned students in ELA in 2010-11, less in 2011-12 and 2012-13). As such, we do believe the roster imbalance in ELA signifies a credible threat to validity.

An additional check on the assumption of student exogeneity with regard to incentive treatment assignment involves conducting auxiliary RD regressions with predicted student achievement in year t+1, based on student observable characteristics (Table A2.5). In most years and subjects, our assumption of exogeneity appears to hold: we can not reject the null hypothesis that, in terms of predicted achievement, students are the same. However, we do see that just-ME teachers have students with predicted achievement that is approximately 0.06 SD higher in specifications 2 and 3 in math in 2012-13, though the imbalance does not hold in specification 4. We run an additional sensitivity check of excluding schools with significantly different predicted achievement

[42] Available from the authors on request.

for E and ME teachers; three schools meet this criteria in math in 2012-13. Our results are robust to the exclusion of these schools.

A final caveat relevant to most RD designs should also be noted here. Our causal estimands are distinctly "local" ones defined for the teachers near the threshold. One aspect to this caveat is the incentives created by a dismissal threat may be uniquely strong for ME-rated teachers who are, by definition, close to being rated Effective. However, it is also true that the incentive contrast we leverage may be muted by the fact that teachers who just barely achieved an E rating may experience a strong incentive to improve. We acknowledge this localness but still stress that these RD estimands constitute a causally credible test of the theory of change motivating such policies. Another caveat associated with the localness of the RD estimand is that they cannot clearly speak to the selection and incentive effects for teachers who are more distal from the threshold. However, we also note that this dimension of localness is still highly policy-relevant because the threshold exists among relatively low-performing teachers, a population of considerable interest.

6. RESULTS

6.1 Graphical evidence. Graphing the regression discontinuity provides a simple and compelling way to examine the relationship between student achievement in year t+1 and teachers' initial IMPACT scores in t, with particular attention paid to the discontinuous "jump" in the functional relationship between the two at the incentive threshold. We first examine math and reading together for each cross-section of data, and then separately by subject as there is important heterogeneity both between the subjects and over time. Following the first year of IMPACT, we see no discontinuity in achievement outcomes at the E/ME threshold in 2010-11 (Figure 2.2, Panel A). However, we do see “jumps” in achievement on the ME side of the threshold in each of the next two years (Figure 2.2, Panels B and C). These effects are similarly sized and appear to be just larger than 0.05 SD of student achievement. As noted earlier, the end of 2010-11 was the first time that twice-ME teachers were separated from the district. As such, 2011-12 is the first year that we observe performance effects once the dismissal threat associated with an ME rating has gained credibility. Thus, it may not be surprising to see larger effects of the dismissal policy in 2011-12 and beyond.

Figure 2.3 displays RD graphs of student achievement by subject for each cross section of data. In Panel A, we see evidence of a drop in students’ math achievement at the ME threshold—the opposite of what might be expected if incentives are theorized to positively affect student achievement. In Panel B we see the opposite: a positive jump in reading achievement outcomes for teachers just rated ME. The magnitude of the effect is similar in both graphs, just larger than 0.05 SD of student achievement. Panels C and D display math and reading graphs respectively for the cross-sections of data with achievement outcomes in 2011-12. In both graphs we see a positive jump in achievement on the ME side of the performance threshold and these effects appear somewhat larger in magnitude than those observed in 2010-11. Finally, Panels E and F display math and reading results for achievement outcomes in 2012-13. In this year, there appears to be a large, positive discontinuity in math achievement outcomes that is greater than 10 percent of a SD of student achievement. In reading, the fitted linear spline across the threshold is virtually straight indicating no effect of the incentive in this subject and year. In sum, there is suggestive evidence of a positive effect in four of our six subject years: 2010-11 in ELA, 2011-12 in both subjects, and 2012-13 in math.

6.2 Parametric Results and Non-Parametric Main Effects. The parametric estimates are largely consistent with the visual evidence provided by the graphs. However, estimating the regression model specified in Equation 1 allows us to obtain estimates of the effects we observe in the graphs, explore additional specifications and run tests of statistical inference. Table 2.2 presents these results by year for math and reading results combined.[43] In 2010-11, we do not see a statistically significant estimate which is consistent with the relatively small discontinuities observed operating in opposite directions that we observe in the graphs. In 2011-12 and 2012-13 we do find statistically detectable, positive effects that are roughly 7 percent of a SD of student achievement in our preferred specification (Specification 3), and estimates are fairly stable across specifications.[44] However, the result in 2011-12 loses statistical significance in specification 4, though this is not the best fitting model according to the AIC. The local linear regression results also suggest that the 2011-12 result may be sensitive to observations that are distal from the threshold—when the bandwidth is narrowed to 40 points surrounding the threshold, the point estimate drops to 0.038 and is no longer significant. In contrast, the 2012-13 result is robust both to inclusion of a higher-order term in the main estimation, and to the local linear specification with a point estimate of 0.127 when the bandwidth is narrowed to 40 IMPACT points. This indicates that, though

[43] In the table, we display results from four increasingly saturated specifications: specification 1 is a simple RD model that only conditions on a linear spline of initial IMPACT score, specification 2 adds school fixed effects, specification 3 adds grade fixed effects, and specification 4 adds a quadratic spline function of initial IMPACT score.

[44] We prefer Specification 3 because it offers a within-school, within-grade interpretation of the results, and is more parsimonious than Specification 4, which includes a higher-order function of the assignment variable. The AIC more often recommends Specification 3 for our main results, however, we also include Specification 4 as a robustness check and include the AIC for model comparison.

we see similarly sized positive effects of the incentive in 2011-12 and 2012-13 (i.e., 7 percent of a SD), the result in 2012-13 is more robust.

Table 2.4 examines these results separately by subject and reveals that the moderate positive effects following the second and third years of the program are largely driven by math, with consistently small results in reading. In math, we continue to see no effect of an ME rating following the first year of IMPACT implementation. We do observe a moderately sized effect in math in AY 2011-12, however it only reaches the level of statistical significance in a single specification and local linear regression results (Table 5) do not support the notion of a positive effect close to the ME threshold. In AY 2012-13 in math, there is a much larger effect (0.13 to 0.23 SD of student achievement) that are robust across specifications and in local linear regressions as well. In reading, we see that the suggested positive effects observed in the graphs in AY 2010-11 and AY 2011-12 are positive but not statistically significant when conditioning on only a linear spline function of the assignment variable in Column 1. The inclusion of school fixed effects in AY 2011-12 results in a statistically detectable 0.09 SD, though this is not robust to the inclusion of grade fixed effects or the higher order assignment variable. As suggested by the graphical evidence, we do not see an effect of an ME rating on student achievement in AY 2012-13 in any of our specifications. LLR results in Table 2.5 are consistent with this story of a null result in reading in all years: there are no estimates that are large or statistically distinguishable from zero when the bandwidth is narrowed around the ME performance threshold. Taken together, this suggests that teachers facing the dismissal threat associated with an ME rating are far more able to respond by improving student outcomes in math than they are in ELA.

6.3 Treatment Effect Heterogeneity. We also investigate the differential effect of the Minimally Effective incentive on the math and reading achievement of students in various demographic groups. As suggested by the theoretical model in this paper, if teachers believe that some students' academic performance may be more responsive to investments of their time and effort than others, highly incentivized teachers may focus their attention toward these students. This could result in weaker or even negative effects of the incentives for traditionally underserved student groups. We explore this treatment effect heterogeneity by including an interaction between our treatment variable and a binary indicator of student status on a number of demographic characteristics (e.g., special education status, free or reduced price lunch eligibility, race and gender) in Table 2.6.

Our analysis indicates that, for the most part, students from traditionally underserved subgroups do not experience lower achievement gains than the general student population when their teachers are faced with strong consequences. In fact, students receiving special education services consistently made larger gains than general education students, and in some years, students receiving Limited English Proficiency (LEP) services also benefitted compared to students not receiving these services. Table 2.6 summarizes treatment heterogeneity for math and reading achievement combined.[45] The first two rows show (1) the main ITT effect for the non-target subgroup and (2) the interaction for the subgroup of interest. When all years and both subjects are combined, special education students improved a statistically-significant 0.143 SD more than general education students. Looking over time, it is clear that this increased effect for

[45] Interestingly, the patterns of treatment heterogeneity are relatively consistent across subjects, despite the fact that math effects are consistently larger than those in reading, as discussed above. The separate math and reading results on treatment heterogeneity can be found in Appendix Tables A2.6 and A2.7.

students receiving special education services is relatively consistent at 0.093 (non-significant), 0.14 (p<0.10) and 0.182 (p<0.05) across years. None of the other aggregate results attain statistical significance, though the heterogeneous effects for LEP students are large and significant in two years (0.356 in 2010-11 and 0.179 in 2011-12), with negative (-0.154) but non-significant heterogeneity in the final year. The heterogeneity we detect for students who receive FRPL is positive but smaller in magnitude and only attains marginal statistical significance in one year, but it is encouraging that lower-income students are, at minimum, not experiencing lower effects that their more advantaged peers. Heterogeneous effects by student race are inconsistent across time, resulting in aggregate null effects: in 2010-11, white and Hispanic students experience much larger effects, while black students experience statistically larger effects in 2012-13. The inconsistency of these results over time makes it difficult to draw definitive conclusions about treatment heterogeneity by student race, particularly in the absence of a theory which describes why we might expect to see these differential effects over the years of our study. We will continue to examine these trends in future research.

7. DISCUSSION

Extensive evidence documents the important effect that teachers have on students' outcomes during their time in school and into adulthood. However, most existing policies that aim to improve teacher effectiveness (e.g., selection, induction, professional development, etc.) demonstrate little effect. Incentive policies that offer teachers financial bonuses for improved test outcomes without offering them information on how to improve are no exception. However, recent evidence suggests that using incentives in conjunction with performance feedback can be effective, and that framing incentives as a loss rather than a gain may be even more powerful. IMPACT's incentives for Minimally Effective teachers, which combine the potential for dismissal with intensive instructional feedback and access to coaching support, share both of these program features. DCPS's combined human capital reforms including evaluation, professional development, compensation and the potential for dismissal make it arguably one of the most comprehensive district-wide efforts to improve teacher effectiveness in the country. The consequence that Minimally Effective teachers face—the possibility of being fired if they do not improve—is also uniquely strong: it is much larger than any positive incentive that has been offered in the U.S. in financial terms, and unquestionably has implications for teachers' professional and personal identity as well. In addition, the five observation and debrief conversations that teachers take part in annually provide uncommonly extensive performance feedback. As a result, DCPS is a particularly compelling setting to examine the potential of incentives policies aimed at teachers to influence student outcomes.

We find that teachers who receive an ME rating and experience the dismissal threat respond by improving student test outcomes by approximately seven percent, conditional on their decision to return to the district following the second and third years of the program. These results are largely driven by gains in math, with smaller and less precise gains in 2011-12, and a large and more robust effect in 2012-13. Results in ELA are mostly null, though some specifications detect an effect in 2011-12. The magnitude of the average effects in 2011-12 and 2012-13 (i.e., roughly 0.07 SD) is educationally significant, equating to between 12 and 20 percent of a year of learning, depending on grade, or roughly seven percent of the black-white achievement gap (Hill, Bloom, Black

& Lipsey, 2007).[46] In the context of these benchmarks, the effect we observe in math in 2012-13 is staggeringly large at 0.23 SD in our preferred specification. However, given the inconsistency of this result with other years and subjects, as well as some evidence of imbalance in student assignment in this year, we caution against over interpretation of its magnitude.

It is noteworthy that we observe these positive effects only after IMPACT had been in place for at least two years, and had come through the political instability of changes in city and school district leadership intact. This is consistent with the null evaluations of shorter-term incentive pilot projects, and lends further credence to the assertion that teachers may not invest in responding to performance measures that they do not expect to continue in the long-term. An alternative explanation for the heterogeneity that we see by year is that learning to respond to new incentive measures can take time. The fact that our results increase over time could indicate that teachers are increasing their ability to respond to IMPACT as the program persists, and that the ability to do so is stronger in math than it is in ELA. This is consistent with a number of studies that detect lower variation and smaller effects of teacher-oriented interventions effects in ELA than in math (Hanushek & Rivkin, 2010; Taylor & Tyler, 2012). One hypothesis is that these subject-level differences arise because math is primarily learned in school whereas reading is also influenced and modeled in home settings (Taylor & Tyler, 2012). While differing results between subjects are interesting in any intervention setting, we find them particularly thought-provoking in the context of individual teacher accountability. Norm-

[46] I use math effect-size grade equivalents for yearly learning which range from 0.32 (7th to 8th grade) to 0.56 SD (4th to 5th grade) since our results are driven by gains in math and reading estimates result in less conservative approximations. The black-white achievement gap in math is estimated to be approximately one SD of student achievement (0.99 in grade 4, 1.04 in grade 8), (Hill, Bloom, Black, & Lipsey, 2007).

referenced value-added distributions by subject ensure that ELA teachers are not at a comparative disadvantage to math teachers overall, but what are the implications of one subject potentially being less malleable, or responsive to effort in the face of incentives? Are there additional supports that might help low-performing ELA teachers who face strong negative consequences realize similar gains to math teachers? We will soon be able to analyze IMPACT's student achievement effects using one additional year of DC CAS data (2013-14), which will provide additional evidence on whether incentive effects continue to be concentrated in math and whether the upward trend we observe in math continues over time.

In future research, we also hope to better understand how IMPACT's incentives effect non-test student outcomes. While it is important to detect whether or not teachers can respond to strong incentives by improving test outcomes (i.e., we learn something important if they cannot), we might be concerned that a positive effect could be driven by emphasizing tested outcomes at the expense of other goals of schooling or by outright cheating. While we cannot wholly rule out either of these explanations, we did obtain a list of potential testing violations from DCPS through 2011-12. During the IMPACT era, DCPS hired independent test-security firms (Caveon Test Security and Alvarez & Marsal) to assess potential violations. They identified critical violations in no more than a dozen classrooms per year. When we exclude the small number of student records identified as having potential testing violations in 2010-11 and 2011-12, our results remain unchanged both in terms of magnitude and statistical significance. Unfortunately, we only have test security data through 2011-12, thus, our final year of data—where we detect the largest effect—has not been subject to this test. Another often-discussed unintended consequence of strong incentives could be more narrowly-focused classrooms, where learning is rote and students are less engaged. We are unlikely to obtain direct measures of student engagement, but we have received student-level data on absences and suspensions that may provide some suggestive evidence on this question and intend to pursue this research soon.

The findings in the previous chapter of this dissertation also provides insight into the nature of the student achievement improvements we observe here. We detect identical patterns of incentive effects on teaching practice and student achievement over the three years: there is no effect in 2010-11, and moderately-sized effects in 2011-12 and 2012-13. These results are consistent with a hypothesis that the student achievement effects result from teachers' improvements in practice. While we cannot identify a causal link between improvements in teaching and improvements in student test outcomes, and the two improvements could be distinct, it seems conceptually unlikely. This implication underscores the importance of performance information and the prioritization of professional support for the development of skills, which is a key component of IMPACT's theory of change.[47]

A final note relates to the generalizability of our findings and of the evaluation reform in DCPS more broadly. IMPACT was introduced in 2009 under exceptionally unique conditions with immense political will invested in district reform by a newly elected mayor with mayoral control of district schools and the unusual situation of evaluation practices not being subject to collective bargaining. In the years since, it has become abundantly clear how unique this context was in the new evaluation and

[47] We may be able to gain additional purchase on this question using data from 2014-15 and 2015-16, when DCPS moved to new Common Core State Standard tests, and student test outcomes were not part of teachers' evaluations for two years.

personnel policies passed by other states and districts. Almost all states (46) and 23 of the largest districts have passed new legislation on teacher evaluation since 2009, but far fewer (60 percent of states and 39 percent of large districts) tie personnel decisions to evaluation results (Steinberg & Donaldson, 2015). Policies that involve the evaluation and dismissal of teachers necessarily imply greater risks to teachers than were faced under the previous system which rarely resulted in dismissal, and will always be contentious. There are legitimate consequences to high-stakes personnel systems that districts must consider such as potential narrowing of educational goals, increased teacher stress, and the opportunity cost of not pursuing other policy priorities. Our study does, however, show that incentives embedded within a comprehensive human capital system can improve student achievement, while simultaneously improving teacher performance as well. Whether or not there is a place for such policies in American education policy is an open question.

Table 2.1. Teacher and student characteristics by year, math and reading
Panel A. Math

	2010-11		2011-12		2012-13	
Variable	Obs	Mean	Obs	Mean	Obs	Mean
Teacher rated ME in t-1	7193	0.22	6084	0.17	4682	0.20
Student count	*	49	*	55	*	50
Teacher group 1 in t	7160	0.96	*	0.98	4643	0.95
Female teacher	*	0.68	*	0.74	*	0.72
Black teacher	*	0.56	*	0.54	*	0.48
White teacher	*	0.31	*	0.31	*	0.33
Male student	*	0.48	*	0.49	*	0.49
White student	*	0.09	*	0.13	*	0.11
Black student	*	0.74	*	0.70	*	0.64
Hispanic student	*	0.14	*	0.13	*	0.21
Special Education	*	0.12	*	0.12	*	0.12
Limited English Proficiency	*	0.07	*	0.06	*	0.09
Free or reduced meals	*	0.70	*	0.63	*	0.56
Elementary grade	*	0.49	*	0.51	*	0.47
School is high poverty	*	0.71	*	0.67	*	0.67
Standardized achievement	*	0.20	*	0.23	*	0.21
Residualized achievement	*	0.05	*	0.02	*	0.01

Panel B. Reading

	2010-11		2011-12		2012-13	
Variable	Obs	Mean	Obs	Mean	Obs	Mean
Teacher rated ME in t-1	7053	0.22	6908	0.17	5387	0.24
Student count	*	47	*	56	*	52
Teacher group 1 in t	7020	0.96	6797	0.98	5297	0.96
Female teacher	*	0.80	*	0.84	*	0.85
Black teacher	*	0.56	*	0.52	*	0.61
White teacher	*	0.32	*	0.34	*	0.29
Male student	*	0.49	*	0.48	*	0.49
White student	*	0.08	*	0.11	*	0.08
Black student	*	0.75	*	0.68	*	0.73
Hispanic student	*	0.15	*	0.17	*	0.15
Special Education	*	0.12	*	0.11	*	0.13
Limited English Proficiency	*	0.07	*	0.07	*	0.07
Free or reduced meals	*	0.73	*	0.66	*	0.63
Elementary grade	*	0.51	*	0.49	*	0.44
School is high poverty	*	0.75	*	0.68	*	0.73
Standardized achievement	*	0.15	*	0.21	*	0.07
Residualized achievement	*	0.03	*	0.00	*	-0.02

Notes: Number of observations is equivalent to number of observations with treatment data, unless otherwise noted.

Table 2.2. Reduced-form RD estimates for the effect of ME rating in year *t* on residualized student achievement in year *t+1*

	n	(1)	(2)	(3)	(4)
2010-11	14,246	0.015	-0.020	-0.011	0.073
		(0.049)	(0.046)	(0.044)	(0.068)
	AIC	-17482	-18151	-18204	-18212
2011-12	12,992	0.073*	0.095**	0.065*	0.062
		(0.043)	(0.040)	(0.037)	(0.041)
		-17274	-18004	-18120	-18118
2012-13	10,069	0.060	0.066**	0.073**	0.091**
		(0.042)	(0.033)	(0.030)	(0.045)
		-12822	-13315	-13371	-13374
School fixed effects			X	X	X
Grade fixed effects				X	X
Quadratic rating variable					X

Notes: Robust standard errors clustered by teacher in parentheses. All models condition on a linear spline function of the assignment variable. *** p<0.01, ** p<0.05, * p<0.1

Table 2.3. Reduced-form RD estimates for the effect of an ME rating in year *t* on residualized achievement in year *t+1*, by alternative bandwidth

	Full sample		$\lvert cr_i \rvert \leq 60$		$\lvert cr_i \rvert \leq 40$	
	n	est.	n	est.	n	est.
2010-11	14,446	-0.009 (0.044)	9,755	-0.045 (0.060)	6,662	0.035 (0.078)
2011-12	12,992	0.065* (0.037)	9,041	0.103** (0.040)	6,313	0.038 (0.042)
2012-13	10,129	0.074** (0.030)	5,989	0.073* (0.041)	4,145	0.127** (0.052)

Notes: Robust standard errors clustered by teacher in parentheses. Results condition on a linear spline function of the assignment variable and school and grade fixed effects.
*** p<0.01, ** p<0.05, * p<0.1

Table 2.4. Reduced-form RD estimates for effect of an ME rating in year *t* on residualized achievement in year *t+1* by subject

		n	(1)	(2)	(3)	(4)
Math	2010-11	7,193	-0.077	-0.026	-0.005	0.021
			(0.057)	(0.064)	(0.063)	(0.111)
		AIC	-8786	-9309	-9317	-9315
	2011-12	6,084	0.087	0.124*	0.093	0.114
			(0.064)	(0.074)	(0.073)	(0.079)
			-8011	-8588	-8622	-8619
	2012-13	4,682	0.130**	0.209***	0.225***	0.232***
			(0.052)	(0.051)	(0.045)	(0.061)
			-5820	-6232	-6276	-6274
ELA	2010-11	7,053	0.079	-0.010	-0.017	0.076
			(0.056)	(0.056)	(0.056)	(0.073)
			-8712	-9141	-9179	-9183
	2011-12	6,908	0.064	0.090**	0.048	-0.014
			(0.044)	(0.042)	(0.036)	(0.051)
			-9269	-9619	-9698	-9698
	2012-13	5,387	0.001	-0.024	-0.022	0.011
			(0.047)	(0.046)	(0.046)	(0.068)
			-7016	-7257	-7264	-7261
School fixed effects				X	X	X
Grade fixed effects					X	X
Quadratic rating var						X

Notes: Robust standard errors clustered by teacher in parentheses. All models condition on a linear spline function of the assignment variable.
*** p<0.01, ** p<0.05, *p<0.1

Table 2.5. Reduced-form RD estimates for effect of an ME rating in year *t* on residualized achievement in year *t+1* by subject, by alternative bandwidth

		Full sample		$\lvert cr_i \rvert \leq 60$		$\lvert cr_i \rvert \leq 40$	
		n	est.	n	est.	n	est.
Math	2010-11	7,234	-0.004	4,910	-0.002	3,379	0.063
			(0.063)		(0.082)		(0.101)
	2011-12	6,084	0.093	4,380	0.187**	2,996	0.061
			(0.073)		(0.082)		(0.100)
	2012-13	4,724	0.224***	2,686	0.201***	1,956	0.257***
			(0.045)		(0.062)		(0.065)
ELA	2010-11	7,212	-0.017	4,845	-0.040	3,283	0.111
			(0.056)		(0.072)		(0.090)
	2011-12	6,908	0.048	4,661	0.068	3,317	-0.031
			(0.036)		(0.049)		(0.055)
	2012-13	5,405	-0.022	3,303	-0.016	2,189	-0.066
			(0.046)		(0.061)		(0.115)

Notes: Robust standard errors clustered by teacher in parentheses. Results condition on a linear spline function of the assignment variable and school and grade fixed effects.
*** p<0.01, ** p<0.05, * p<0.1

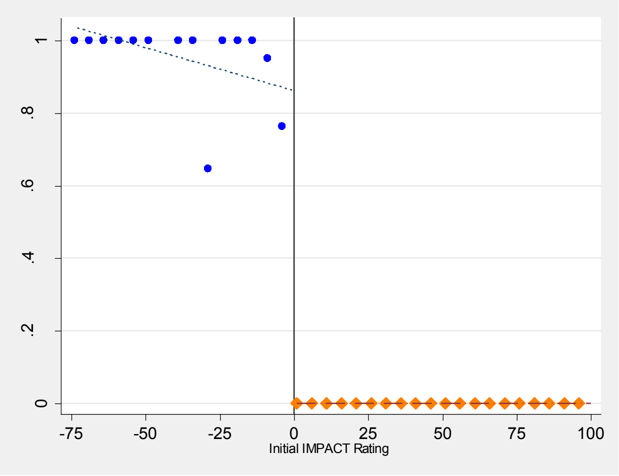
Table 2.6. Heterogeneity in main effects by student subgroup, math and reading

		(1)	(2)	(3)	(4)	(5)	(6)	(7)
	VARIABLES	Spec. Ed.	LEP	FRPL	Male	White	Black	Hispanic
All years	ITT	0.023	0.039	0.010	0.048**	0.044*	0.052	0.041*
		(0.025)	(0.024)	(0.032)	(0.023)	(0.024)	(0.046)	(0.025)
	ITT * Student Char.	0.143***	0.100	0.044	-0.010	-0.012	-0.007	0.042
		(0.048)	(0.078)	(0.030)	(0.022)	(0.082)	(0.048)	(0.049)
2010-11	ITT	-0.025	-0.024	-0.071	-0.002	-0.015	0.241***	-0.021
		(0.048)	(0.044)	(0.079)	(0.046)	(0.044)	(0.080)	(0.045)
	ITT * Student Char.	0.093	0.356**	0.073	-0.017	0.499***	-0.266***	0.249***
		(0.082)	(0.177)	(0.066)	(0.043)	(0.095)	(0.081)	(0.089)
2011-12	ITT	0.047	0.053	-0.008	0.069*	0.071*	0.062	0.052
		(0.037)	(0.036)	(0.048)	(0.038)	(0.039)	(0.052)	(0.038)
	ITT * Student Char.	0.140*	0.179**	0.103*	-0.009	-0.078	0.003	0.073
		(0.077)	(0.085)	(0.053)	(0.035)	(0.058)	(0.060)	(0.055)
2012-13	ITT	0.052	0.084***	0.080**	0.069**	0.076**	-0.013	0.090***
		(0.033)	(0.031)	(0.040)	(0.030)	(0.030)	(0.058)	(0.033)
	ITT * Student Char.	0.182**	-0.154	-0.011	0.009	-0.060	0.117**	-0.085
		(0.090)	(0.121)	(0.035)	(0.033)	(0.122)	(0.057)	(0.059)

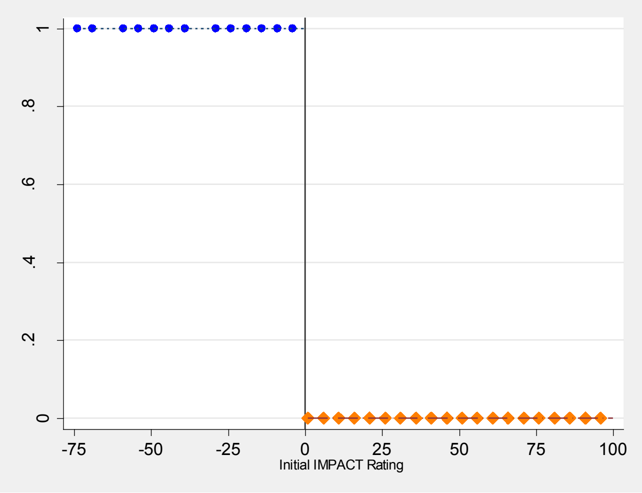
Notes: Robust standard errors clustered by teacher in parentheses. Results condition on a linear spline function of the assignment variable and school and grade fixed effects. ***p<0.01, **p<0.05, *p<0.1

Figure 2.1. First Stage, Intent-to-Treat at Minimally Effective Performance Threshold

Panel A. 2009-10



Panel B. 2010-11



Panel C.2012-13

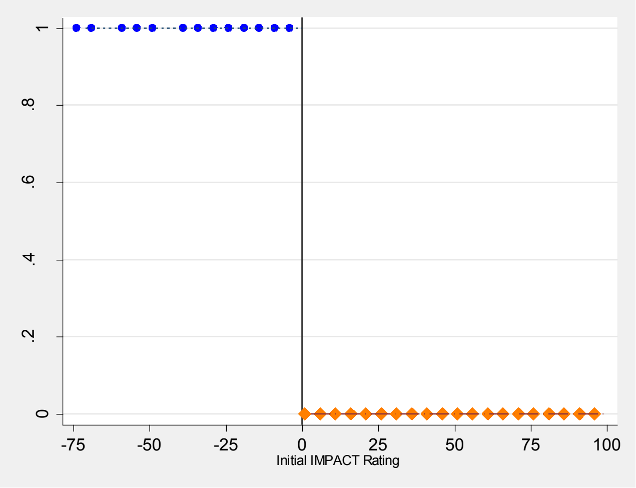
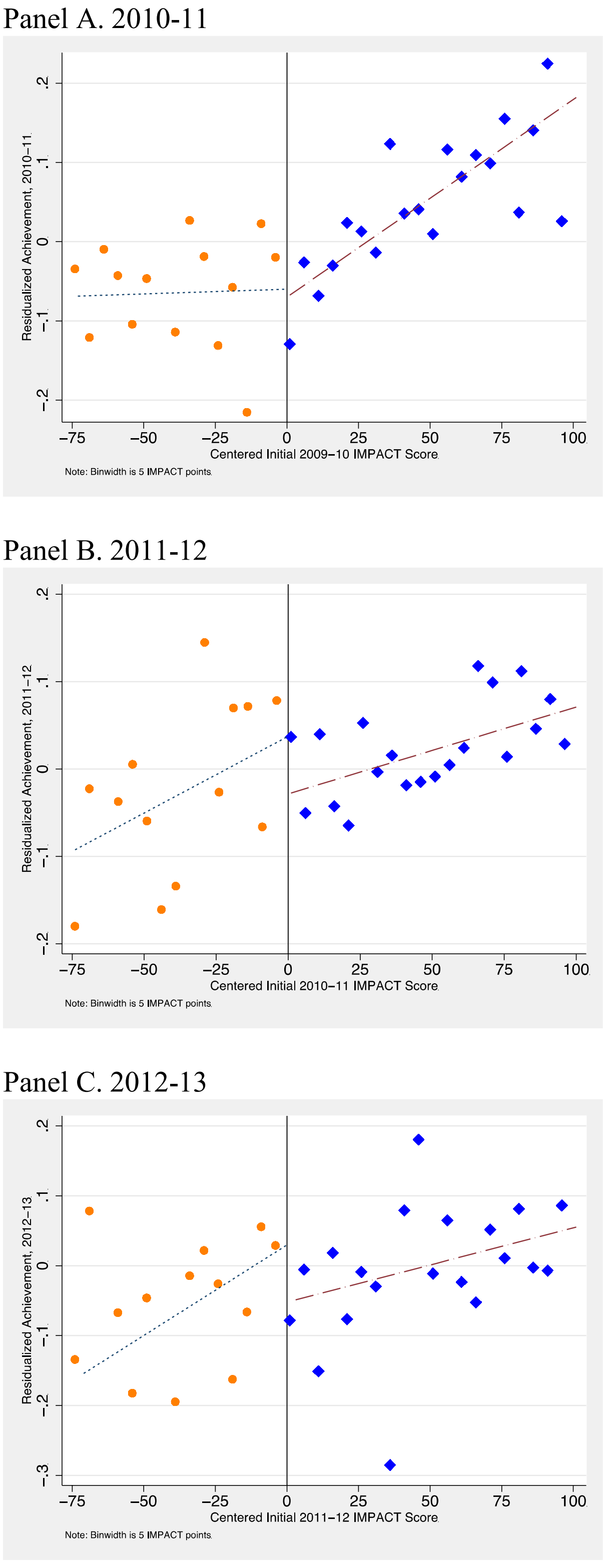


Figure 2.2. Incentive Effects at Minimally Effective Performance Threshold, Both Subjects

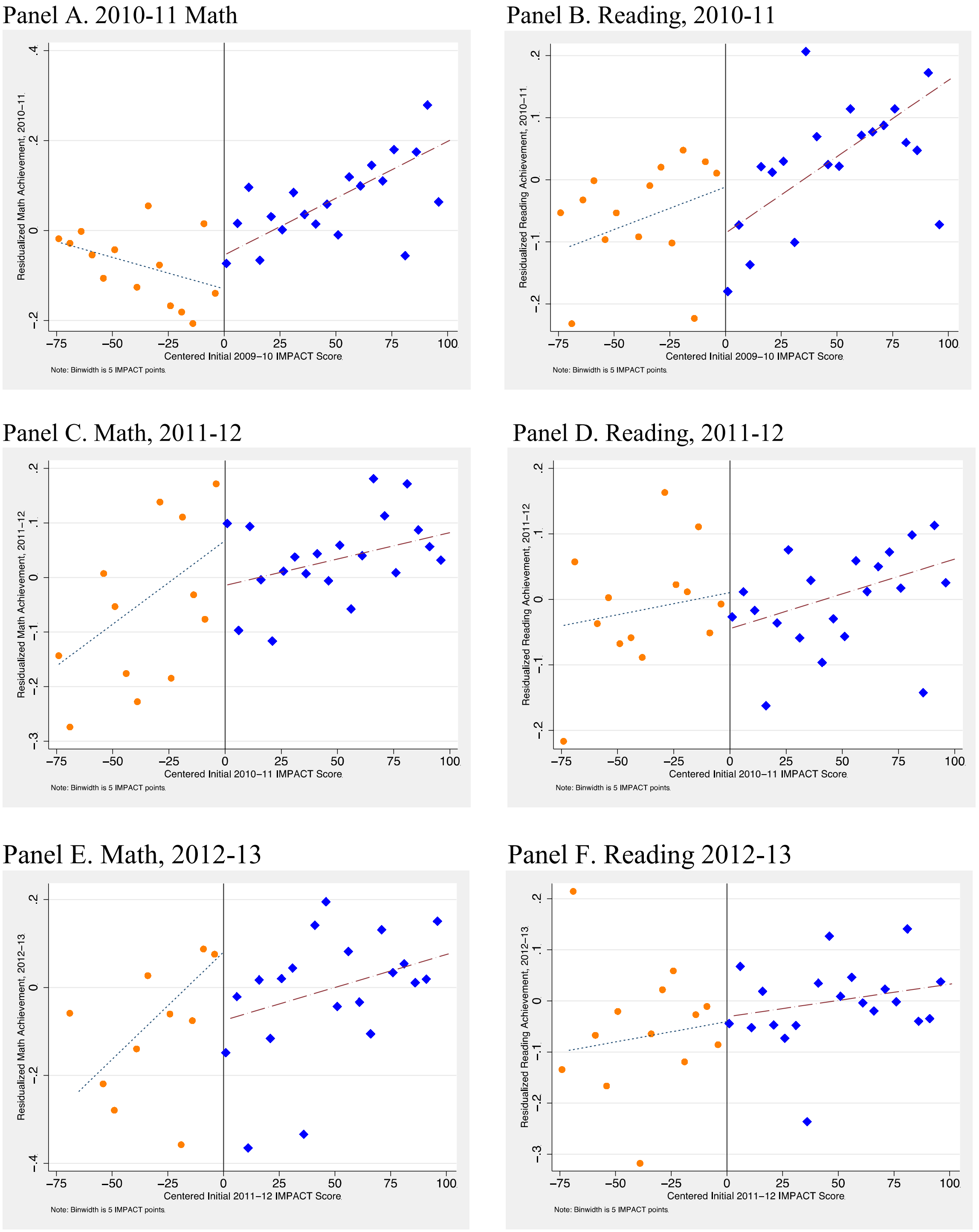


Panel A. 2010-11

Panel B. 2011-12

Panel C. 2012-13

Figure 2.3. Incentive Effects at Minimally Effective Performance Threshold, by Subject



Panel A. 2010-11 Math

Panel B. Reading, 2010-11

Panel C. Math, 2011-12

Panel D. Reading, 2011-12

Panel E. Math, 2012-13

Panel F. Reading 2012-13

CHAPTER 3

Identifying Profiles of Instruction from Classroom Observation in the District of Columbia Public Schools and the Measures of Effective Teaching Project

New teacher evaluation systems employ standardized observation protocols that provide information on teachers' performance on discrete instructional practices. This information is used to assess teachers' overall performance and also to provide formative feedback to teachers that may help them improve. However, beyond the accountability and improvement goals for this information, the advent of large-scale classroom observation data also provides a new opportunity to examine different typologies of teaching or "profiles of instruction" using granular information on classroom practice. I employ latent profile analysis to identify these profiles using data from two contexts: the District of Columbia Public Schools (DCPS) IMPACT teacher evaluation system, and the Measures of Effective Teaching (MET) research project. I find that the profiles of instruction derived in MET and DCPS are meaningfully different. While instructional profiles in the MET data provide some information on relative strengths and weaknesses, profiles in DCPS simply yield a leveled taxonomy of teaching effectiveness. This result is driven by the greater dimensionality of information from MET observation instruments than from DCPS's Teaching and Learning Framework. I discuss reasons that we may observe more unidimensional observation results in a real-world, high-stakes evaluation system than in a research setting, and the implications of this for policy and practice.

1. Introduction

Research has shown that there are large differences between teachers in their ability to help students achieve, and that these differences have large and persistent effects for students (Chetty, Friedman, & Rockoff, 2014; Jackson, 2012; Rockoff, 2004; Rivkin, Hanushek, & Kain, 2005). In recognition of these differences, many districts have introduced new teacher evaluation systems that incorporate standards-based classroom observation as a major component (Steinberg & Donaldson, 2015). While one goal of these systems is to provide an overall measure of teacher performance that can be used for accountability purposes, another, perhaps more important goal, is to use the individualized feedback on specific aspects of instruction to help teachers improve (Hill & Grossman, 2013; Papay 2012). For instance, in the District of Columbia Public Schools (DCPS) where teachers are observed five times a year on the Teaching and Learning Framework (TLF), evaluators provide teachers with feedback following each observation, and instructional coaches help teachers work toward improving particular aspects of their practice.

While emerging evidence suggests that identifying and targeting specific areas of instruction is an improvement over "one-size-fits-all" traditional models of PD (e.g., Allen, et al., 2011; Papay, Taylor, Tyler & Laski, 2016), this method may importantly neglect how competencies are interrelated within a teachers' overall professional practice. In fact, we know very little about teachers' patterns of practice, or "instructional profiles." If there are groups of teachers who share similar characteristics in their instructional practice (e.g., weak ability to provide strong content explanations, but strong classroom management), it may be helpful to explicitly identify these profiles in order to better coordinate professional development across multiple teaching competencies. There is increasing empirical and theoretical research to support this conception, but until recently the dearth of detailed data on classroom performance has made it impossible to explore patterns of practice empirically on a large scale.

This study uses data on teachers' classroom practice from the District of Columbia Public Schools (DCPS) IMPACT teacher evaluation system and the Measures of Effective Teaching Project (MET) to describe teachers' profiles of instruction using an innovative analytic technique, Latent Profile Analysis (LPA). LPA is an individual-centered latent variable measurement model that empirically identifies distinct behavioral patterns within a larger population (Collins & Lanza, 2010; Muthén 2004). Halpin and Kieffer (2015) illustrate the potential for the use of this technique to analyze classroom observation data, and I extend their work by applying it in two data sets—one importantly from actual evaluations of teachers in DCPS—to investigate how differences in classroom observation in research vs. real world settings may lead to differences in the instructional profiles observed.

Briefly, I find that the profiles of instruction derived in MET and DCPS data to be meaningfully different, in ways that are salient for their use in supporting teachers' professional needs. While instructional profiles in the MET data provide some information on relative strengths and weaknesses that could be used, for instance, to coordinate professional support, profiles in DCPS simply yield a leveled taxonomy of teaching. That is, the defining characteristic of the instructional profiles in DCPS is simply being "high" or "low" on all aspects of practice, which makes their construction less practically relevant for applications such as improving the coordination of professional development across multiple components of practice.

2. Background / Conceptual Model

Identification of instructional profiles that describe meaningfully different types of teaching relies on two factors: first, that different in patterns of teaching practice truly exist, and second, that the observation tool and process used to capture classroom practice produces information that can distinguish these differences. I discuss each of these below.

2.1 Differences in teachers' impacts on students. There is considerable evidence that teachers impact students' ability to attain a multitude of goals, from the development of academic and social skills to preparation for college and workforce participation. It is a nearly stylized fact in education research that teachers meaningfully impact students' academic outcomes and that there are big differences between teachers in their ability to help students achieve (Aaronson, Barrow, & Sanders, 2007; Chetty, Friedman, & Rockoff, 2014; Kane & Staiger, 2008; Rivkin, Hanushek, & Kain, 2005; Rockoff, 2004). Recently, the field has also become increasingly invested in understanding teachers' effects on non-test outcomes including students' attendance, GPA, on-time grade progression, higher-order thinking and motivation and mindset regarding academic work (Blazar, 2015; Gershenson 2016; Grissom, Loeb, & Doss, 2016; Jackson, 2013; Jennings & DiPrete, 2010; Kraft & Grace, 2016; Ladd & Sorenson, 2014; Ruzek, Domina, Conley, Duncan, & Karabenick, 2014). One implication of this work is that it provides a more detailed, empirically-derived representation of the multidimensional nature of teaching, where teachers can impact students on multiple dimensions and need not be equally skilled at helping students attain all goals.[48]

2.2 Multiple Dimensions of Teaching Practice. If teachers are differentially effective at supporting students in attaining different types of educational goals, it follows that they likely employ different kinds of instructional practice as well. For instance, a teacher who excels at supporting students in attaining high test outcomes might be focused on instructional strategies and maximizing time for instruction, while a teacher who improves students' socio-emotional skills may excel at building relationships with students and promoting a positive classroom climate. A new generation of standardized classroom observation instruments has greatly expanded our ability to observe these differences in teaching in recent years. If the differences in teachers' ability to help students achieve educational goals are caused by differences in observed instructional practice, then there are clear advantages for both research and practice in the use of these tools.

Many of the more recently developed frameworks for effective practice have strong theoretical foundations that support a multidimensional conceptualization of the teaching process, and suggest that meaningful differences in instructional profiles exist. Two frameworks that are widely used and that I employ as outcome measures in this study are the Teaching Through Interactions Framework, assessed through the Classroom Assessment Scoring System (CLASS; Pianta, LaParo & Hamre, 2008) and Charlotte

[48] Jackson (2013) formalizes this assertion in a two-dimensional latent factor model and shows that teachers who affect students' "non-cognitive" skills but not their "cognitive" (i.e., test score) skills can still have large impacts on long-term effects such as college attendance. Similarly, Kraft and Grace's (2016) recent analysis of teachers' effects on state standardized tests, complex performance tasks, and social-emotional measures finds only weak relationships between these measures, suggesting multidimensionality in teacher job performance.

Danielson's Framework for Teaching (FFT; Danielson, 2011). Each of these tools characterizes effective teaching across subjects using conceptually distinct domains.[49] Both of these frameworks will be reviewed in greater detail in the next section. However, I use CLASS as an illustrative example of the empirical research that supports our understanding of the dimensionality in teaching practice, because extensive analyses have been conducted on this, relative to other, observation tools.

CLASS is based on the importance of interactions between students and teachers as a primary mechanism for both academic and socio-emotional goals for students (Hamre et al., 2013). The framework is organized into three primary domains: Emotional Support, Instructional Support, and Classroom Organization. Each of these domains is grounded in developmental theory on the importance of interactions for students' learning and development (e.g., Bronfenbrenner & Morris, 1998). The developers of CLASS and other researchers have explored the instrument's dimensionality through extensive factor analyses. Hamre and colleagues (2013) investigate the factor structure of CLASS using data from over 4,000 classrooms. They validate that the three domain structure fits the data better than one- or two-factor solutions, but also find that the covariance between the three domains is high suggesting a general element of effective teaching. They further explore and validate the existence of this general factor in a bi-factor analysis (Hamre, Hatfield, Pianta, & Jamil, 2014), a finding that is consistent with MET analyses (Kane & Staiger, 2012). However, meaningful secondary factors are also identified in both analyses. A recent investigation of the CLASS-Secondary instrument

[49] In addition, content-specific classroom observation instruments characterize effective teaching through a subject-specific lens, which provides yet another potential dimension of instruction to consider. For instance, the Protocol for Language Arts Teaching Observation (PLATO; Grossman et al., 2013) draws on theory on adolescent literacy instruction for it's four domain description of effective language arts instruction.

across several large studies of middle and high school classrooms confirms that a three-factor structure fits the data better than a one- or two-factor structure in upper grades as well (Hafen, Hamre, Allen, Bell, Gitomer, & Pianta, 2015).

2.3. The influence of classroom observation features on observed instructional profiles. The research discussed above provides a strong empirical and theoretical basis for the existence of meaningfully different profiles of instruction among educators. However, features of classroom observation such as the observation instrument used, the scoring protocol, and the purpose of the observation all may influence the nature of the instructional profiles that we detect in this analysis. First, consider the intended dimensionality the of observation instrument itself. While many observational frameworks are grounded in theory and designed expressly to capture multiple dimensions of teaching as described above, other tools may be more strongly pervaded by a single philosophy for effective practice. For example, a "home-grown" observation rubric created by a school or district that strongly emphasizes the importance of high expectations for students across multiple competencies is more likely to yield a one-dimensional depiction of teaching. Frameworks that employ more distinct domains of effective instruction will logically produce item-level results that are more multidimensional in nature.

Differences in the protocol for scoring observations (i.e., the length of observations and observation setting) also play an important role. We know from prior work that the cognitive load of assigning observation ratings is high: in studies of observers "thinking aloud," Gitomer et al. (2014) report that observers face significant challenges both in mastering the observation protocol, and in applying it. Kane and Staiger (2012) frame the problem of rater cognition this way:

> Suppose that an observer has learned, perhaps even subconsciously, that virtually all of those who scored well (or poorly) in questioning strategies also scored well (or poorly) in classroom management. When that happens, an observer may have a hard time noticing the few cases in which a teacher with good questioning skills showed weak classroom management. Once they form an impression of one competency, they may subconsciously infer that the other was also present, even if they did not see it explicitly.

It is reasonable to expect that the limits on rater cognition are more pronounced when raters are required to process more pieces of information from longer observations. Likewise, settings where observers may be simultaneously processing additional information (e.g., the principal subtly monitoring student well-being during an observation) or scoring a lesson after the fact, may strain rater cognition and amplify the tendency of raters to assign higher or lower ratings across the board.

Finally, the "stakes" (i.e., rewards and consequences) associated with the observation will affect the quality of the information from the classroom observation as well. While the goal of observers in low-stakes settings (e.g. researchers) is presumably to rate each lesson as accurately and reliably as possible, observers in high-stakes settings (e.g., principals conducting performance evaluations) may also have other goals for observation ratings, such as producing an overall score that results in a particular consequence, "nudging" teachers to focus on a particular area of instruction, or even avoiding a difficult conversation. These managerial uses of classroom observation may be in conflict with the singular goal of accurate and reliable scores on observation items and may work against the construction of instructional profiles that capture "true" multidimensionality in teachers' practice.

3. Data and Context: DCPS IMPACT and the MET Project

This project utilizes data from two sources: administrative teacher evaluation records from DCPS, and records from the large-scale MET research project. Both sources of data provide detailed information on classroom practice. However, the way classroom observation data is captured in DCPS and MET differs considerably in ways that are meaningful for this analysis, so I outline these differences below.

3.1 Structure of DCPS IMPACT. DCPS began to implement IMPACT, a high-stakes teacher evaluation and compensation system, in the 2009-10 school year. Under IMPACT, teachers are assessed on multiple measures of performance, which depending on subject and grade may include: classroom observation on a district-created protocol, teacher individual value-added (IVA) in tested grades and subjects[50], teacher-assessed student achievement (TAS), measures of teacher core professionalism (CP) and contribution to school community (CSC), and school value-added (SVA). At the end of the year, these measures are aggregated to an overall score that assigns teachers to one of four rating categories: Ineffective, Minimally Effective, Effective or Highly Effective.[51] Teachers rated Ineffective are dismissed immediately, Minimally Effective teachers must improve to an Effective rating in the next year or be dismissed, and Highly Effective teachers receive a one-time bonus and are eligible for a permanent salary increase in the next year.

3.2 Teaching and Learning Framework. Particularly important for the present study is IMPACT's classroom observation measure, the Teaching and Learning

[50] Value-added models seek to isolate a teacher's contribution to student test score growth by controlling for observable characteristics of students and schools.

[51] In 2012-13, DCPS made two changes to the way they assign overall performance rating that reweights the components and creates a Developing rating to differentiate teachers within the Effective rating category. However, this study only includes data through 2011-12 and these changes don't apply.

Framework or TLF. Classroom observation is a major component of IMPACT, not only because it makes up the majority of the overall evaluation score for the majority of teachers, but also because it is the primary mechanism through which teachers receive feedback about their instructional practice.[52] TLF was originally designed by teachers, school leaders and central office staff in 2008-09 and was streamlined after the first year of implementation, yielding the nine teaching standards in Table 3.1, Panel A. TLF draws on instructional research including FFT and CLASS, as well as Wiggins and McTighe's Understanding by Design (2005). Teachers are observed using TLF five times each year: three times by a school administrator, and twice by a "master educator"—a content area expert who is employed by the district expressly for the purpose of conducting evaluations. Observations are at least thirty minutes. The first observation by an administrator is announced, and all subsequent observations are unannounced. Preliminary internal psychometric analyses of TLF suggest that the reliability and validity of the instrument are comparable to those in MET, but these results have not been publicly released.

In this analysis, I aggregate classroom observation data to the teacher-level, both in DCPS as well as in the MET data. This has two advantages: first, since LPA is a cross-sectional data analytic technique, this allows me to use a more stable measure of teachers' practice that is less subject to measurement error. Second, this mimics the annual data that districts would hypothetically use to determine profile membership for groupings of teachers. I focus the DCPS analysis on ratings assigned by master educators,

[52] For general education teachers without value-added scores (83% in this study), 75% of their overall IMPACT score is based primarily on TLF with minor weight given to TAS (10%), CSC (10%) and SVA (5%).For teachers in tested grades and subjects for whom IVA can be calculated (approximately 17%), this measure comprises 50% of the overall score, with TLF (35%), CSC (10%) and school value-added (5%) making up the rest.

both for internal consistency in my dissertation, and also because their expertise in rating is more likely to yield helpful instructional profiles. However, results that incorporate both administrator and Master Educator ratings are discussed as well and included in the appendix.

3.3 The Measures of Effective Teaching Project. The MET project was a large-scale research project funded by the Bill and Melinda Gates Foundation that took place between 2009 and 2012, involving more than 3,000 teachers in six districts across the country.[53] Teachers volunteered to participate in MET and were compensated for their time and effort. The goal of the project was to determine the extent to which effective teaching could be measured accurately and reliably using a variety of measures of teacher performance, and how that information could best be used to help teachers improve. MET project data included student and teacher surveys, a content knowledge for teaching measure, student achievement outcomes on standardized tests and alternative open-ended assessments, and videos of teachers' classroom performance that were scored using five instructional rubrics.[54]

MET data contains rich measures of classroom observation, similar to those in DCPS. However, the context for these scores could hardly be more different. Rather than having evaluators in the classroom conducting observations as is true in DCPS and most school districts, MET teachers recorded and submitted videos of their teaching, captured using specialized 360 degree cameras. Videos were scored by trained, external

[53] The participating districts were: Charlotte-Mecklenburg, NC; Dallas, TX; Denver, CO; Hillsborough County, FL; Memphis, TN; and New York City, NY. Pittsburg, PA served as a pilot district for MET, but no data from this district was included in the final MET data set.

[54] Rubrics employed in MET include: FFT, CLASS, the Mathematical Quality of Instruction (MQI; Hill et al., 2010), Protocol for Language Arts Teaching Observation (PLATO; Grossman et al., 2010) and UTeach Teacher Observation Protocol (UTOP; Marder & Walkington, 2010). For additional information on the MET project, please see http://metproject.org/.

raters who passed an initial certification test and were continuously calibrated in their ratings by expert “scoring leaders” (White, Rowan, Alter, & Greene, 2014). All videos were assessed on the cross-subject instruments, CLASS and FFT, and were also scored using subject-specific instruments in the relevant subject. This analysis is focused on CLASS and FFT. While assessing how the inclusion of content-specific measures of classroom practice affects instructional profiles is also an important area of research, I’ve purposely narrowed the scope to content-agnostic instruments to avoid introducing a content-specific dimension in MET data that is not evaluated in DCPS. CLASS and FFT, and the MET scoring protocol for each instrument, are described in more detail below.

3.4 CLASS. The Classroom Assessment Scoring System (Pianta, LaParo, & Hamre, 2008) was originally developed by researchers at the University of Virginia to assess the quality of teacher-student interactions in early childhood classrooms.[55] As discussed in the previous section, CLASS is organized into three primary domains: Emotional Support, Instructional Support, and Classroom Organization, with one additional item for Student Engagement. All items, organized by domain, included in Table 3.1, Panel B. CLASS is scored on a 7-point scale, with descriptions of high, mid, and low practice. In the MET project, classroom videos were split into 15 minute segments, and each segment was scored by a different rater on all 12 items. For this analysis, I aggregate CLASS item scores to the lesson-level, and then to the teacher-level (i.e., average scores within-lesson first, then within-teacher, by year), to better approximate the evaluation score structure in DCPS and other districts.[56]

[55] There are now grade-specific versions of CLASS that span Pre-K through secondary school; in this study, the upper elementary (grades 4-6) and secondary (grades 7-9) versions are used.

[56] The lesson-level aggregation is only important because some CLASS videos have three segments scored while others have two.

Relative to other available observation instruments, extensive research has been done on the psychometric properties of CLASS, much of it conducted by the instrument's designers. In addition to the research on the factor structure of CLASS detailed in the previous section, there is also extensive evidence on the reliability of the instrument and its relationship to student outcomes in the MET project and other research (Hamre & Pianta, 2005; Howes et al., 2008; Kane & Staiger, 2012; Ponitz et al., 2009; Rimm-Kaufman, Curby, Grimm, Nathanson, & Brock, 2009).

3.5 Framework for Teaching. The Framework for Teaching was originally written for use as part of the Praxis teacher licensing system, and has been used to assess teaching for nearly 20 years (Danielson, 2011). The full Framework contains 22 items in four domains: Planning and Preparation, Classroom Environment, Instruction, and Professional Responsibilities. In the MET study, only eight items within two domains (Classroom Environment and Instruction) are rated. Videos were scored on FFT as follows: raters watch a segment of the lesson from 0-15 minutes, the video skips minutes 15-25, and resumes for minutes 25-35. Thus, 25 discontinuous minutes of video are scored by a single rater.

Despite FFT's long history in the field, there is less psychometric research publicly available on it than on CLASS. However, MET has produced information on the validity and reliability of FFT, and a number of field-based studies have explored the relationship of FFT to measures student achievement. These descriptive studies establish a moderate correlation between FFT and student achievement measures (Gallagher, 2004; Kane, Taylor, Tyler, & Wooten, 2011; Kimball, White, Milanowski, & Borman, 2004; Milanowski, 2004). In MET, Kane and Staiger establish a causal relationship between a composite measures of teacher effectiveness that includes FFT and student achievement outcomes (2012), but Garrett and Steinberg are unable to establish a causal link between FFT alone and student achievement due to significant noncompliance with randomization (2015).

The DCPS and MET data sets complement one another and lend more strength to this analysis than either data set would alone. The classroom observation measures contained in the MET data have the advantage of validated classroom observation protocols and unbiased raters. Classroom observation data from DCPS on the other hand, is subject to a number of influences such as the high-stakes nature of the evaluation and personal relationships between evaluators and teachers, and exemplifies classroom observation data as it exists in the "real world." As such, constructing instructional profiles across the two data sets may provide more clarity about patterns of instructional practice and insight into the potential of this technique than the use of either data set on its own.

4. Empirical Strategy

4.1 Latent Variable Mixture Models. Latent Profile Analysis (LPA) is a measurement modeling technique used to categorize individuals into unobserved subgroups using item response patterns. (Collins & Lanza, 2010; Vermunt & Magidson, 2002). LPA is a specific, relatively straightforward application within the wider class of latent variable "mixture" modeling, which recognize the "mixture" of subgroups within the broader population. Latent profile models—also called Latent Class (LC) models in the case of categorical response items—date back to the 1960s, and are most commonly used in psychological and health sciences for diagnostic purposes (Collins & Lanza, 2010, Lazearfield & Henry 1968). The appeal of mixture models in this field is intuitive: it is a common goal in medicine to discern a diagnosis, or unobserved condition, from an observed set of symptoms. Latent profile models have been a fruitful analytic technique in this area. There is a natural application for LPA in education as well: diagnostic assessments for students and classroom observations for teachers provide a set of discrete indicators on performance. While we could treat each indicator individually, if there are patterns of student understanding or teacher practice that exist, uncovering and responding to these may improve our holistic treatment of the individual.

While mixture modeling is rare in education, another latent variable model, factor analysis, is much more common. Similar to factor analysis, LPA attempts to understand latent constructs using the covariance structure of observed item-level data. However, unlike factor analysis, which looks for variables that cluster together in the data, LPA looks for individuals who cluster together. This difference can also be understood through the nature of the latent variable: factor analyses conceptualize a *continuous* latent trait, while latent profile models conceptualize a *categorical* latent variable that explains the common variance between indicators.

4.2 Model specification and estimation. LPA treats item-level data for each individual as a vector, such that X = [X_1, X_2, … X_j]. Following Halpin and Kieffer (2015), the first statistical assumption in LPA is that some categorical latent variable Y (i.e., instructional profile), exists such that the joint distribution of p(X,Y) is well-defined. C is the total number of latent profiles, or classes.

$$p(X) = \sum_{c=1}^{C} p(Y = Y_c)\, p(X \mid Y = y_c) \tag{1}$$

The equation above indicates that the probability of observing an individual's particular set of classroom observation outcomes, p(X), is a weighted average of the class-specific probabilities of observing X. The second assumption of LPA is that, conditional on Y, items are statistically independent within class.

$$p(X\,|Y = y_c) = \prod_{j=1}^{J} p\left(X_j \mid Y = y_c\right) \tag{2}$$

This assumption, also implicit in factor analysis, assumes that the latent variable Y explains all correlation between the items. This means that the covariance matrix is decomposed into shared variance accounted for by Y and uncorrelated residuals (Bauer & Curran, 2004). This assumption can be relaxed in a more general mixture model, the finite normal mixture model, and I use this alternate model to test the robustness of my findings in the standard LPA model (Bauer & Curran, 2004; Uebersax, 1999).

By substituting Equation (2) into Equation (1), the overall model is as follows:

$$p(X) = \sum_{c=1}^{C} p(Y = Y_c) \prod_{j=1}^{J} p(X_j|\ Y = y_c) \tag{3}$$

Finally, since I am using teacher-level classroom observation items (i.e., items have been collapsed over repeated occasions), the outcome variables are continuous rather than categorical, as represented above. Thus, the model can be represented:

$$f(X) = \sum_{c=1}^{C} P(Y_c)\, f(X|\mu_c\,, \sigma_c) \tag{4}$$

Where f(x) is the joint density function, or the "mixture" of the class-specific densities (Vermunt & Magidson, 2002; Shulka, 2015). Put differently, f(x) is a weighted average of each of the class-specific density functions with a separate mean μ_c, and covariance matrix, σ_c for each class. I estimate this as an unconditional model because we would not want to control for these differences in performance before assigning teachers to professional development, the intended use of the profiles for the current study.

Estimation of LPA proceeds in two phases: (1) parameters of the model are estimated using maximum-likelihood estimation given a fixed number of latent classes, C, and (2) a combination of theory and fit statistics are used to determine the number of classes that best fit the data. From this point, class-specific parameters (i.e., mean and variance) for each item are calculated to determine what the profiles look like in terms of typical instruction. For each individual, a posterior probability of belonging to each instructional profile can be calculated as well. In an analysis such as this with teachers clustered within schools, cluster-robust standard errors are used in hypothesis testing to account for the nested structure of the data.

4.3. Auxiliary Exploratory Factor Analysis. In addition to the main LPA analysis, I also include a pairwise correlation matrix and auxiliary exploratory factor analysis (EFA) for each observation instrument. This allows me to explore the dimensionality of the instruments that are used to construct the instructional profiles. While analyses of the factor structure of the observation instruments included in MET have already been conducted (e.g., Hamre, et al., 2013; Hamre, Hatfield, Pianta, & Jamil, 2014; Lockwood, Savitsky, & McCaffrey, 2015; Kane & Staiger, 2012), there has not yet been a factor analysis of the observation rubric in DCPS, and I employ the same straightforward EFA to all instruments for the sake of consistency. I use a standard principal factor analysis to extract the factors, and retain factors using a combination of the Kaiser rule (i.e., retain factors with an eigenvalue above 1; Kaiser, 1960) and Cattell's scree test (1966; Fabrigar, 2012). I orthogonally rotate the factors (using varimax rotation) to examine the extent to which observation captures information on uncorrelated dimensions of teaching practice (DeCoster, 1998).

5. Analytic Sample

This analysis employs two years of cross-sectional classroom observation data from each source. I describe and compare the DCPS and MET analytic samples below.

5.1 DCPS Analytic Sample. To construct the analytic sample in DCPS, I employ two sample restrictions. First, I limit the sample to general education teachers to ensure that only TLF and not another observation framework is used. Next, I limit the timeframe of my analysis to the years of 2010-11 and 2011-12, because there were no changes in the evaluation program or the TLF rubric language during this time. This yields a sample of 2,694 teachers in 2010-11, and 2,669 teachers in 2011-12. Table 3.2 describes teacher demographic characteristics and average classroom observation scores for the analytic sample. The demographics of the sample are fairly stable across years. Nearly half of teachers (44-46 percent) work in an elementary school, with the rest working in secondary schools (36-37 percent) and educational campuses (18-19 percent) which are mostly K-8 (and some 6-12) schools. Thirty-one percent of teachers in both years are categorized as novice with three or fewer years of experience, and roughly two-thirds have a Masters degree. More than half of teachers are Black (57 percent), a third are White (35 percent) and three quarters are female. Panel B includes the mean and standard deviation (SD) for each “Teach” standard on TLF. There is considerable variation in average performance between the standards, with Teach 7 (Develop higher-level understanding through effective questioning) and Teach 3 (Engage students at all learning levels in rigorous work) having the lowest averages and Teach 9 (Build a supportive, learning-focused classroom) rated highest on average.

5.2 MET Analytic Sample. In constructing the MET analytic sample I use the Inter-university Consortium for Political and Social Research (ICPSR) Core Files with observation videos scored using both the FFT and CLASS. This follows the recommendation in the User Guide to the MET Longitudinal Database that "strongly recommends that researchers base their analyses on the Core Files whenever possible" (p. 6, White, Rowan, Alter, & Greene, 2014). I use data from both years of the MET project (2009-10 and 2010-11), and I restrict the sample to teachers in the core data files who have observation data for both FFT and CLASS. This yields a sample of 1554 teachers in 2009-10 and 1280 in 2010-11.

Table 3.3 describes the MET analytic sample, and provides some insight into differences between teachers in MET and DCPS. First, teachers in MET teach either Math (48 percent) or ELA (52 percent) in grades 4-9. This subsample was purposively chosen in order to ensure that all effectiveness measures (in particular, student achievement measures) were available for all teachers included in the study. In contrast, DCPS analyses include teachers in all grades and subjects. Teachers in the MET data also differ from teachers in DCPS in their demographic profile. For the subsample of teachers for whom we have information on classroom experience, it appears they are less likely to be novice (24 percent vs. 31 percent). They are also less likely to have a Masters degree, though this is likely an artifact of differences in the data collection, where in DCPS the Masters degree designation includes having a Bachelors degree + 30 credit hours. Teachers who participated in MET are more often White and less often Black than teachers in DCPS, with a similarly low proportion of teachers who are Hispanic.

Panel B includes the mean and SD for FFT and CLASS. As in DCPS, there are meaningful differences between the measures in average performance, though interestingly, some of the most challenging areas of practice in DCPS are also low-scoring in MET. "Using Questioning and Discussion Technique" is the lowest-scoring item on FFT (similar to DCPS), followed by "Using Assessment in Instruction." One CLASS item, "Negative Culture," has been reverse coded (e.g., 1=7, 7=1, etc.) to facilitate the interpretation of LPA item plots. On CLASS, the lowest-scoring item is "Analysis and Problem Solving" followed by "Regard for Student Perspectives." The highest-scoring items relate to behavior management on both rubrics—this is similar to the highest-scoring item in DCPS, "Build a supportive, learning-focused classroom," which includes indicators related to behavior management as well.

6. Results.

6.1 Correlation Matrices and EFA. I begin with correlation tables and exploratory factor analyses for each of the three instruments, and for CLASS and FFT together. Since latent variable analyses are, in essence, an interpretation of the correlation between these variables, looking at the raw correlation matrices and factor structure provides a transparent foundation for subsequent analyses.

Tables 3.4, 3.5 and 3.6 provides the correlation matrices for teacher-by-year scores for TLF, FFT and CLASS.[57] These tables make clear that there are strong relationships amongst items within all three observation instruments: across observation protocols, item-scores are moderately to highly correlated with each other. For TLF (Table 3.4), correlations range from 0.448 to 0.692, with an average inter-item correlation of 0.60. Correlations for FFT (Table 3.5) are similar, from 0.507 to 0.792.

[57] All tables include both years of data, but are virtually identical when analyzed separately by year.

The average inter-item correlation for FFT is slightly higher at 0.67. CLASS (Table 3.6) provides the broadest set of correlation coefficients, ranging from 0.212 to 0.869, with an average inter-item correlation of 0.558. It is noteworthy that the one CLASS component that is reverse coded, Negative Climate (NC) has a much lower positive correlation than other items. There are also several clusters of items that, while still moderately correlated (e.g., 0.3-0.4), are not highly correlated (e.g., 0.6-0.7) as are most items on TLF and FFT. For instance, there is only a 0.39 correlation between Behavior Management and Instructional Dialogue.

Correlations between CLASS And FFT items (Table 3.7) are also moderately high, ranging from 0.245 to 0.738 with an inter-item correlation of 0.496. In general, items on the two scales that seek to capture similar teacher competencies do appear to be more highly correlated (e.g., “Managing Student Behavior” and “Behavior Management” are correlated at 0.738) than items from dissimilar domains (e.g., “Managing Student Behavior” and “Analysis and Problem Solving”, r=0.396). The overall average correlation for the scales together is lower than either separately, indicating that multiple dimensions of practice may more likely be captured by the combination of the two scales than by either scale alone.

The clustering of variables observed in the correlation tables is explored more formally in exploratory factor analyses, which largely confirm what the correlation matrices suggest. EFA of both TLF and FFT find a single latent factor underlying each observation instrument. For both instruments, the first extracted factor explains a large proportion of the variance, with a barely detectable second factor (the eigenvalue for the second TLF factor is 0.22, eigenvalue for the second FFT factor is 0.45). Both the Kaiser rule (Kaiser, 1960) and scree plot examinations (Cattell, 1966) indicate a one-factor solution. Both tables of rotated factor loadings (Tables 3.8 and 3.9), indicate that all variables are approximately equally weighted in each instruments' single underlying factor. The similarities in the factor structure for these two instruments makes sense given the level of correspondence between them and the extent to which TLF is modeled on FFT.

Exploratory Factor Analysis for CLASS reveals a large first factor (eigenvalue=7.45) with a second factor (eigenvalue=1.01) that could arguably be excluded or retained. I retain it to investigate the composition of a two-factor class solution.[58] Orthogonal rotation reveals one factor that is characterized by positive Emotional Support items and Instructional Support, and one that includes Classroom Organization items and Negative Climate. Student Engagement is related to both factors. Finally, a factor analysis of CLASS and FFT together confirms that there are two definitive, uncorrelated factors captured by the two rubrics together (eigenvalue of the first extracted factor is 11.42, second factor is 1.90). Factor loadings (Table 3.11) reveal one factor that is characterized by primarily instructional components (particularly those in CLASS) and emotional support. The second factor contains all items that relate to behavior management, and also the instructional items from FFT. A stricter division into "Instruction" and "Behavior" factors would be consistent with a recent, more complex Bayesian EFA of four observation rubrics (CLASS, FFT, MQI and PLATO) which also finds support for two primary factors across all rubrics that are grouped into instruction and behavior clusters (Lockwood et al., 2015). However, in this more simplistic EFA, it

[58] A one-factor solution is an equally-weighted composite of all variables except for Negative Climate, which contributes less to the latent factor.

appears that shared variance at the rubric level—which is modeled for explicitly in the Lockwood analysis—is also being picked up by the factor structure in this analysis.

6.2 Latent Profile Analysis. Latent Profile Analyses were conducted using TLF data (9 items) in DCPS, and CLASS and FFT data together (20 items) in MET. I chose to explore instructional profiles using the multiple frameworks together in MET to leverage the multi-dimensionality uncovered in the factor analysis. I also conduct LPA for each instrument separately, and these results are included in the Appendix. Table 3.12 contains fit statistics for models containing between three and eight profiles for DCPS and MET. Fit statistics include the Bayesian Information Criterion (BIC), considered one of the best fit statistics for LPA models (Nylund, Asparouhov, & Muthen, 2007), and entropy and average classification probability (ACP) which both judge the certainty with which individuals can be classified into profiles (Muthen, 2004). The BIC in DCPS and MET data consistently indicate that model fit improves as additional profiles are added (lower BICs indicate better model fit). The improvement in fit declines rapidly after adding a fifth profile in DCPS and sixth profile in MET, however, suggesting that a model in this range may strike a good balance between classifying individuals precisely and overfitting the data.[59] A five-profile solution also has superior ACP and entropy, especially in DCPS. Finally, it is important to consider practicality in choosing between appropriately fitting models. Since adding profiles results in relatively small teacher counts in additional categories after five categories (e.g., a 76-person

[59] While the ideal model fit would be denoted by finding a minimum BIC, there is not a detectable minimum BIC in fitting up to 10 models in DCPS or MET. This is the case in many latent profile analyses, specifically the two I have read that involve constructing profiles of teacher performance (Halpin & Kieffer, 2015; Morin & Marsh, 2013). I also conduct a bootstrap likelihood ratio test (Nylund et al., 2007), and results provide no additional guidance beyond what is offered by the BIC.

category in the MET 6-profile solution), I chose the five profile solution. Moreover, adding a sixth profile does not change the trends I observe in the data.

Figure 3.1 shows mean item scores on TLF for five instructional profiles in DCPS in 2010-11.[60] Each line represents a single profile, and the differences in the heights of each TLF item indicate differences in the mean scores across profiles. As can be clearly seen, when LPA is applied in DCPS, the profiles that emerge are driven by differences in overall level of effectiveness. That is, there is really no difference in the "shape" of the five profiles; no indication of differences in relative strengths and weaknesses that theory and empirical evidence suggest may exist in teaching practice. Rather, strengths and weaknesses in instruction remain consistent across profiles, and differences between teachers are simply characterized as having more or less effective scores on all items. Across all five profiles, teachers perform worst on Teach 3 and 7. The error bars on Figure 3.1 show 95% confidence intervals on the mean estimates, indicating that each profile is statistically difference from the next across all TLF items. This finding is extremely robust across all models that I explored in the DCPS data. No matter how many profiles are included, we simply see a "leveled" profile result. Even relaxing the local independence assumption and allowing items to correlate within a profile (which can sometimes allow for better detection of differences in profile "shape"; Morin & Marsh, 2013; Uebersax, 1999), still results in leveled profiles in the best-fitting model.

As can be seen in Figure 3.2, the results of a latent profile analysis in FFT and CLASS using MET data are much different than those in DCPS. The first eight items on

[60] I construct profiles using a single cross-section of data, as this is the convention in LPA (Muthen & Muthen, 2000). This also offers the opportunity to check the robustness of results in a second year of data. In both DCPS and MET data, findings are very consistent from year to year. Tables corresponding to the figures can be found in the Appendix.

the left of the graph are FFT items, the remaining 12 are CLASS items. Figure 2 indicates that both “level” differences and “shape” differences characterize the instructional profiles in MET. Profiles do tend toward being higher or lower across most CLASS and FFT items. For example, Profile 1, which includes 7.9 percent of individuals, has a lower mean score than the next profile (Profile 2, with 17 percent of individuals) on all FFT and CLASS items. However, unlike in DCPS, mean scores for Profile 1 are not significantly lower than those in Profile 2 on all items.

This is difficult to see in Figure 3.2, so Figure 3.3 compares mean item scores for two profiles at a time. In Panel A, the similarities between Profiles 1 and 2 become more clear: there are statistically significant differences between the two profiles on all FFT items and the CLASS Negative Climate, Behavior Management, and Productivity items, but the two profiles are indistinguishable on Positive Climate, Teacher Sensitivity, Regard for Student Perspectives, and all of the Instructional Support items. These are the groupings of items that appeared in the EFA on CLASS and FFT as well, so we could say that these two profiles do not differ on Factor 1 (Positive Emotional Support/CLASS Instruction), but they do differ on Factor 2 (Danielson, Classroom Organization and Negative Climate). Panel B provides a second illustration of differences between instructional profiles in MET, and there is evidence of relative strengths and weaknesses between these two profiles. Profile 2 is characterized by stronger scores for behavior- and organization-oriented items, while Profile 3 is statistically better on CLASS indicators of instruction (though indistinguishable on FFT instructional indicators). Though these teachers exhibit meaningful differences in the nature of their practice, if these items were simply averaged to create a composite score, they would be virtually indistinguishable. It is worth noting that we do not see profiles with patterns of opposing strengths and weaknesses when CLASS and FFT are analyzed separately, as illustrated by Appendix Figures A3.2 and A3.3. Given the lower dimensionality of the independent frameworks that was evident in the factor analysis, it is not surprising that these independent MET instrument results look somewhat more similar to the findings in DCPS. However, both still exhibit greater differences in shape and weaker level-effects than the profiles constructed from TLF.

7. Discussion and Policy Implications

Classroom observation is widely employed in teacher evaluation and development efforts as a primary mechanism for teacher improvement. There is increasing empirical support for this model: recent evidence shows that providing teachers with individualized feedback following observation (Allen et al., 2011; Taylor & Tyler, 2012), aligning professional development with observation (Cohen, et al., in press), and matching teachers for peer support based on observed strengths and weaknesses (Papay, Taylor, Tyler, & Laski, 2016) have all been effective in improving outcomes for students. This is promising evidence, and suggests that classroom observation can be a powerful tool for improvement. However, we still know very little about the highest leverage ways to employ these tools. This study contributes to the ongoing effort to understand information from classroom observation and how it may be used to improve teaching practice. If classroom observations capture meaningful patterns of strengths and weaknesses over multiple dimensions of performance, then professional support could be organized so that teachers with similar (or complementary) areas for growth could work together. Indeed, a recent study finds that high-quality collaboration leads to improvements in teacher practice and student achievement (Ronfeldt, Owens Farmer, McQueen & Grissom, 2015).

This study provides evidence that, while latent profile analysis may hold promise for classroom observation instruments and processes that capture multiple dimensions of practice, it is of limited use for those that capture only a single dimension. In this case, profile analysis is unlikely to yield helpful insights because instructional profiles will simply report differences in level for that single dimension. As indicated in the conceptual model of this paper, there are a number of reasons that we may observe more limited dimensionality in DCPS observation scores than in the MET research, and these factors have significant implications for the use of observation outcomes as part of consequential evaluation systems. Perhaps the most obvious difference between observations in DCPS and the MET project is the context itself: TLF scores in DCPS are attached to strong consequences and rewards including the potential for dismissal, while MET scores are used for research purposes only. As a result, observers in these two settings likely face different goals for observation measures. By using the scores assigned by master educators in DCPS I have tried to minimize the potential influence of principal's "managerial" use of classroom observation, though the fact remains that DCPS observers are still assigning consequential ratings and conducting debrief conversations with teachers.

There is little research on the observation scoring process in high-stakes settings, though a recent study of observers in the Los Angeles Unified School District finds that there are few who operate "by-the-book," or use the narrow reasoning they were taught in training to score observations (Bell et al., 2015). The researchers report that four out of five observers used strategies to assign observation scores that were not formally supported by the observation protocol, such as using internal criteria (i.e., personal judgments not aligned to the scoring criteria). Another study in Dade County Public Schools in Florida compared principals' assigned observation scores with a low-stakes assessment of teacher performance that was shared with researchers only and found that principals facing accountability pressure inflate assessments relative to their true beliefs (Grissom & Loeb, 2014). These phenomena–-using internal criteria and score inflation—are almost certainly at play in DCPS as well, especially for school administrators, and contribute to observation scores that tend toward rating teachers as "high" or "low" across all areas of instruction. Indeed, in comparing the average inter-item correlation for all ratings (r=0.71) vs. those assigned only by master educators (r=0.60), it appears that administrators are much more likely to assign highly-correlated item scores.

Differences in observation medium and scoring protocol in MET and DCPS likely contribute to differences in the profiles observed in this study as well. As described earlier, observations were scored in MET by highly-trained external raters watching video segments—sometimes for as short as 15 minutes—while DCPS observation scores result from classroom visits during the course of the school day. One likely implication of these differences may be that the use of video allows observers to focus in on assigning ratings rather than on other factors that may be at play in the classroom. Additionally, while MET observers watch short segments and immediately assign ratings using a far more standardized protocol, district- and school-based evaluators are likely to approach the scoring process in a less regimented way.

Finally, the observation framework used will also importantly influence the dimensionality of teaching that is observed. Both FFT and CLASS identify domains of practice that are not necessarily orthogonal to one another, but that are conceptually distinct. In CLASS, in particular, these discrete domains have been empirically validated. It is also noteworthy that the dimensionality of the information from CLASS and FFT increases when both frameworks are analyzed together and which suggests that new information is added when frameworks based on different conceptualizations of effective teaching are combined. In DCPS, observation scores are based on only one domain (the "Teach" section of TLF) though, interestingly, there are two other domains included in TLF (the "Plan" domain and the "Increase Effectiveness" domain) which were never assessed. It is possible that if these domains were included in the DCPS observation process, more dimensionality in teaching effectiveness in the district may have emerged.

There are a number of additional analyses that could provide additional insight into to the profiles of instruction observed in high- and low-stakes settings. First, extending the current analysis to include a conditional LPA model that controls for covariates known to influence observation ratings (e.g., student attributes, teacher experience) could yield insights about the extent to which the differences we observe in the constructed instructional profiles are due to observable traits. This could uncover differences in relative strengths and weaknesses, conditional on a set of observable factors, that are less readily apparent in the unconditional model. Future research in this area could also explore the use of a more complex latent variable model, known as a factor mixture model (FMM), for modeling latent constructs in classroom observation data. Factor mixture models are hybrid models that incorporate both continuous and categorical latent variables (Bauer & Curran, 2004; Muthen & Asparouhov, 2006). FMMs can allow researchers to more clearly distinguish "shape" from "level" of profiles (Morin and Marsh, 2013), and is similar to a bi-factor analytic strategy (Chen, Hayes, Carver, Laurenceau, & Zhang, 2012; Hamre, et al., 2014). A continuous common factor is extracted first, and then the "shape" (i.e., relative strengths and weaknesses of the profile) is more readily apparent. It is possible that this model would fit the data better than a simple LPA model and reveal more about teachers' relative strengths and weaknesses, conditional on their overall effectiveness level. However, given the already complex nature of the analysis, this analysis should be undertaken with caution due to the increasingly complex statistical assumptions of FMMs (Curran & Bauer, 2004).

Another promising extension of this work is the examination of teachers' "profiles of instructional growth" rather than their relative strengths and weaknesses at a specific point in time. This research question asks whether there are groups of individuals who improve (or regress) in their instruction in similar ways over time. Theoretical research on learning to teach (e.g., Grossman et al., 2009) suggests that there may be important differences in the acquisition of knowledge for teaching, but this question hasn't been explored empirically on a large-scale. Recent developments to latent profile modeling and the expansion of large-scale observation data make this investigation more feasible than it has been in the past.

In conclusion, while researchers and practitioners using classroom observation data should consider whether a categorical latent variable might help to uncover meaningful profiles of instruction, this analysis suggests that this approach may have limited utility when the observation instrument primarily captures a general factor of effectiveness. However, when the observation is well-positioned to capture multiple dimensions of job performance through an instrument with multiple, conceptually distinct constructs and with a scoring protocol that supports rater cognition, there may be additional value gained by considering how individuals are grouped together in terms of the relative strengths and weaknesses of their instructional practice. The extent to which classroom observation tools used in real-world or high-stakes evaluation capture only one dimension of effective teaching is a challenge not only for the construction of instructional profiles, but for the goal of providing useful formative feedback to teachers more broadly. As observation-focused evaluation systems move to the forefront of districts' efforts to improve teaching, understanding the nature of information on teaching practice derived from classroom observation will continue to be an important area for further research.

Table 3.1. TLF, CLASS, and FFT Classroom Observation Instruments

Domain	Item
Panel A. TLF	
(no domains assigned)	1. Lead well-organized, objective-driven lessons
	2. Explain content clearly
	3. Engage students at all learning levels in rigorous work
	4. Provide students multiple ways to engage with content
	5. Check for student understanding
	6. Respond to student misunderstandings
	7. Develop higher-level understanding through effective questioning
	8. Maximize instructional time
	9. Build a supportive, learning-focused classroom
Panel B. CLASS	
Emotional Support	1. Positive climate (PC)
	2. Negative climate (NC)
	3. Teacher sensitivity (TS)
	4. Regard for student perspectives (RSP)
Classroom Organization	5. Behavior management (BM)
	6. Productivity (P)
	7. Instructional learning formats (ILF)
Instructional Support	8. Content understanding (CU)
	9. Analysis and problem solving (APS)
	10. Instructional dialogue (ID)
	11. Quality of feedback (QF)
Student engagement	12. Student engagement (SE)
Panel C. FFT	
Classroom environment	1. Creating an environment of respect and rapport (CERR)
	2. Establishing a culture for learning (ECL)
	3. Managing classroom procedures (MCP)
	4. Managing student behavior (MSB)
Instruction	5. Communicating with students (CS)
	6. Using questioning and discussion technique (UQDT)
	7. Engaging Students in Learning (ESL)
	8. Using assessment in instruction (UAI)

Notes: The 8-item version of FFT used in MET is abridged; for full version with 22 components, see https://www.danielsongroup.org/framework/.

Table 3.2. Sample Description, DCPS Analytic Sample

	2010-11			2011-12		
	N	Mean	SD	N	Mean	SD
Panel A. Demographics						
Elementary School	2694	0.44	0.50	2669	0.46	0.50
Secondary School	2694	0.37	0.48	2669	0.36	0.48
Educational Campus	2694	0.19	0.39	2669	0.18	0.38
Novice	2651	0.31	0.46	2617	0.31	0.46
Masters Degree?	2453	0.65	0.48	2582	0.68	0.47
Female	2610	0.74	0.44	2646	0.76	0.43
White	2426	0.35	0.48	2411	0.35	0.48
Black	2426	0.57	0.49	2411	0.57	0.50
Hispanic	2426	0.03	0.17	2411	0.04	0.20
Panel B. TLF Scores (Master Educator)						
TEACH 1	2686	3.03	0.59	2662	3.05	0.58
TEACH 2	2686	2.99	0.66	2662	3.01	0.63
TEACH 3	2686	2.71	0.67	2662	2.72	0.66
TEACH 4	2686	3.08	0.68	2662	3.11	0.63
TEACH 5	2686	3.04	0.63	2662	3.10	0.60
TEACH 6	2094	2.92	0.79	1612	2.90	0.79
TEACH 7	2686	2.48	0.75	2662	2.54	0.70
TEACH 8	2686	3.06	0.71	2662	3.10	0.67
TEACH 9	2686	3.19	0.61	2662	3.20	0.59

Notes: Educational campuses are primarily K-8 schools. Novice status indicates 3 or fewer years of experience. Masters degree categorization also includes BA+30 credit hours.

Table 3.3. Sample Description, MET Analytic Sample

	2009-10			2010-11		
	N	Mean	SD	N	Mean	SD
Panel A. Demographics						
Math	1554	0.48	0.43	1280	0.48	0.43
ELA	1554	0.52	0.43	1280	0.52	0.43
Elem (4-5)	1554	0.56	0.50	1280	0.54	0.50
Secondary (6-9)	1554	0.44	0.50	1280	0.46	0.50
Novice	688	0.24	0.43	581	0.24	0.43
Masters Degree?	1170	0.37	0.48	954	0.36	0.48
Female	1495	0.82	0.39	1218	0.82	0.39
White	1493	0.59	0.49	1216	0.60	0.49
Black	1493	0.34	0.47	1216	0.32	0.47
Hispanic	1493	0.05	0.22	1216	0.06	0.23
Panel B. Framework for Teaching						
CERR	1554	2.65	0.43	1280	2.63	0.42
ECL	1554	2.46	0.42	1280	2.48	0.41
MCP	1554	2.62	0.43	1280	2.64	0.40
MSB	1554	2.72	0.44	1280	2.71	0.40
CS	1554	2.59	0.37	1280	2.59	0.36
USDT	1554	2.18	0.40	1280	2.19	0.39
ESL	1554	2.39	0.40	1280	2.42	0.40
UAI	1554	2.25	0.41	1280	2.28	0.38
Panel C. CLASS						
PC	1554	4.36	0.77	1280	4.43	0.71
NC	1554	5.21	1.36	1280	5.16	1.22
TS	1554	4.09	0.65	1280	4.03	0.59
RSP	1554	3.08	0.72	1280	3.14	0.64
BM	1554	5.80	0.77	1280	5.77	0.65
PD	1554	5.71	0.64	1280	5.65	0.55
ILF	1554	4.11	0.68	1280	4.07	0.59
CU	1554	3.76	0.67	1280	3.82	0.59
APS	1554	2.60	0.63	1280	2.65	0.59
QF	1554	3.49	0.72	1280	3.50	0.69
ID	1554	3.21	0.73	1280	3.24	0.67
SE	1554	4.74	0.69	1280	4.77	0.63

Notes: Novice status indicates 3 or fewer years of experience.

Table 3.4. Pairwise correlation matrix for DCPS Teaching and Learning Framework

	TLF1	TLF2	TLF3	TLF4	TLF5	TLF6	TLF7	TLF8	TLF9
TLF1	1								
TLF2	0.6334	1							
TLF3	0.5989	0.6627	1						
TLF4	0.6337	0.658	0.7103	1					
TLF5	0.6016	0.651	0.6491	0.681	1				
TLF6	0.4482	0.5544	0.5459	0.528	0.5603	1			
TLF7	0.5577	0.5972	0.6381	0.593	0.5962	0.4958	1		
TLF8	0.56	0.5586	0.6339	0.6392	0.6043	0.4596	0.5314	1	
TLF9	0.532	0.5324	0.5457	0.5896	0.5742	0.4243	0.5026	0.6912	1

Notes: N= 5,348. Teacher-by-year scores assigned by Master Educators for 2010-11 and 2011-12.[61] All correlations are significant at p<0.001 level.

[61] Separate matrices by year are very similar and available by request.

Table 3.5. Pairwise correlation matrix for Framework for Teaching

	CERR	ECL	MCP	MSB	CS	USDT	ESL	UAI
CERR	1							
ECL	0.732	1						
MCP	0.743	0.662	1					
MSB	0.789	0.652	0.785					
CS	0.693	0.719	0.641	0.608	1			
USDT	0.593	0.686	0.539	0.507	0.659	1		
ESL	0.672	0.792	0.614	0.590	0.704	0.714	1	
UAI	0.606	0.701	0.567	0.534	0.663	0.694	0.724	1

Notes: N=2,834. Teacher-by-year FFT scores for 2009-10 and 2010-11. All correlations are significant at p<0.001 confidence level.

Table 3.6. Pairwise correlation matrix for CLASS

	PC	NC	TS	RSP	BM	PD	ILF	CU	APS	QF	ID	SE
PC	1											
NC	0.305	1										
TS	0.777	0.273	1									
RSP	0.731	0.212	0.668	1								
BM	0.494	0.446	0.450	0.324	1							
PD	0.511	0.373	0.503	0.369	0.809	1						
ILF	0.718	0.284	0.717	0.673	0.533	0.638	1					
CU	0.654	0.260	0.682	0.640	0.476	0.574	0.817	1				
APS	0.643	0.229	0.618	0.750	0.387	0.459	0.070	0.734	1			
QF	0.757	0.262	0.760	0.694	0.440	0.532	0.775	0.801	0.765	1		
ID	0.731	0.248	0.695	0.779	0.390	0.482	0.753	0.772	0.812	0.869	1	
SE	0.751	0.313	0.682	0.662	0.615	0.680	0.770	0.699	0.665	0.735	0.747	1

Notes: N=2,834. Teacher-by-year scores in 2009-10 and 2010-11. All correlations are significant at p<0.001 confidence level.
NC (Negative climate) item is reverse-coded.

Table 3.7. Pairwise correlation matrix for CLASS and FFT

	CERR	ECL	MCP	MSB	CS	USDT	ESL	UAI
PC	0.562	0.568	0.426	0.439	0.516	0.501	0.564	0.489
NC	0.364	0.31	0.337	0.370	0.301	0.245	0.284	0.237
TS	0.481	0.507	0.388	0.392	0.474	0.48	0.531	0.491
RSP	0.44	0.508	0.336	0.325	0.446	0.507	0.555	0.435
BM	0.664	0.553	0.669	0.738	0.517	0.421	0.511	0.456
PD	0.614	0.540	0.637	0.633	0.529	0.455	0.524	0.484
ILF	0.559	0.565	0.513	0.493	0.538	0.532	0.594	0.521
CU	0.518	0.543	0.494	0.467	0.537	0.521	0.540	0.517
APS	0.461	0.509	0.419	0.396	0.492	0.532	0.538	0.470
QF	0.516	0.561	0.449	0.427	0.539	0.560	0.568	0.548
ID	0.498	0.557	0.416	0.395	0.510	0.564	0.566	0.505
SE	0.595	0.612	0.538	0.546	0.539	0.512	0.613	0.515

Notes: N=2,834. Teacher-by-year scores in 2009-10 and 2010-11. All correlations are significant at $p<0.001$. NC (Negative climate) item is reverse coded.

Table 3.8. Exploratory Factor Analysis Factor Loadings, TLF

Variable	Factor1	Uniqueness
TLF1	0.849	0.280
TLF2	0.874	0.236
TLF3	0.872	0.240
TLF4	0.868	0.247
TLF5	0.873	0.237
TLF6	0.768	0.411
TLF7	0.827	0.316
TLF8	0.848	0.280
TLF9	0.799	0.362

Notes: N=3,706. Teacher-by-year TLF scores assigned by Master Educators for 2010-11 and 2011-12. Varimax rotated factor loadings.

Table 3.9. Exploratory Factor Analysis Factor Loadings, FFT

Variable	Factor 1	Uniqueness
CERR	0.852	0.274
ECL	0.870	0.243
MCP	0.801	0.358
MSB	0.794	0.370
CS	0.816	0.335
USDT	0.770	0.414
ESL	0.847	0.283
UAI	0.782	0.388

Notes: N=2,834. Analysis uses teacher-by-year scores in 2009-10 and 2010-11. Varimax rotated factor loadings.

Table 3.10. Exploratory Factor Analysis Factor Loadings, CLASS

Variable	Factor 1	Factor 2	Uniqueness
PC	0.765	0.358	0.287
NC	0.163	0.427	0.791
TS	0.745	0.332	0.335
RSP	0.830	0.134	0.294
BM	0.241	0.836	0.244
PD	0.336	0.815	0.223
ILF	0.758	0.445	0.226
CU	0.773	0.363	0.271
APS	0.818	0.209	0.287
QF	0.863	0.289	0.172
ID	0.897	0.211	0.151
SE	0.692	0.531	0.239

Notes: N=2,834. Teacher-by-year scores in 2009-10 and 2010-11.
Varimax rotated factor loadings below 0.4 appear in grey.

Table 3.11. EFA Factor Loadings, CLASS and FFT

Variable	Factor 1	Factor 2	Uniqueness
CERR	0.337	0.797	0.251
ECL	0.428	0.718	0.301
MCP	0.242	0.805	0.294
MSB	0.210	0.835	0.259
CS	0.394	0.681	0.381
USDT	0.465	0.565	0.465
ESL	0.474	0.661	0.338
UAI	0.414	0.615	0.451
PC	0.775	0.330	0.290
NC	0.177	0.382	0.823
TS	0.765	0.280	0.337
RSP	0.816	0.179	0.302
BM	0.262	0.756	0.360
PD	0.385	0.669	0.404
ILF	0.778	0.387	0.246
CU	0.777	0.341	0.279
APS	0.803	0.245	0.295
QF	0.864	0.286	0.172
ID	0.887	0.236	0.158
SE	0.715	0.460	0.277

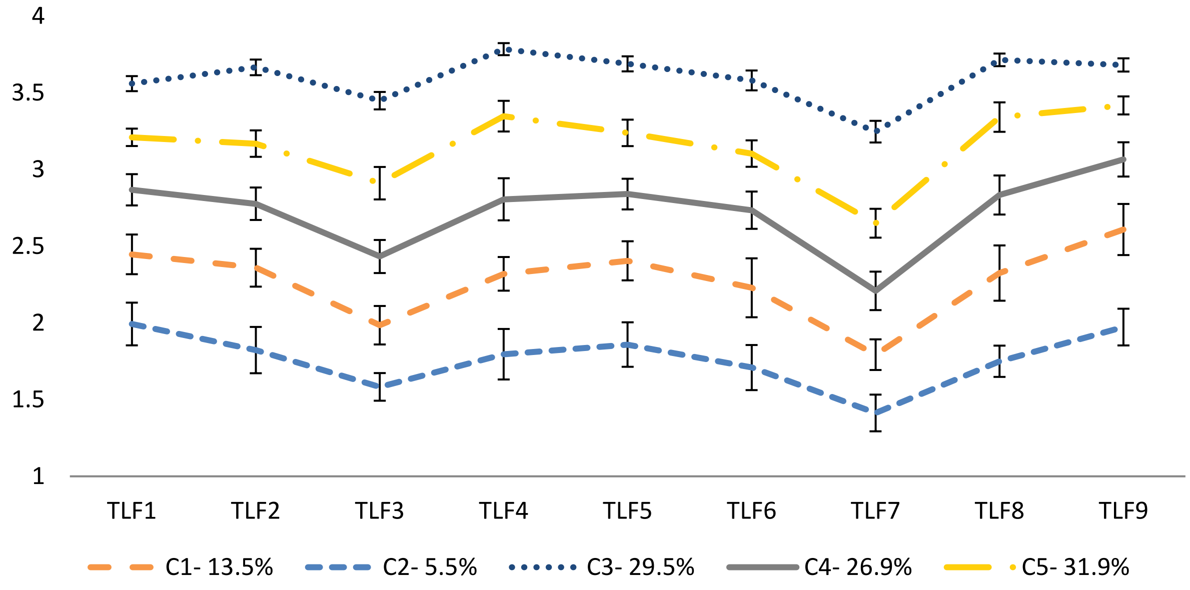
Notes: N=2,834. Teacher-by-year scores in 2009-10 and 2010-11. Varimax rotated factors below 0.4 appear in grey.

Table 3.12. LPA Model Fit Statistics, DCPS and MET

	DCPS, 2010-11				MET, 2009-10			
#	BIC	Δ in BIC	Entropy	ACP	BIC	Δ in BIC	Entropy	ACP
3	35374	3191	0.887	0.948	38145		0.939	0.976
4	34267	1106	0.856	0.921	36130	-2015	0.930	0.963
5	33995	272	0.826	0.886	34695	-1435	0.926	0.954
6	33900	95	0.79	0.838	33439	-1256	0.927	0.950
7	33808	93	0.786	0.812	32636	-803	0.925	0.950
8	33712	95	0.779	0.816	32178	-458	0.924	0.94

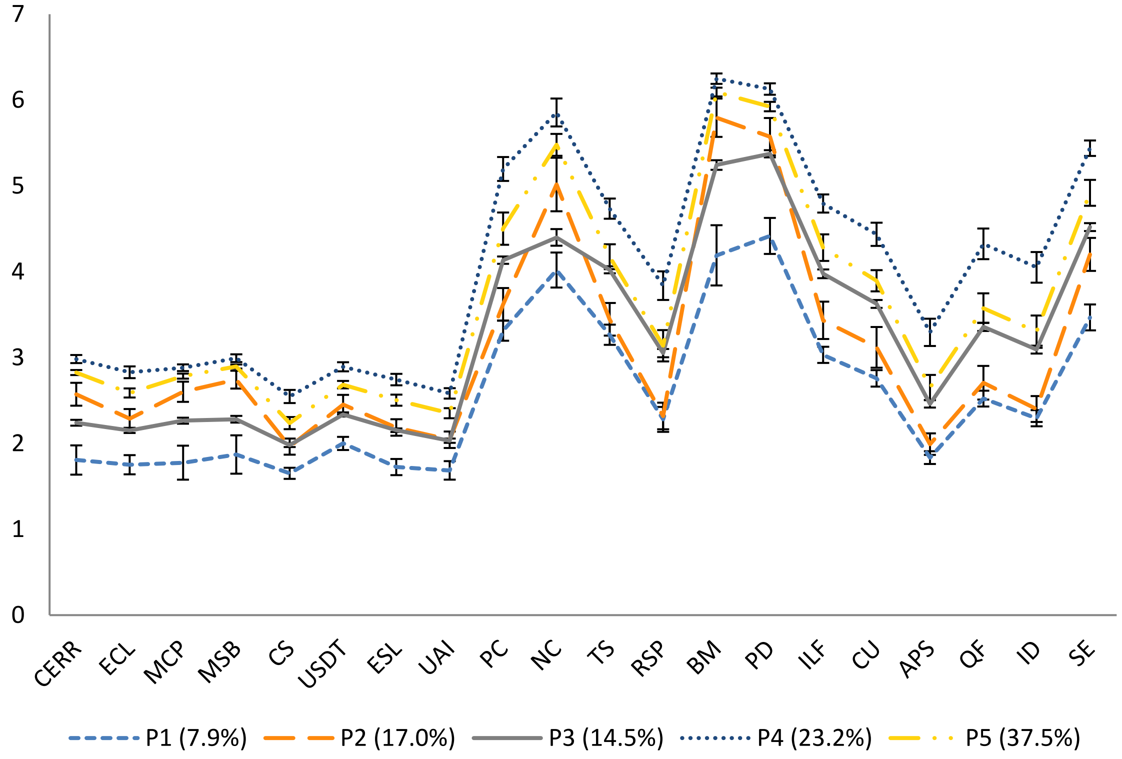
Notes: N=2,693 in DCPS, N=1,554 in MET. BIC is Bayesian Information Criteria, ACP is Average Classification Probability

Figure 3.1. Mean Item Scores for TLF, Five Class Solution (2010-11)



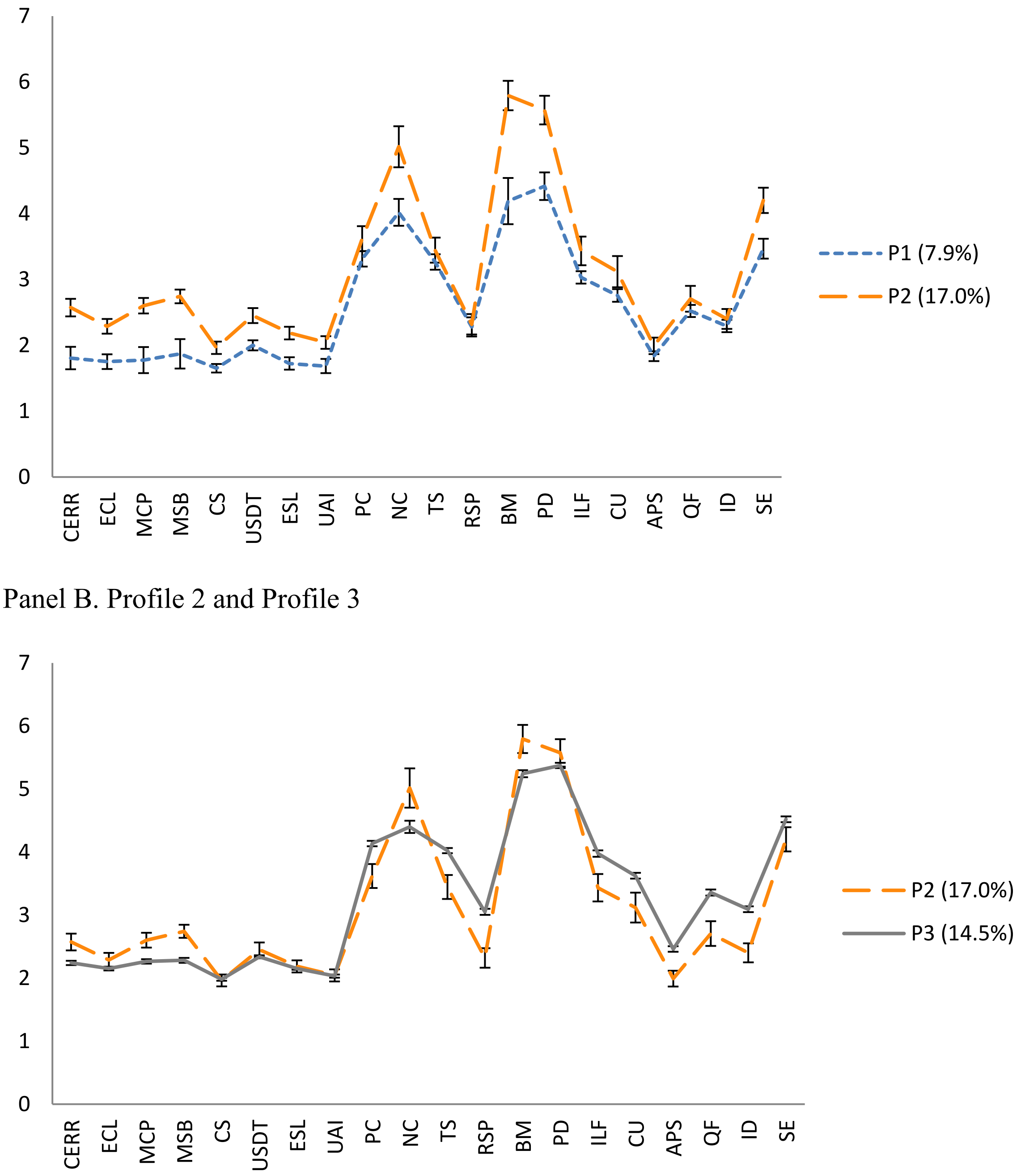
Notes: N=2,686. Error bars show 95% confidence intervals on mean estimates for each profile. Ratings based on scores assigned by master educators. Comparison is between height of each item; slopes are irrelevant.

Figure 3.2. Mean Item Scores for CLASS and FFT, Five Class Solution (2009-10)



Notes: N=1,554. Error bars show 95% confidence intervals on mean estimates for each profile. Comparison is between height of each item; slopes are irrelevant.

Figure 3.3. Mean Item Scores for CLASS and FFT Five Class Solution, Detail (2009-10)



Panel A. Profile 1 and Profile 2

Panel B. Profile 2 and Profile 3

Notes: N=1,554. Error bars show 95% confidence intervals on mean estimates for each profile. Comparison is between height of each item; slopes are irrelevant

REFERENCES

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and Student Achievement in the Chicago Public Schools. *Journal of Labor Economics* 25(1) 95-135.

Allen, J. P., Pianta, R. C., Gregory, A., Mikami, A. Y., & Lun, J. (2011). An interaction-based approach to enhancing secondary school instruction &student achievement. *Science*, *333*(6045), 1034-1037.

Atkinson, J. W. (1957). Motivational determinants of risk-taking behavior. *Psychological review*, *64*(6p1), 359.

Atteberry, A., Loeb, S. and Wyckoff, J. (2013). "Do First Impressions Matter? Improvement in Early Career Effectiveness," CALDER Working Paper No. 90, February 2013.

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., Ravitch, D., Rothstein, R., Shavelson, R. & Shepard, L. A. (2010). Problems with the Use of Student Test Scores to Evaluate Teachers. EPI Briefing Paper# 278. *Economic Policy Institute*.

Baker, G. (1992). Incentive Measures and Performance Measurement. *Journal of Political Economy*, 100, 598-614.

Baker, G. P., Gibbons, R. and Murphy, K. J. (1994). Subjective Performance Measures in Optimal Incentive Contracts. *Quarterly Journal of Economics*, 109, 1125-1156.

Ball, D. L., & Forzani, F. M. (2009). The work of teaching and the challenge for teacher education. *Journal of teacher education*, *60*(5), 497-511.

Bandiera, O., Barankay, I., and Rasul, I. (2007). Incentives for Managers and Inequality Among Workers: Evidence from a Firm-Level Experiment. *Quarterly Journal of Economics*, 122(2), 729-773.

Bandura, A. (1997). *Self-efficacy: The exercise of control*. Macmillan.

Bauer, D. J., & Curran, P. J. (2004). The integration of continuous and discrete latent variable models: potential problems and promising opportunities. *Psychological methods*, *9*(1), 3.

Bell, C. Qi, Y., Jones, N., Lewis, J., Kirui, D., Stickler, L., Redash, A. (2015). "Administrators' Strategies for Evaluating Teaching Practice." Paper presented at the Association for Education Finance and Policy, February 2015.

Bell, C. A., Qi, Y., Croft, A. J., Leusner, D., Gitomer, D. H., McCaffrey, D. F., & Pianta, R. C. (2013). Improving observational score quality: Challenges in observer thinking. *Measures of Effective Teaching. San Francisco: Jossey-Bass*.

Blazar, D. (2015). Effective teaching in elementary mathematics: Identifying classroom practices that support student achievement. *Economics of Education Review*, *48*, 16-29.

Boyd, D., Grossman, P., Lankford, H., Loeb, S., & Wyckoff, J. (2009). Teacher preparation and student achievement. *Educational Evaluation and Policy Analysis*, *31*(4), 416-440.

Bronfenbrenner, U., & Morris, P. A. (1998). The ecology of developmental processes.

Campbell, D. T. (1969). Reforms as experiments. *American psychologist*, *24*(4), 409.

Cannon, M. D., & Witherspoon, R. (2005). Actionable feedback: Unlocking the power of learning and performance improvement. *The Academy of Management Executive*, *19*(2), 120-134.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate behavioral research*, *1*(2), 245-276.

Chen, F. F., Hayes, A., Carver, C. S., Laurenceau, J. P., & Zhang, Z. (2012). Modeling general and specific variance in multifaceted constructs: A comparison of the bifactor model to other approaches. *Journal of Personality*, *80*(1), 219-251.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood. *The American Economic Review*, *104*(9), 2633-2679.

Chiang, H., Wellington, A., Hallgren, K., Speroni, C., Herrmann, M., Glazerman, S., Constantine, J. (2015). Evaluation of the Teacher Incentive Fund: Implementation and Impacts of Pay-for-Performance After Two Years. *Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance*

Cohen, J. (2015). The challenge of identifying high-leverage practices. *Teachers College Record*, *117*(8).

Cohen, J. Chambers Schuldt, L., Brown, L., & Grossman, P. (in press). Leveraging Observation Tools for Instructional Improvement: Exploring Variability in Uptake of Ambitious Instructional Practices.

Cohodes, S. (2015). Teaching to the Student: Charter School Effectiveness in Spite of Perverse Incentives. *Journal of Education Policy and Finance.*

Collins, L. M., & Lanza, S. T. (2010). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences* (Vol. 718). John Wiley & Sons.

Danielson, C. (2011). *Enhancing Professional Practice: A Framework for Teaching.* Alexandria, VA: Association for Supervision and Curriculum Development.

Darling-Hammond, L., Wei, R. C., Andree, A., Richardson, N., & Orphanos, S. (2009). Professional learning in the learning profession. *Washington, DC: National Staff Development Council*.

Deci, E. L., Koestner, R., & Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin, 125*(6), 627.

DeCoster, J. (1998). Overview of factor analysis. Retrieved February 18, 2016 from: http://www.stat-help.com/factor.pdf

Dee, T. S. & Wyckoff, J. H. (2015). Incentives, selection, and teacher performance: Evidence from IMPACT. *Journal of Policy Analysis and Management*, 34(2), 267-297.

Doherty, K. M., Jacobs, S. (2015). *State of the States: Evaluating Teaching, Leading, and Learning.* Washington D.C.: National Council on Teacher Quality. Retrieved November 20, 2015 from: http://www.nctq.org/dmsView/StateofStates2015.

Donaldson, M. & Papay, J. (2015). Teacher Evaluation. In Helen F. Ladd and Margaret E. Goertz (Eds.), *Handbook of Research in Education Finance and Policy* (2nd ed., pp. 174-193). New York: Routledge.

Dweck, C. (2006). *Mindset: The new psychology of success*. Random House.

Eccles, J. S., & Wigfield, A. (1995). In the mind of the achiever: The structure of adolescents' academic achievement related-beliefs and self-perceptions. *Personality and Social Psychology Bulletin*, *21*(3), 215-225.

Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology*, *53,* 109-132.

Eccles (Parsons), J., Adler, T. F., Futterman, R., Goff, S. B., Kaczala, C. M., Meece, J. L., & Midgley, C. (1983). Expectancies, values, and academic behaviors. In J. T. Spence (Ed.), *Achievement and achievement motivation* (pp. 75–146). San Francisco, CA: W. H. Freeman.

Eraut, M. (1994). *Developing professional knowledge and competence*. Psychology Press.

Fabrigar, L. R., & Wegener, D. T. (2012). Understanding statistics: Exploratory factor analysis. *New York, NY: Oxford University*.

Fang, M., & Gerhart, B. "How Does Pay for Individual Performance Influence Intrinsic Interest and Motivation Orientation?" Manuscript. University of Wisconsin, Madison, 2005.

Fryer, R. (2013). "Teacher Incentives and Student Achievement: Evidence form New York City Public Schools," *Journal of Labor Economics* 31(2): 373-427.

Fryer, R., Levitt, S., List, J., & Sadoff, S. (2012). Enhancing the efficacy of teacher incentives through loss aversion. NBER Working Paper No. 18237. Cambridge, MA: National Bureau of Economic Research.

Gallagher, H. A. (2004). Vaughn Elementary's innovative teacher evaluation system: Are teacher evaluation scores related to growth in student achievement?. *Peabody Journal of Education*, *79*(4), 79-107.

Garet, M. S., Cronen, S., Eaton, M., Kurki, A., Ludwig, M., Jones, W., . . . & Sztejnberg, L. (2008). *The impact of two professional development interventions on early reading instruction and achievement* (NCEE 2008-4030). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Garet, M. S., Porter, A. C., Desimone, L., Birman, B. F., & Yoon, K. S. (2001). What makes professional development effective? Results from a national sample of teachers. *American Educational Research Journal*, *38*(4), 915–945.

Garet, M. S., Wayne, A. J., Stancavage, F., Taylor, J., Eaton, M., Walters, K., . . . & Doolittle, F. (2011). *Middle School Mathematics Professional Development*

Impact Study: Findings after the second year of implementation (NCEE 2011-4025). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Garrett, R., & Steinberg, M. P. (2015). Examining Teacher Effectiveness Using Classroom Observation Scores Evidence From the Randomization of Teachers to Students. *Educational Evaluation and Policy Analysis*, *37*(2), 224-242.

Gershenson, S. (2016). Linking Teacher Quality, Student Attendance, and Student Achievement. Education Finance and Policy. Springr 2016, Vol. 11, No. 2: 125–149.

Gill, B., Bruch, J., & Booker, K. (2013). Using Alternative Student Growth Measures for Evaluating Teacher Performance: What the Literature Says. REL 2013-002. *Regional Educational Laboratory Mid-Atlantic*.

Gitomer, D. H., Bell, C. A., Qi, Y., McCaffrey, D. F., Hamre, B. K., & Pianta, R. C. (2014). The instructional challenge in improving teaching quality: Lessons from a classroom observation protocol. *Teachers College Record*, *116*(6).

Glazerman, S., E. Isenberg, S. Dolfin, M. Bleeker, A. Johnson, M. Grider, and M. Jacobus. (2010). *Impacts of Comprehensive Teacher Induction: Final Results From a Randomized Controlled Study* (NCEE 2010-4028). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Glazerman, S. and Seifullah, A. (2012). An Evaluation of the Chicago Teacher Advancement Program (Chicago TAP) After Four Years. Princeton: Mathematica Policy Research.

Goe, L., & Croft, A. (2009). Methods of Evaluating Teacher Effectiveness. Research-to-Practice Brief. *National Comprehensive Center for Teacher Quality*.

Goldhaber, D. (2015). Teachers matter, but effective teacher quality policies are elusive. In Helen F. Ladd and Margaret E. Goertz (Eds.), *Handbook of Research in Education Finance and Policy* (2nd ed). New York: Routledge.

Goldhaber, D. (2015). Exploring the Potential of Value-Added Performance Measures to Affect the Quality of the Teacher Workforce. *Educational Researcher*, *44*(2), 87-95.

Goldhaber, D. and Hansen, M. (2010). "Is It Just a Bad Class? Assessing the Stability of Measured Teacher Performance." CEDR Working Paper 2010-3. University of Washington, Seattle, WA.

Goldhaber, D., Liddle, S., & Theobald, R. (2013). The gateway to the profession: Assessing teacher preparation programs based on student achievement. *Economics of Education Review*, *34*, 29-44.

Goldhaber, D., & Chaplin, D. (2012). Assessing the 'Rothstein Falsification Test': Does It Really Show Teacher Value-Added Models Are Biased? *Center for Education Data & Research Working Paper*.

Goldring, E., Grissom, J. A., Rubin, M., Neumerski, C. M., Cannata, M., Drake, T., & Schuermann, P. (2015). Make Room Value Added Principals' Human Capital

Decisions and the Emergence of Teacher Observation Data.*Educational Researcher*, *44*(2), 96-104.

Goldring, E., Grissom, J. A., Rubin, M., Neumerski, C. M., Cannata, M., Drake, T., & Schuermann, P. (2015). Make Room Value Added Principals' Human Capital Decisions and the Emergence of Teacher Observation Data.*Educational Researcher*, *44*(2), 96-104.

Grissom, J. A., Kalogrides, D., & Loeb, S. (2015). Using Student Test Scores to Measure Principal Performance. *Educational Evaluation and Policy Analysis*, *37*(1), 3-28.

Grissom, J. A., Loeb, S., & Doss, C. (2015). The Multiple Dimensions of Teacher Quality. *Improving Teacher Evaluation Systems: Making the Most of Multiple Measures*, 37.

Grossman, P. L. (1992). Why models matter: An alternate view on professional growth in teaching. *Review of Educational Research*, 171-179.

Grossman, P. Loeb, S. Cohen, J. & Wyckoff, J. (2013). Measure for Measure: The Relationship between Measures of Instructional Practice in Middle School English Language Arts and Teachers' Value-Added Scores. *American Journal of Education*. 119(3). 445-470.

Grossman, P. L., Smagorinsky, P., & Valencia, S. (1999). Appropriating tools for teaching English: A theoretical framework for research on learning to teach. *American Journal of Education*, 1-29.

Grossman, P. L., Valencia, S. W., Evans, K., Thompson, C., Martin, S., & Place, N. (2000). Transitions into teaching: Learning to teach writing in teacher education and beyond. *Journal of Literacy Research, 32*, 631-662.

Hafen, C. A., Hamre, B. K., Allen, J. P., Bell, C. A., Gitomer, D. H., & Pianta, R. C. (2015). Teaching through interactions in secondary school classrooms revisiting the factor structure and practical application of the classroom assessment scoring system–secondary. *The Journal of Early Adolescence*, *35*(5-6), 651-680.

Halpin, P. F., & Kieffer, M. J. (2015). Describing Profiles of Instructional Practice A New Approach to Analyzing Classroom Observation Data. *Educational Researcher*, 44(5), 263-277.

Hamre, B. K., & Pianta, R. C. (2005). Can instructional and emotional support in the first grade classroom make a difference for children at risk of school failure? *Child Development*, **76**(5), 949 –967.

Hamre, B. K., Pianta, R. C., Downer, J. T., DeCoster, J., Mashburn, A. J., Jones, S. M., ... & Brackett, M. A. (2013). Teaching through interactions. *The Elementary School Journal*, *113*(4), 461-487.

Hamre, B., Hatfield, B., Pianta, R., & Jamil, F. (2014). Evidence for General and Domain- Specific Elements of Teacher–Child Interactions: Associations With Preschool Children's Development. *Child Development*, *85*(3), 1257-1274.

Hanushek, E. A., Kain, J. F., O'Brien, D. M., & Rivkin, S. G. (2005). *The market for teacher quality* (No. w11154). National Bureau of Economic Research.

Hanushek, E. A., & Lindseth, A. A. (2009). *Schoolhouses, courthouses, and statehouses: Solving the funding-achievement puzzle in America's public schools*. Princeton University Press.

Harris, D. & Harrington, C. (2015). Editors' Introduction: The Use of Teacher Value-Added Measures in Schools: New Evidence, Unanswered Questions, and Future Prospects. *Educational Researcher*. March 2015 44: 71-76.

Harris, D. N., & Sass, T. R. (2011). Teacher training, teacher quality and student achievement. *Journal of public economics*, *95*(7), 798-812.

Heckman, J. J., Stixrud, J., & Urzua, S. (2006). *The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior* (No. w12006). National Bureau of Economic Research.

Heneman III, H. G. (1998). Assessment of the motivational reactions of teachers to a school-based performance award program. *Journal of Personnel Evaluation in Education*, *12*(1), 43-59.

Hill, C.J, H.S. Bloom, A, R. Black, and M.W. Lipsey (2007). *Empirical Benchmarks for Interpreting Effect Sizes in Research*, MDRC Working Papers on Research Methodology, New York, N.Y.: MDRC. Available at: http://onlinelibrary.wiley.com/doi/10.1111/j.1750-8606.2008.00061.x/pdf

Hill, H. C., Beisiegel, M., Jacob, R. (2013). Professional Development Research: Consensus, Crossroads, and Challenges. *Education Researcher, 42*(9), 476-487.

Hill, H., & Grossman, P. (2013). Learning from teacher observations: Challenges and opportunities posed by new teacher evaluation systems. *Harvard educational review*, *83*(2), 371-384.

Holmstrom, B. & Milgrom, P. (1991). Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design. *Journal of Law, Economics, and Organization*, 7, 24-52.

Howes, C., Burchinal, M., Pianta, R., Bryant, D., Early, D., Clifford, R., & Barbarin, O. (2008). Ready to learn? Children's pre-academic achievement in pre-kindergarten programs. *Early Childhood Research Quarterly*, **23**(1), 27–50.

Hulleman, C. S., & Barron, K. E. (2010). Performance Pay and Teacher Motivation: Separating Myth from Reality. *Phi Delta Kappan*, *91*(8), 27-31.

Imbens, G., & Kalyanaraman, K. (2011). Optimal bandwidth choice for the regression discontinuity estimator. *The Review of Economic Studies*, rdr043.

Imbens, G. W., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics, 142*(2), 615-635.

Jackson, C. K. (2012). *Non-cognitive ability, test scores, and teacher quality: Evidence from 9th grade teachers in North Carolina* (No. w18624). National Bureau of Economic Research.

Jacob, B. A., & Lefgren, L. (2004). The impact of teacher training on student achievement quasi-experimental evidence from school reform efforts in Chicago. *Journal of Human Resources*, *39*(1), 50-79.

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

Johnson, S. M. and The Project on the Next Generation of Teachers (2007). *Finders and Keepers: Helping New Teachers Survive and Thrive in Our Schools*. Jossey-Bass, An Imprint of Wiley. Indianapolis, IN.

Kane, T., & Cantrell, S. (2010). Learning about teaching: Initial findings from the measures of effective teaching project. *MET Project Research Paper, Bill & Melinda Gates Foundation*, 9.

Kane, T., Kerr, K., & Pianta, R. (2014). *Designing teacher evaluation systems: New guidance from the measures of effective teaching project*. John Wiley & Sons.

Kane, T. J. & Ho, A. (2013). Ensuring Fair and Reliable Measures of Effective Teaching: Culminating Findings from the MET Project's Three-Year Study. The Bill and Melinda Gates Foundation, January 2013.

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

Kane, T. J., & Staiger, D. O. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation* (No. w14607). National Bureau of Economic Research.

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

Kimball, S. M., White, B., Milanowski, A. T., & Borman, G. (2004). Examining the relationship between teacher evaluation and student assessment results in Washoe County. *Peabody Journal of Education*, *79*(4), 54-78.

Kluger, A. ,& DeNisi. (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.

Koedel, C., & Betts, J. R. (2011). Does Student Sorting Invalidate Value-Added Models of Teacher Effectiveness? An Extended Analysis of the Rothstein Critique. *Education Finance & Policy*,*6*, 18-42.

Koedel, C., & Betts, J. (2010). Value added to what? How a ceiling in the testing instrument influences value-added estimation. *Education*, *5*(1), 54-81.

Kraft MA, Grace S. (2016). Teaching for Tomorrow's Economy? Teacher Effects on Complex Cognitive Skills and Social-Emotional Competencies. Working Paper.

Kraft, M. A., & Papay, J. P. (2014). Can Professional Environments in Schools Promote Teacher Development? Explaining Heterogeneity in Returns to Teaching Experience. *Educational Evaluation and Policy Analysis*, January 30, 2014.

Ladd, H. F., & Sorensen, L. C. (2014). Returns to teacher experience: Student achievement and motivation in middle school. *Center for Analysis of Longitudinal Data in Education Research Working Paper*, *112*.

Lazear, E. P. (1995). *Personnel Economics*, Cambridge: MIT Press.

--. (2000). Performance Pay and Productivity. *American Economic Review*, 90(5), 1346-1361.

Lazear, E. P., & Oyer, P. (2009). Personnel Economics. Robert Gibbons and John Roberts (Eds.), *Handbook of Organizational Economics.* Princeton, NJ: Princeton University Press

Lazarsfeld, P. F., Henry, N. W., & Anderson, T. W. (1968). *Latent structure analysis*. Boston: Houghton Mifflin.

Lockwood, J. R., Savitsky, T. D., & McCaffrey, D. F. (2015). Inferring constructs of effective teaching from classroom observations: An application of Bayesian exploratory factor analysis without restrictions. *The Annals of Applied Statistics*, *9*(3), 1484-1509.

Locke, E. A., & Latham, G. P. (2002). Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *American psychologist*, *57*(9), 705.

Lee, D.S. & Lemieux, T. (2009). Regression discontinuity designs in economics. *Journal of Economic Literature*, *48*, 281-355.

McCaffrey, D. F., Sass, T. R., Lockwood, J. R., & Mihaly, K. (2009). The intertemporal variability of teacher effect estimates. *Education*, *4*(4), 572-606.

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: a density test. *Journal of Econometrics, 142(2)*, 698-714.

Measures of Effective Teaching (MET) Project (2013). Ensuring Fair and Reliable Measures of Effective Teaching: Culminating Findings from the MET Project's Three-Year Study. The Bill and Melinda Gates Foundation, January 2013.

Mihaly, K., McCaffrey, D. F., Staiger, D. O., & Lockwood, J. R. (2013). A composite estimator of effective teaching. *Seattle, WA: Bill & Melinda Gates Foundation*.

Milanowski, A. (2004). The relationship between teacher performance evaluation scores and student achievement: Evidence from Cincinnati. *peabody Journal of Education*, *79*(4), 33-53.

Milgrom, P., & Roberts, J. (1992). Organization and Management. *Englewood Cliffs*.

Morin, A.J.S., & Marsh, H.W. (2013). Disentangling Shape from Levels Effects in Person- Centered Analyses: An Illustration Based University Teacher Multidimensional Profiles of Effectiveness. *Structural Equation Modeling*.

Murnane, R., & Cohen, D. (1986). Merit pay and the evaluation problem: Why most merit pay plans fail and a few survive. *Harvard educational review*, *56*(1), 1-18.

Muthén, B. (2004). Latent variable analysis. *The Sage handbook of quantitative methodology for the social sciences. Thousand Oaks, CA: Sage Publications*, 345-68.

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural equation modeling*, *14*(4), 535-569.

Papay, J. P. (2012). Refocusing the debate: Assessing the purposes and tools of teacher evaluation. *Harvard Educational Review*, *82*(1), 123-141.

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

Papay, J. P., Taylor, E. S., Tyler, J., & Laski, M. (2015). *Learning Job Skills from Colleagues at Work: Evidence from a Field Experiment Using Teacher Performance Data*. NBER Working Paper.

Pianta, R. C., & Hamre, B. K. (2015). Implementing Rigorous Observation of Teachers. *Improving Teacher Evaluation Systems: Making the Most of Multiple Measures*, 22.

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher, 38*(2), 109.

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). Classroom assessment scoring system. Baltimore: Paul H. Brookes.

Pianta, R. C. (2012). Implementing Observation Protocols: Lessons for K-12 Education from the Field of Early Childhood. *Center for American Progress*.

Pianta, R. (2011). *Teaching children well: New evidence-based approaches to teacher professional development and training*. Washington D.C.: Center for American Progress.

Ponitz, C. C., Rimm-Kaufman, S. E., Grimm, K. J., & Curby, T. W. (2009). Kindergarten classroom quality, behavioral engagement, and reading achievement. *School Psychology Review*, **38**, 102–120.

Reardon, S. F., & Robinson, J. P. (2012). Regression discontinuity designs with multiple rating-score variables. *Journal of Research on Educational Effectiveness*, *5*(1), 83-104.

Rice, J.K. (2009). Investing in human capital through teacher professional development. In D. Goldhaber & J. Hannaway (Eds.), *Creating a new teaching profession*. Washington, DC: Urban Institute.

Rice, J. K., Malen, B., Jackson, C., & Hoyer, K. M. (2015). Time to Pay Up Analyzing the Motivational Potential of Financial Awards in a TIF Program. *Educational Evaluation and Policy Analysis*, *37*(1), 29-49.

Rimm-Kaufman, S. E., Curby, T. W., Grimm, K., Nathanson, L., & Brock, L. L. (2009). The contribution of children's self-regulation and classroom quality to children's adaptive behaviors in the kindergarten classroom. *Developmental Psychology*, **45**, 958 –972.

Rivkin, S., E. Hanushek, and J. Kain (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. *American Economic Review, 94*(2), 247-252.

Ryan R.M., Mims V., & Koestner R. 1983. Relation of reward contingency and interpersonal context to intrinsic motivation: a review and test using cognitive evaluation theory. *Journal of Personality and Social Psychology* 45(4):736–50

Rynes, S. L., Gerhart, B. & Parks, L. (2005). Performance Evaluation and Pay for Performance. *Annual Review of Psychology* 56 (February): 571-600.

Ruzek, E. A., Domina, T., Conley, A. M., Duncan, G. J., & Karabenick, S. A. (2014). Using value-added models to measure teacher effects on students' motivation and achievement. *The Journal of Early Adolescence*, 0272431614525260.

Rothstein, J. (2009). *Student sorting and bias in value added estimation: Selection on observables and unobservables* (No. w14666). National Bureau of Economic Research.

Rothstein, J. (2008). *Teacher quality in educational production: Tracking, decay, and student achievement* (No. w14442). National Bureau of Economic Research.

Sanders, W. L., & Horn, S. P. (1996). Research findings from the Tennessee Value-Added Assessment System (TVAAS) database: Implications for educational evaluation and research. *Journal of Personnel Evaluation in Education*, *12*(3), 247-256.

Sartain, L., Stoelinga, S. R., & Brown, E. R. (2011). *Rethinking Teacher Evaluation in Chicago: Lessons Learned from Classroom Observations, Principal-Teacher Conferences, and District Implementation. Research Report*. Consortium on Chicago School Research.

Sass, T. R., Hannaway, J., Xu, Z., Figlio, D. N., & Feng, L. (2012). Value added of teachers in high-poverty schools and lower poverty schools. *Journal of Urban Economics*, *72*(2), 104-122.

Sawchuk, S. (2015, October 7). Teacher Evaluation Heads to the Courts. *Education Week*. Retrieved online November 20, 2015 from: http://www.edweek.org/ew/section/multimedia/teacher-evaluation-heads-to-the-courts.html

Shearer, B. (2004). Piece Rates, Fixed Wages and Incentives: Evidence from a Field Experiment. *Review of Economic Studies*, 71(2), 513-534.

Shukla, K.(2015). Applications of Latent Class Models: Profiles of School Climate and Invalid Respondents in Self-reports. Retrieved from http://libra.virginia.edu/catalog/libra- oa:9182

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher, 15*(2), 4-14.

Springer, M., Ballou, D., Hamilton, L., Le, V., Lockwood, J.R., McCaffrey, D., Pepper, M., Stecher, B. (2010). Teacher Pay for Performance, Experimental Evidence from the Project on Incentives in Teaching, Nashville, TN: National Center on Performance Incentives at Vanderbilt University.

Springer, M., Pane, J., Le, V., McCaffrey, D., Burns, S., Hamilton, L., & Stecher, B. (2012). Team pay for performance: Experimental evidence from the round rock pilot project on team incentives. Educational Evaluation and Policy Analysis, 34, 367–390.

Steinberg, M.P. & Donaldson, M. (2015). The new educational accountability: Understanding the landscape of teacher evaluation in the post-NCLB era. *Education Finance and Policy*.

Steinberg, M. & Sartain, L. (2015). Does teacher evaluation improve school performance? Experimental evidence from Chicago's Excellence in Teaching Project. *Education Finance and Policy*.

Taylor, E.S. and Tyler, J.H. (2012). The effect of evaluation on teacher performance. *American Economic Review*, 102(7), pp. 3628-3651.

TNTP (2015). The Mirage: Confronting the Hard Truth About Our Quest for Teacher Development. Brooklyn, NY. Retrieved October 15, 2015 from: http://tntp.org/assets/documents/TNTP-Mirage_2015.pdf

Tschannen-Moran, M., & Hoy, A. W. (2001). Teacher efficacy: Capturing an elusive construct. *Teaching and Teacher Education, 17*(7), 783-805.

Tyler, J. H., Taylor, E. S., Kane, T. J., & Wooten, A. L. (2010). Using student performance data to identify effective classroom practices. *The American Economic Review*, 256-260.

Uebersax, J. S. (1999). Probit latent class analysis with dichotomous or ordered category measures: conditional independence/dependence models. *Applied Psychological Measurement*, *23*(4), 283-297.

Umut, Ö. (2014). A Closer Look at the Student Achievement Trends in the District of Columbia between 2006-07 and 2012-13. CALDER Working Paper 119.

U.S. Department of Education (2009). Race to the Top Program Executive Summary. Washington D.C. Retrieved 10/31/2014 from: http://www2.ed.gov/programs/racetothetop/executive-summary.pdf

Vermunt, J. K., & Magidson, J. (2002). Latent class cluster analysis. *Applied latent class analysis*, *11*, 89-106.

Wayne, A. J., & Youngs, P. (2003). Teacher characteristics and student achievement gains: A review. *Review of Educational research*, *73*(1), 89-122.

Weingarten, R. (2010). A New Path Forward: Four Approaches to Quality Teaching and Better Schools. *American Educator*, *34*(1), 36-39.

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). The Widget Effect. *Brooklyn, NY: The New Teacher Project*.

White, M., Rowan, B., Alter, G., & Greene, C. (2014). User guide to the measures of effective teaching longitudinal database (MET LDB). *Ann Arbor, MI: Inter-University Consortium for Political and Social Research, The University of Michigan*.

Whitehurst, G. J., Chingos, M. M., & Lindquist, K. M. (2014). Evaluating teachers with classroom observations: Lessons learned in four districts. Brown Center on Education Policy at Brookings Policy Report.

Wigfield, A., & Eccles, J. S. (1992). The development of achievement task values: A theoretical analysis. *Developmental review*, *12*(3), 265-310.

-- (2000). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology*, *25*, 68-81.

Wiggins, G. P., & McTighe, J. (2005). *Understanding by design*. Ascd.

Wong, V. C., Steiner, P. M., & Cook, T. D. (2013). Analyzing regression-discontinuity designs with multiple assignment variables a comparative study of four estimation methods. *Journal of Educational and Behavioral Statistics*, *38*(2), 107-141.

Xu, Z., Özek, U., & Hansen, M. (2014). Teacher Performance Trajectories in High-and Lower-Poverty Schools. *Educational Evaluation and Policy Analysis*.

Yoon, K. S., Duncan, T., Lee, S. W. Y., Scarloss, B., & Shapley, K. L. (2007). Reviewing the Evidence on How Teacher Professional Development Affects Student Achievement. Issues & Answers. REL 2007-No. 033. *Regional Educational Laboratory Southwest (NJ1)*.

Yuan, K., Le, V. N., McCaffrey, D. F., Marsh, J. A., Hamilton, L. S., Stecher, B. M., & Springer, M. G. (2013). Incentive Pay Programs Do Not Affect Teacher Motivation or Reported Practices Results From Three Randomized Studies. *Educational Evaluation and Policy Analysis*, *35*(1), 3-22.

APPENDIX TABLES AND FIGURES

Table A1.1. Descriptions of Effective Practice in the Teaching and Learning Framework

Teach Standard	Description for Effective Rating (3 out of 4)
Teach 1: Lead well-organized, objective-driven lessons	• The lesson objective is specific, measurable, and aligned to standards; it conveys what students are learning and what they will be able to do by the end of the lesson. • The objective of the lesson is clear to students. For example, the teacher might clearly state and explain the objective, or students might demonstrate through their actions that they understand what they will be learning and doing. • The teacher ensures that students understand the importance of the objective. For example, the teacher might effectively explain its importance, or students might demonstrate through their comments that they understand the importance of what they are learning. • The lesson builds on students' prior knowledge in a significant and meaningful way, as appropriate to the objective. • The lesson is well-organized: All parts of the lesson are connected to each other and aligned to the objective, and each part significantly moves students toward mastery of the objective.
Teach 2: Explain Content Clearly	• Explanations of content are clear and coherent, and they build student understanding of content. • The teacher uses developmentally appropriate language and explanations. • The teacher gives clear, precise definitions and uses specific academic language as appropriate. • The teacher emphasizes key points when necessary. • When an explanation is not effectively leading students to understand the content, the teacher adjusts quickly and uses an alternative way to effectively explain the concept. • Students ask relatively few clarifying questions because they understand the explanations. However, they may ask a number of extension questions because they are engaged in the content and eager to learn more about it.
Teach 3: Engage students at all learning levels in rigorous work	• The teacher makes the lesson accessible to almost all students; there is evidence that the teacher knows each student's level and ensures that the lesson meets almost all of students where they are. For example, if necessary, the teacher might differentiate content, process, or product (using strategies that might include, for example, flexible grouping, leveled texts, or tiered assignments) in order to ensure that students are able to access the lesson. • The teacher makes the lesson challenging to almost all students; there is evidence that the teacher knows each student's level and ensures that the lesson pushes almost all students forward from where they are. For example, the teacher might

	ask more challenging questions, assign more demanding work, or provide extension assignments in order to ensure that all students are challenged by the lesson. • There is an appropriate balance between teacher-directed instruction and rigorous student-centered learning during the lesson, such that students have adequate opportunities to meaningfully practice, apply, and demonstrate what they are learning.
Teach 4: Provide students multiple ways to engage with content	• The teacher provides students more than one way to engage with content, as appropriate, and all ways are matches to the lesson objective. For particular types of lessons, this may only entail giving students two ways to engage with content (for example, a Socratic seminar might involve verbal/linguistic and interpersonal ways), while for many lessons, this may involve three or more. • The ways students engage with content all promote student mastery of the objective.
Teach 5: Check for student understanding	• The teacher checks for understanding of content at almost all key moments (when checking is necessary to inform instruction going forward, such as before moving on to the next step of the lesson or partway through the independent practice). • The teacher gets an accurate "pulse" of the class's understanding from almost every check, such that the teacher has enough information to adjust subsequent instruction if necessary. • If a check reveals a need to make a whole-class adjustment to the lesson plan (for example, because most of the students did not understand a concept just taught), the teacher makes the appropriate adjustment in an effective way.
Teach 6: Respond to student misunderstandings	• The teacher responds to most student misunderstanding with effective scaffolding. • When possible, the teacher uses scaffolding techniques that enable students to construct their own understandings (for example, by asking leading questions) rather than simply re-explaining a concept. • If an attempt to address a misunderstanding is not succeeding, the teacher, when appropriate, responds with another way of scaffolding.
Teach 7: Develop higher-level understanding through effective questioning	• The teacher frequently develops higher-level understanding through effective questioning. • Nearly all of the questions used are effective in developing higher-level understanding. • The teacher uses a variety of questions.

Teach 8: Maximize instructional time	• Routines and procedures run smoothly with some prompting from the teacher; students generally know their responsibilities. Transitions are generally smooth with some teacher direction. • Students are only idle for very brief periods of time while waiting for the teacher (for example, while the teacher takes attendance or prepares materials). • The teacher spends an appropriate amount of time on each part of the lesson. • The lesson progresses at a quick pace, such that students are almost never disengaged or left with nothing meaningful to do (for example, after finishing the assigned work, or while waiting for one student to complete a problem in front of the class). • Inappropriate or off-task student behavior rarely interrupts or delays the lesson.
Teach 9: Build a supportive, learning-focused classroom community	• Students are invested in their work and value academic success. For example, students work hard, remain focused on learning without frequent reminders, and persevere through challenges. • The classroom is a safe environment for students to take on challenges and risk failure. For example, students are eager to answer questions, feel comfortable asking the teacher for help, and do not respond negatively when a peer answers a question incorrectly. • Students are always respectful of the teacher and their peers. For example, students listen and do note interrupt when their peers ask or answer questions. • The teacher meaningfully reinforces positive behavior and good academic work as appropriate. • The teacher has a positive rapport with students, as demonstrated by displays of positive affect, evidence of relationship building, and expressions of interest in students’ thoughts and opinions.

Table A1.2 Auxiliary RD, Teacher Covariate Balance in Year *t*

	Minimally Effective Threshold				Highly Effective Threshold			
	All yrs	2010-11	2011-12	2012-13	All yrs	2010-11	2011-12	2012-13
White	-0.019	-0.002	-0.016	-0.040	0.087***	0.102**	0.043	0.060
	(0.026)	(0.041)	(0.048)	(0.049)	(0.029)	(0.044)	(0.063)	(0.055)
Black	-0.029	-0.059	-0.022	0.025	-0.096***	-0.097**	-0.047	-0.097*
	(0.030)	(0.049)	(0.054)	(0.060)	(0.029)	(0.045)	(0.061)	(0.054)
Female	-0.022	-0.007	-0.029	0.028	0.016	-0.035	0.106*	0.027
	(0.029)	(0.048)	(0.049)	(0.058)	(0.026)	(0.041)	(0.060)	(0.047)
Group 1	0.014	-0.065	0.007	0.125**	0.006	-0.001	-0.026	0.000
	(0.026)	(0.040)	(0.046)	(0.056)	(0.022)	(0.031)	(0.039)	(0.048)
Exp 0-1	-0.018	-0.074*	0.037	-0.028	0.010	-0.006	0.064	0.004
	(0.027)	(0.042)	(0.049)	(0.054)	(0.021)	(0.031)	(0.055)	(0.041)
Exp 2-4	-0.010	0.042	-0.047	-0.044	0.014	-0.007	-0.033	0.066
	(0.022)	(0.034)	(0.039)	(0.045)	(0.027)	(0.037)	(0.062)	(0.054)
Exp 5-9	0.031	0.015	0.063	0.029	0.010	0.021	-0.080	0.016
	(0.024)	(0.039)	(0.041)	(0.048)	(0.027)	(0.043)	(0.054)	(0.050)
Exp 10-14	0.020	0.005	0.020	0.054	-0.008	-0.017	0.021	-0.020
	(0.020)	(0.030)	(0.036)	(0.045)	(0.023)	(0.033)	(0.053)	(0.041)
Exp 15-19	0.013	0.014	0.000	0.027	-0.023	-0.040	0.075	-0.044
	(0.020)	(0.032)	(0.035)	(0.042)	(0.021)	(0.034)	(0.049)	(0.031)

Notes: Each cell reports the results of a separate regression with the indicated dependent variable. Results condition on a smooth linear spline function of centered initial IMPACT score. Results also condition on teacher covariates and school fixed effects. Robust standard errors in parentheses.
***p<0.01, **p<0.05, *p<0.1.

Table A1.3. Reduced-form RD Estimates at Minimally Effective Threshold, Teach Standards (All Rater Averages)

	All years		2010-11		2011-12		2012-13	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
TLF 1	0.033	-0.103	-0.076	-0.376***	0.109	0.182	0.196	-0.011
	(0.065)	(0.092)	(0.099)	(0.137)	(0.108)	(0.153)	(0.142)	(0.202)
TLF 2	0.090	-0.006	-0.138	-0.270*	0.206*	0.192	0.380***	0.247
	(0.066)	(0.092)	(0.101)	(0.140)	(0.112)	(0.161)	(0.137)	(0.187)
TLF 3	0.101	0.050	-0.017	-0.173	0.260**	0.381**	0.216	0.097
	(0.065)	(0.093)	(0.100)	(0.141)	(0.116)	(0.165)	(0.138)	(0.190)
TLF 4	0.095	0.085	0.011	-0.086	0.173	0.333**	0.172	0.099
	(0.065)	(0.090)	(0.099)	(0.135)	(0.118)	(0.167)	(0.133)	(0.178)
TLF 5	0.117*	0.049	-0.138	-0.267*	0.228**	0.449***	0.488***	0.176
	(0.065)	(0.094)	(0.098)	(0.141)	(0.111)	(0.160)	(0.135)	(0.188)
TLF 6	0.075	0.037	-0.040	-0.126	0.055	0.183	0.329**	0.180
	(0.071)	(0.102)	(0.107)	(0.147)	(0.133)	(0.192)	(0.143)	(0.207)
TLF 7	0.045	-0.006	-0.109	-0.286**	0.096	0.281*	0.322**	0.136
	(0.064)	(0.091)	(0.095)	(0.136)	(0.114)	(0.161)	(0.138)	(0.191)
TLF 8	0.044	-0.064	-0.119	-0.309**	0.197	0.296*	0.149	0.001
	(0.068)	(0.096)	(0.102)	(0.145)	(0.121)	(0.168)	(0.148)	(0.207)
TLF 9	0.045	0.021	-0.138	-0.236	0.229*	0.281*	0.213	0.209
	(0.070)	(0.097)	(0.107)	(0.153)	(0.117)	(0.157)	(0.154)	(0.210)
Linear spline	X		X		X		X	
Quadratic spline		X		X		X		X

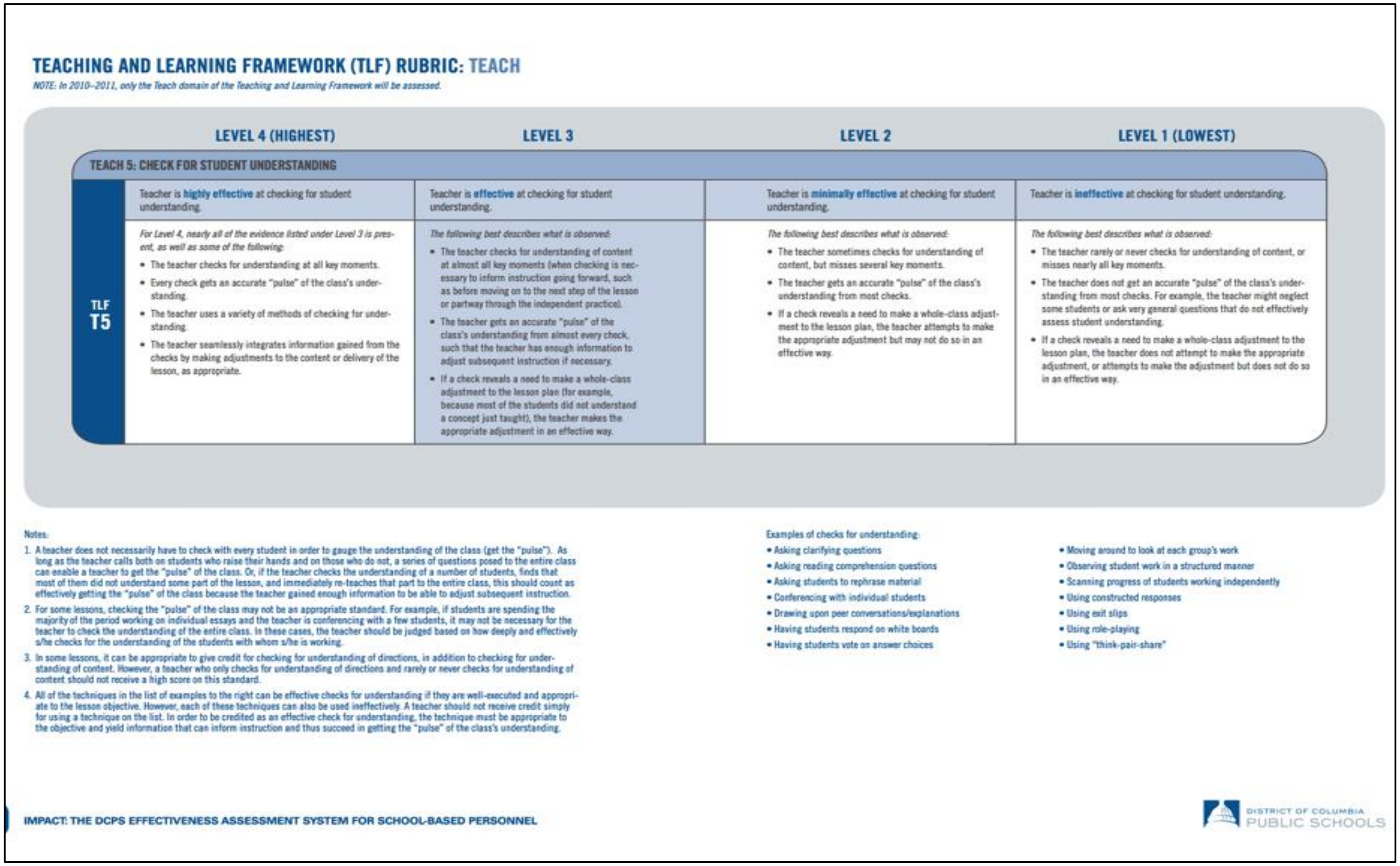
Notes: Each cell reports the results of a separate regression with the indicated dependent variable. Results condition on a smooth function of centered initial IMPACT score, specifications with a linear spline and quadratic spline are both shown above. Results also condition on teacher covariates and school fixed effects. Robust standard errors in parentheses. ***p<0.01, **p<0.05, *p<0.1

Table A1.4 Reduced-form RD Estimates at Highly Effective Threshold, TEACH Standards (All Rater Averages)

	All years		2010-11		2011-12		2012-13	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
TLF1	0.178***	0.046	0.359***	0.136	-0.027	-0.051	0.118	0.024
	(0.054)	(0.074)	(0.074)	(0.101)	(0.127)	(0.167)	(0.100)	(0.146)
TLF 2	0.157***	0.025	0.246***	0.075	0.091	0.045	0.114	-0.035
	(0.051)	(0.070)	(0.070)	(0.096)	(0.116)	(0.156)	(0.101)	(0.146)
TLF 3	0.089*	-0.005	0.181**	0.044	-0.068	-0.079	0.139	0.016
	(0.054)	(0.074)	(0.078)	(0.107)	(0.116)	(0.156)	(0.106)	(0.151)
TLF 4	0.038	-0.030	0.179***	0.069	-0.215**	-0.284**	0.062	-0.006
	(0.048)	(0.065)	(0.065)	(0.088)	(0.108)	(0.135)	(0.103)	(0.149)
TLF 5	0.053	-0.088	0.170**	0.033	-0.195	-0.266	0.060	-0.179
	(0.053)	(0.073)	(0.073)	(0.101)	(0.132)	(0.164)	(0.099)	(0.141)
TLF 6	0.085	-0.057	0.116	-0.064	-0.082	-0.174	0.193*	0.051
	(0.056)	(0.077)	(0.080)	(0.112)	(0.140)	(0.184)	(0.104)	(0.148)
TLF 7	0.119**	0.005	0.233***	0.041	0.097	0.225	0.069	-0.165
	(0.053)	(0.073)	(0.082)	(0.111)	(0.108)	(0.138)	(0.097)	(0.136)
TLF 8	0.107**	0.032	0.256***	0.170*	-0.053	-0.059	0.061	-0.086
	(0.049)	(0.069)	(0.068)	(0.093)	(0.112)	(0.150)	(0.091)	(0.135)
TLF 9	0.070	0.048	0.187***	0.177**	0.065	0.120	-0.018	-0.128
	(0.047)	(0.063)	(0.064)	(0.087)	(0.103)	(0.140)	(0.092)	(0.127)
Linear spline	X		X		X		X	
Quadratic spline		X		X		X		X

Notes: Each cell reports the results of a separate regression with the indicated dependent variable. Results condition on a smooth function of centered initial IMPACT score, specifications with a linear spline and quadratic spline are both shown above. Results also condition on teacher covariates and school fixed effects. Robust standard errors in parentheses. ***p<0.01, **p<0.05, p<0.1.

Figure A1.1. Example of Teaching and Learning Framework Standard, Teach 5



TEACHING AND LEARNING FRAMEWORK (TLF) RUBRIC: TEACH

NOTE: In 2010–2011, only the Teach domain of the Teaching and Learning Framework will be assessed.

	LEVEL 4 (HIGHEST)	LEVEL 3	LEVEL 2	LEVEL 1 (LOWEST)
TEACH 5: CHECK FOR STUDENT UNDERSTANDING				
TLF T5	Teacher is **highly effective** at checking for student understanding.	Teacher is **effective** at checking for student understanding.	Teacher is **minimally effective** at checking for student understanding.	Teacher is **ineffective** at checking for student understanding.
	For Level 4, nearly all of the evidence listed under Level 3 is present, as well as some of the following: • The teacher checks for understanding at all key moments. • Every check gets an accurate "pulse" of the class's understanding. • The teacher uses a variety of methods of checking for understanding. • The teacher seamlessly integrates information gained from the checks by making adjustments to the content or delivery of the lesson, as appropriate.	*The following best describes what is observed:* • The teacher checks for understanding of content at almost all key moments (when checking is necessary to inform instruction going forward, such as before moving on to the next step of the lesson or partway through the independent practice). • The teacher gets an accurate "pulse" of the class's understanding from almost every check, such that the teacher has enough information to adjust subsequent instruction if necessary. • If a check reveals a need to make a whole-class adjustment to the lesson plan (for example, because most of the students did not understand a concept just taught), the teacher makes the appropriate adjustment in an effective way.	*The following best describes what is observed:* • The teacher sometimes checks for understanding of content, but misses several key moments. • The teacher gets an accurate "pulse" of the class's understanding from most checks. • If a check reveals a need to make a whole-class adjustment to the lesson plan, the teacher attempts to make the appropriate adjustment but may not do so in an effective way.	*The following best describes what is observed:* • The teacher rarely or never checks for understanding of content, or misses nearly all key moments. • The teacher does not get an accurate "pulse" of the class's understanding from most checks. For example, the teacher might neglect some students or ask very general questions that do not effectively assess student understanding. • If a check reveals a need to make a whole-class adjustment to the lesson plan, the teacher does not attempt to make the appropriate adjustment, or attempts to make the adjustment but does not do so in an effective way.

Notes:

1. A teacher does not necessarily have to check with every student in order to gauge the understanding of the class (get the "pulse"). As long as the teacher calls both on students who raise their hands and on those who do not, a series of questions posed to the entire class can enable a teacher to get the "pulse" of the class. Or, if the teacher checks the understanding of a number of students, finds that most of them did not understand some part of the lesson, and immediately re-teaches that part to the entire class, this should count as effectively getting the "pulse" of the class because the teacher gained enough information to be able to adjust subsequent instruction.
2. For some lessons, checking the "pulse" of the class may not be an appropriate standard. For example, if students are spending the majority of the period working on individual essays and the teacher is conferencing with a few students, it may not be necessary for the teacher to check the understanding of the entire class. In these cases, the teacher should be judged based on how deeply and effectively s/he checks for the understanding of the students with whom s/he is working.
3. In some lessons, it can be appropriate to give credit for checking for understanding of directions, in addition to checking for understanding of content. However, a teacher who only checks for understanding of directions and rarely or never checks for understanding of content should not receive a high score on this standard.
4. All of the techniques in the list of examples to the right can be effective checks for understanding if they are well-executed and appropriate to the lesson objective. However, each of these techniques can also be used ineffectively. A teacher should not receive credit simply for using a technique on the list. In order to be credited as an effective check for understanding, the technique must be appropriate to the objective and yield information that can inform instruction and thus succeed in getting the "pulse" of the class's understanding.

Examples of checks for understanding:

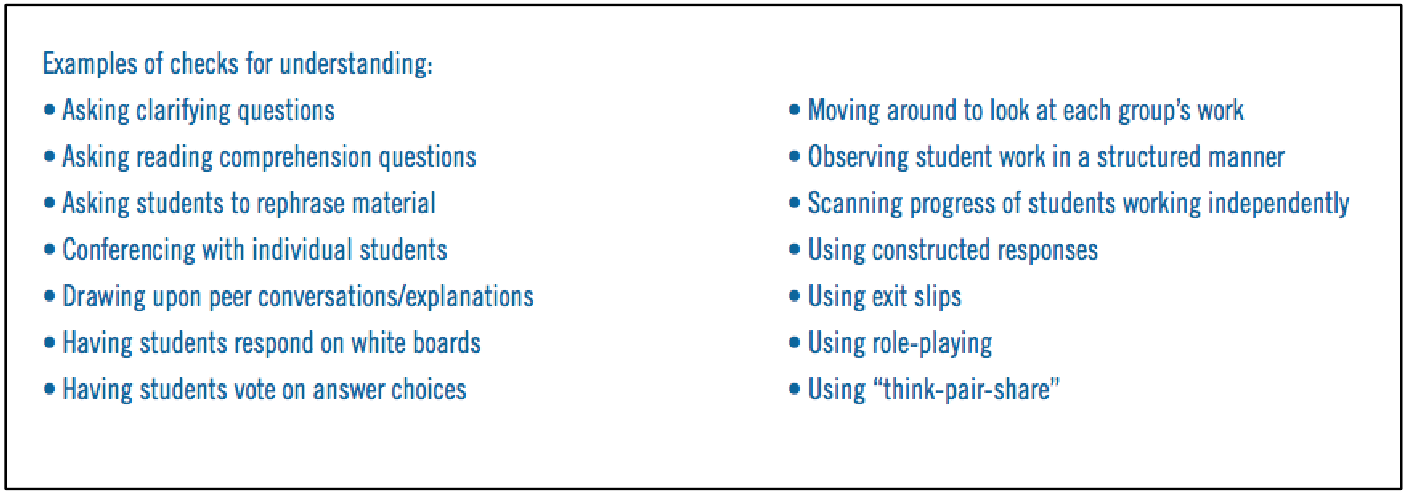
- Asking clarifying questions
- Asking reading comprehension questions
- Asking students to rephrase material
- Conferencing with individual students
- Drawing upon peer conversations/explanations
- Having students respond on white boards
- Having students vote on answer choices
- Moving around to look at each group's work
- Observing student work in a structured manner
- Scanning progress of students working independently
- Using constructed responses
- Using exit slips
- Using role-playing
- Using "think-pair-share"

IMPACT: THE DCPS EFFECTIVENESS ASSESSMENT SYSTEM FOR SCHOOL-BASED PERSONNEL

DISTRICT OF COLUMBIA PUBLIC SCHOOLS

Source: DCPS IMPACT Guidebook for Group 1 Teachers, 2010-11.

Figure A1.2. Example of Strategy List from Teaching and Learning Framework Standard, Teach 5

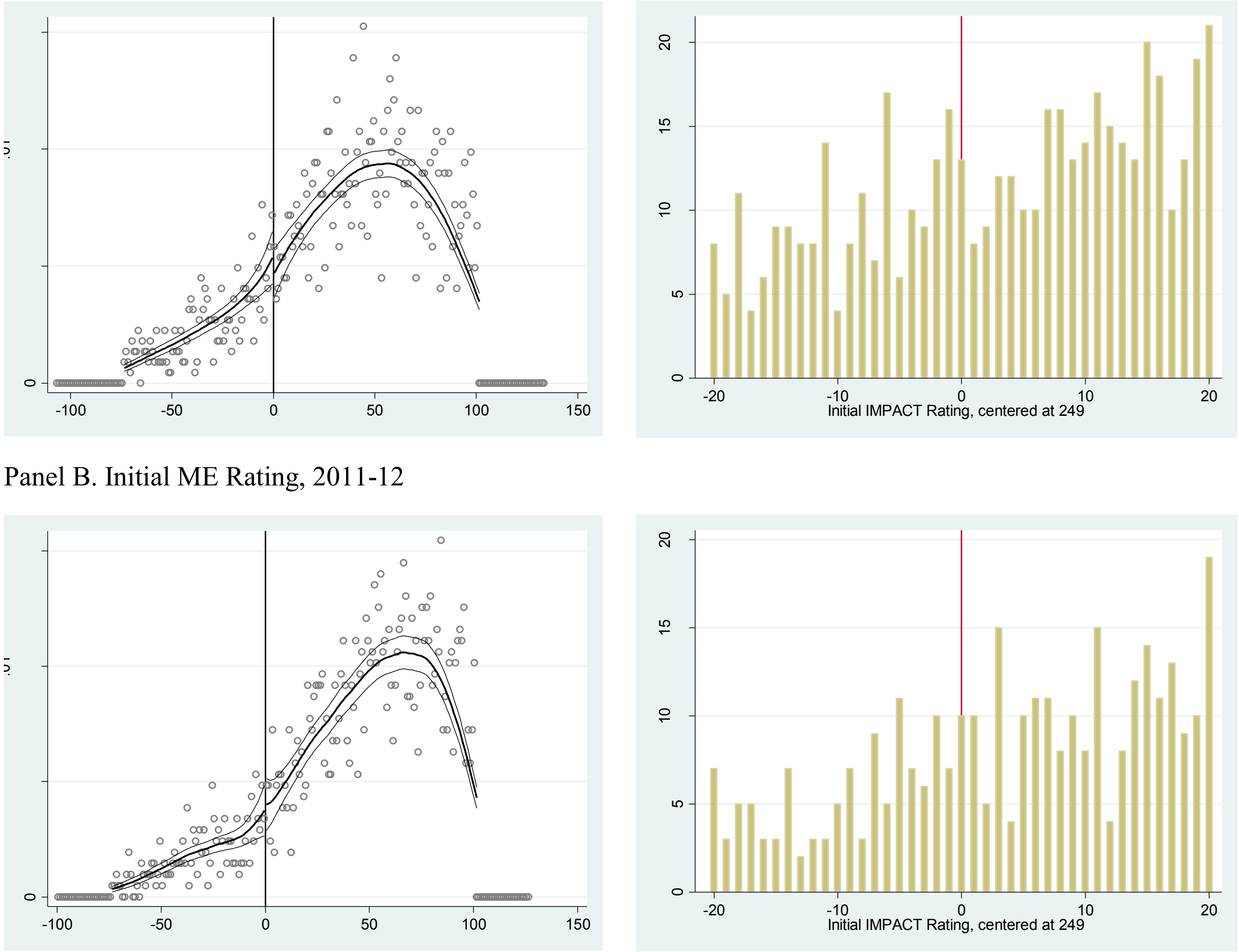


Examples of checks for understanding:

- Asking clarifying questions
- Asking reading comprehension questions
- Asking students to rephrase material
- Conferencing with individual students
- Drawing upon peer conversations/explanations
- Having students respond on white boards
- Having students vote on answer choices
- Moving around to look at each group's work
- Observing student work in a structured manner
- Scanning progress of students working independently
- Using constructed responses
- Using exit slips
- Using role-playing
- Using "think-pair-share"

Source: DCPS IMPACT Guidebook for Group 1 Teachers, 2010-11.

Figure A1.3. Density Check for Assignment Variable, Initial IMPACT Rating



Panel A. Initial ME Rating, 2010-11

Panel B. Initial ME Rating, 2011-12

Panel C. Initial HE Rating, 2010-11

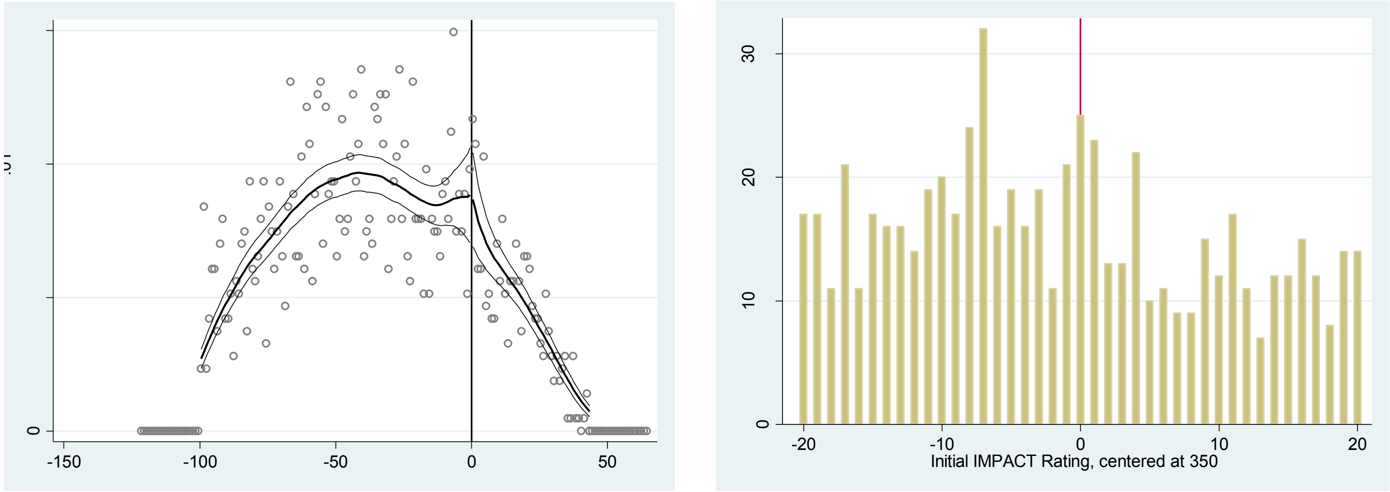


Table A2.1. First-stage, Intent-To-Treat at Minimally Effective Performance Threshold

		n	(1)	(2)	(3)
Math	2009-10	7,193	0.954*** (0.032)	0.983*** (0.017)	0.987*** (0.014)
	2010-11	6,084	1.0*** (0)	1.0*** (0)	1.0*** (0)
	2012-13	4,682	1.0*** (0)	1.0*** (0)	1.0*** (0)
Reading	2009-10	7,053	0.818*** (0.127)	0.898*** (0.070)	0.889*** (0.068)
	2011-12	6,908	1.0*** (0)	1.0*** (0)	1.0*** (0)
	2012-13	5,387	1.0*** (0)	1.0*** (0)	1.0*** (0)
School fixed effects				X	X
Grade fixed effects					X

Notes: Robust standard errors clustered by teacher in parentheses. All models condition on a linear spline function of the assignment variable. *** p<0.01, ** p<0.05 * p<0.1

Table A2.2. Auxiliary RD Examining Teacher Covariate Balance at Minimally Effective Threshold in Year *t*

		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	n	female	black	white	grad	exp	exp	exp	exp	group 1
			tchr	tchr	degree	0-1	2-4	5-8	10-14	t+1
2010-11	426	-0.064	0.033	-0.040	0.001	-0.068	0.005	-0.005	0.043	0.045
		(0.106)	(0.107)	(0.100)	(0.111)	(0.099)	(0.084)	(0.084)	(0.075)	(0.121)
2011-12	441	0.042	-0.051	-0.040	0.072	0.057	-0.093	0.053	0.064	0.051
		(0.089)	(0.100)	(0.095)	(0.097)	(0.089)	(0.080)	(0.087)	(0.066)	(0.116)
2012-13	358	-0.012	0.193	-0.146	-0.038	-0.088	-0.043	0.132	-0.045	-0.020
		(0.104)	(0.118)	(0.109)	(0.121)	(0.108)	(0.101)	(0.092)	(0.091)	(0.134)

Notes: Robust standard errors in parentheses. Results condition on a linear spline of the assignment variable and school fixed effects. ***p<0.01, **p<0.05, *p<0.1

Table A2.3. Auxiliary RD, Propensity to Remain in Group 1 in Year *t+1*

		n	(1)	(2)	(3)	(4)
Subjects stacked	2010-11	14,180	-0.004	-0.018	-0.012	0.027
			(0.050)	(0.051)	(0.048)	(0.071)
			-45580	-50129	-50661	-50741
	2011-12	12,881	0.069*	0.119***	0.116**	0.063
			(0.039)	(0.045)	(0.045)	(0.058)
			-48766	-54379	-54520	-55204
	2012-13	9,940	-0.071	-0.011	-0.023	-0.137*
			(0.075)	(0.034)	(0.037)	(0.073)
			-31847	-38543	-39470	-39924
Math	2010-11	7,160	-0.054	-0.029	-0.015	0.054
			(0.063)	(0.073)	(0.067)	(0.095)
			-23095	-25827	-26125	-26182
	2011-12	6,084	0.072*	0.202***	0.200***	0.126
			(0.042)	(0.071)	(0.073)	(0.090)
			-22846	-25638	-25724	-26482
	2012-13	4,643	-0.054	-0.042	-0.085	-0.188**
			(0.080)	(0.044)	(0.053)	(0.086)
			-14573	-17766	-18485	-18774
ELA	2010-11	7,020	0.037	0.019	0.026	0.074
			(0.047)	(0.057)	(0.055)	(0.078)
			-22529	-24810	-25053	-25101
	2011-12	6,797	0.067*	0.117**	0.117**	0.095
			(0.038)	(0.047)	(0.047)	(0.068)
			-25927	-29254	-29324	-29566
	2012-13	5,297	-0.088	-0.019	-0.018	-0.160
			(0.083)	(0.048)	(0.050)	(0.104)
			-17294	-21208	-21485	-21739
School fixed effects				X	X	X
Grade fixed effects					X	X
Quadratic rating variable						X

Notes: Robust standard errors clustered by teacher in parentheses. All models condition on a linear spline function of the assignment variable. *** p<0.01, ** p<0.05, * p<0.1

Table A2.4. Auxiliary RD, Number of Rostered Students in Year *t+1*

		n	(1)	(2)	(3)	(4)
Subjects stacked	2010-11	14,246	-13.648	-4.979	-2.642	10.750*
			(9.032)	(4.949)	(4.301)	(5.605)
		AIC	98975	81965	78410	78051
	2011-12	12,992	-22.653**	-1.268	-4.628	-10.452*
			(9.791)	(4.996)	(4.538)	(5.999)
			92889	71613	69203	68964
	2012-13	10,069	11.227	-4.965	-6.959	-16.300**
			(11.918)	(5.866)	(5.186)	(8.066)
			67098	53158	47147	46847
Math	2010-11	7,193	-33.206***	-19.134***	-14.429***	-0.735
			(9.488)	(5.470)	(5.387)	(6.146)
			49940	41044	39564	39446
	2011-12	6,084	-28.276**	4.608	-4.017	-9.605
			(13.339)	(6.959)	(5.203)	(7.012)
			43559	32247	30282	30203
	2012-13	4,682	24.125	1.739	**-4.110**	-5.025
			(17.441)	(6.436)	(4.732)	(6.815)
			30842	23868	19043	19044
ELA	2010-11	7,053	2.146	9.163	9.274*	19.578***
			(11.693)	(6.523)	(5.187)	(6.557)
			48800	39943	38077	37911
	2011-12	6,908	-17.452	-5.168	-8.019	-15.925*
			(13.137)	(6.399)	(6.635)	(9.167)
			49291	37583	37014	36873
	2012-13	5,387	1.940	-9.752	-11.745	-27.167***
			(9.380)	(9.164)	(8.679)	(10.337)
			35786	27377	25302	24875
School fixed effects				X	X	X
Grade fixed effects					X	X
Quadratic rating variable						X

Notes: Robust standard errors clustered by teacher in parentheses. All models condition on a linear spline function of the assignment variable. *** p<0.01, ** p<0.05, * p<0.1

Table A2.5. Auxiliary RD, Predicted Achievement Balance in Year *t+1*

		n	(1)	(2)	(3)	(4)
Subjects stacked	2010-11	14,246	0.012	-0.002	-0.004	0.038
			(0.044)	(0.035)	(0.035)	(0.056)
			-19884	-23747	-23750	-23750
	2011-12	12,992	0.065	-0.039	-0.030	0.019
			(0.087)	(0.032)	(0.031)	(0.035)
			-16280	-21241	-21245	-21248
	2012-13	10,069	0.109	0.025	0.021	-0.008
			(0.076)	(0.036)	(0.037)	(0.044)
			-12736	-15691	-15692	-15695
Math	2010-11	7,193	-0.014	-0.047	-0.055	-0.038
			(0.061)	(0.034)	(0.035)	(0.048)
			-10100	-12338	-12338	-12337
	2011-12	6,084	-0.013	-0.052	-0.028	0.021
			(0.083)	(0.038)	(0.033)	(0.043)
			-7652	-10288	-10290	-10288
	2012-13	4,682	0.120	0.065**	0.060*	0.020
			(0.127)	(0.030)	(0.034)	(0.035)
			-5722	-7452	-7446	-7445
ELA	2010-11	7,053	0.034	-0.023	-0.020	0.063
			(0.048)	(0.069)	(0.064)	(0.076)
			-9784	-11494	-11489	-11498
	2011-12	6,908	0.142	-0.031	-0.027	0.049
			(0.134)	(0.048)	(0.047)	(0.054)
			-8658	-11027	-11026	-11031
	2012-13	5,387	0.094	0.004	0.021	0.045
			(0.057)	(0.072)	(0.079)	(0.079)
			-7107	-8320	-8324	-8321
School fixed effects				X	X	X
Grade fixed effects					X	X
Quadratic rating variable						X

Notes: Robust standard errors clustered by teacher in parentheses. All models condition on a linear spline function of the assignment variable. *** p<0.01, ** p<0.05, * p<0.1

Table A2.6. Heterogeneity in Main Effects by Student Subgroup, Math

		(1) Spec. Ed.	(2) LEP	(3) FRPL	(4) Male	(5) White	(6) Black	(7) Hispanic
All years	ITT	0.018	0.036	0.039	0.068**	0.039	0.045	0.045
		(0.036)	(0.036)	(0.044)	(0.033)	(0.036)	(0.055)	(0.037)
	ITT * Student Char.	0.169**	0.084	0.005	-0.056*	0.081	0.004	0.004
		(0.075)	(0.099)	(0.039)	(0.032)	(0.127)	(0.064)	(0.066)
2010-11	ITT	-0.032	-0.022	-0.124	0.023	-0.011	0.161	-0.014
		(0.071)	(0.063)	(0.095)	(0.064)	(0.062)	(0.103)	(0.064)
	ITT * Student Char.	0.185	0.229	0.152*	-0.052	0.473***	-0.178*	0.170*
		(0.138)	(0.197)	(0.083)	(0.058)	(0.073)	(0.105)	(0.100)
2011-12	ITT	0.075	0.089	0.024	0.138*	0.096	0.069	0.094
		(0.075)	(0.075)	(0.078)	(0.073)	(0.074)	(0.081)	(0.077)
	ITT * Student Char.	0.130	0.082	0.089	-0.092*	-0.077	0.032	-0.007
		(0.098)	(0.112)	(0.067)	(0.053)	(0.459)	(0.091)	(0.099)
2012-13	ITT	0.203***	0.233***	0.248***	0.231***	0.237***	0.075	0.267***
		(0.046)	(0.049)	(0.058)	(0.052)	(0.046)	(0.072)	(0.056)
	ITT * Student Char.	0.262*	-0.102	-0.040	-0.013	-0.123	0.245***	-0.171*
		(0.143)	(0.170)	(0.047)	(0.051)	(0.141)	(0.081)	(0.100)

Notes: Robust standard errors clustered by teacher in parentheses. Results condition on a linear spline function of the assignment variable and school and grade fixed effects. ***p<0.01, **p<0.05, *p<0.1.

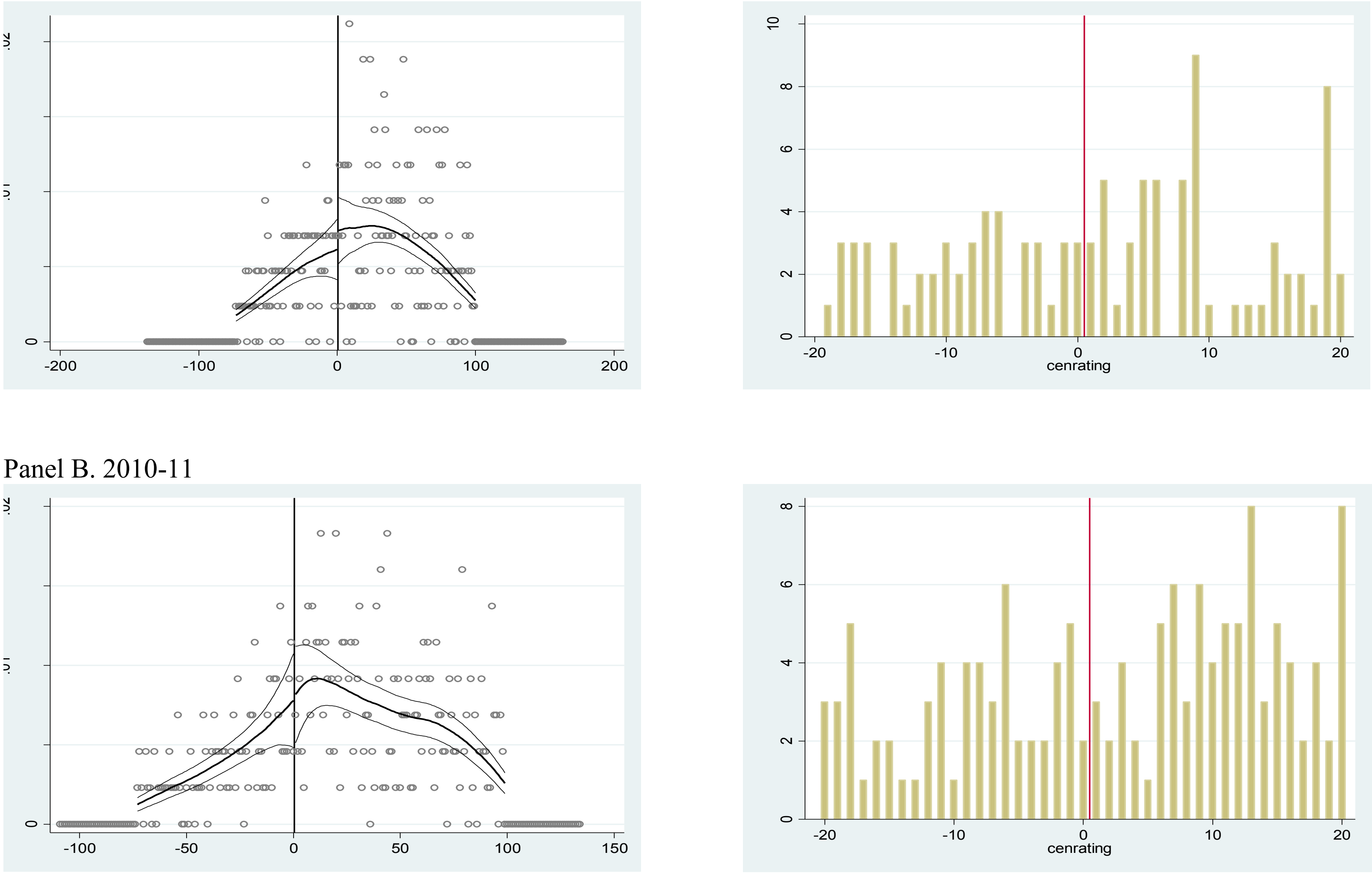
Table A2.7. Heterogeneity in Main Effects by Student Subgroup, Reading

		(1) Spec. Ed.	(2) LEP	(3) FRPL	(4) Male	(5) White	(6) Black	(7) Hispanic
All years	ITT	0.025 (0.028)	0.038 (0.027)	-0.018 (0.035)	0.028 (0.028)	0.044 (0.027)	0.049 (0.055)	0.035 (0.028)
	ITT * Student Char.	0.114* (0.060)	0.083 (0.086)	0.076** (0.034)	0.029 (0.029)	-0.043 (0.093)	-0.010 (0.058)	0.064 (0.048)
2010-11	ITT	-0.026 (0.057)	-0.037 (0.054)	-0.025 (0.085)	-0.027 (0.058)	-0.024 (0.055)	0.345*** (0.083)	-0.041 (0.055)
	ITT * Student Char.	0.032 (0.093)	0.548*** (0.202)	0.010 (0.070)	0.023 (0.050)	-0.149*** (0.038)	-0.394*** (0.083)	0.376*** -0.119
2011-12	ITT	0.034 (0.036)	0.039 (0.035)	-0.030 (0.049)	0.016 (0.038)	0.060 (0.038)	0.008 (0.054)	0.037 (0.036)
	ITT * Student Char.	0.124 (0.120)	0.150 (0.106)	0.125** (0.055)	0.064 (0.040)	-0.100** (0.050)	0.060 (0.061)	0.056 (0.041)
2012-13	ITT	-0.047 (0.054)	-0.004 (0.047)	-0.073 (0.068)	-0.037 (0.046)	-0.017 (0.046)	-0.093 (0.117)	-0.014 (0.046)
	ITT * Student Char.	0.163 (0.106)	-0.367** (0.176)	0.066 (0.052)	0.030 (0.050)	-0.548*** (0.034)	0.081 (0.106)	-0.049 (0.111)

Notes: Robust standard errors clustered by teacher in parentheses. Results condition on a linear spline function of the assignment variable and school and grade fixed effects. ***p<0.01, **p<0.05, *p<0.1.

Figure A2.1. Density check for Assignment Variable, Initial IMPACT Score

Panel A. 2009-10



Panel B. 2010-11

Panel C. 2011-12

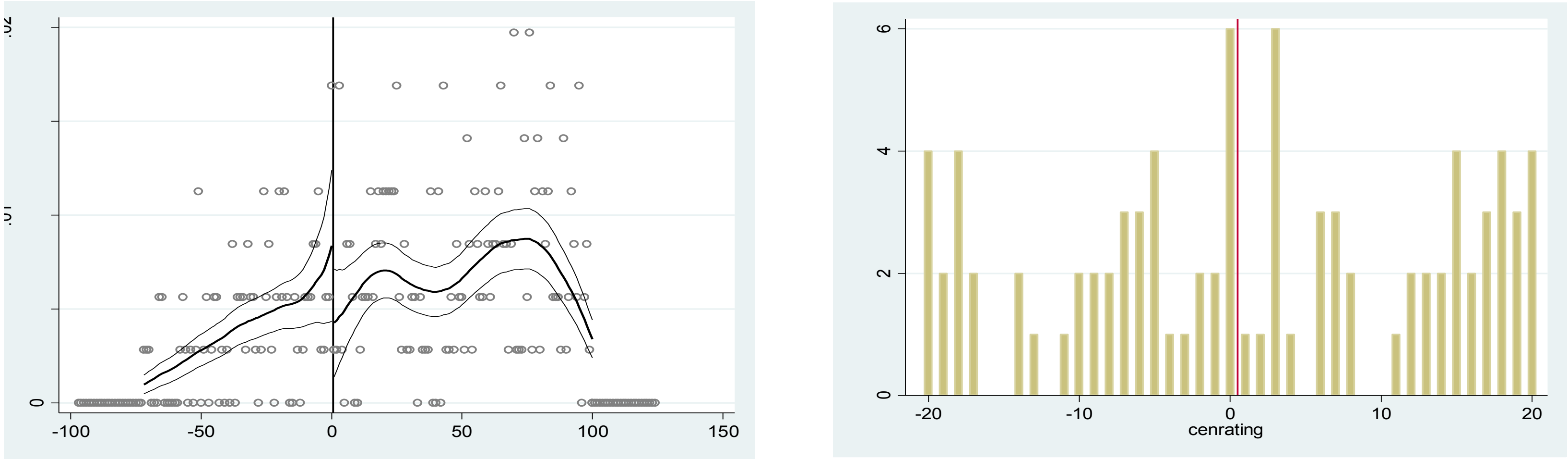


Table A3.1. Pairwise correlation matrix for DCPS TLF (All rater averages)

	TLF1	TLF2	TLF3	TLF4	TLF5	TLF6	TLF7	TLF8	TLF9
TLF1	1								
TLF2	0.775	1							
TLF3	0.738	0.768	1						
TLF4	0.747	0.755	0.791	1					
TLF5	0.743	0.771	0.761	0.77	1				
TLF6	0.635	0.692	0.673	0.66	0.699	1			
TLF7	0.715	0.736	0.744	0.72	0.734	0.654	1		
TLF8	0.721	0.724	0.737	0.74	0.733	0.624	0.670	1	
TLF9	0.672	0.678	0.671	0.68	0.696	0.602	0.647	0.779	1

Notes: N=5,363. Teacher-by-year TLF scores assigned by master educators and administrators for 2010-11 and 2011-12. All correlations are significant at p<0.001.

Table A3.2. Exploratory Factor Analysis Factor Loadings, TLF (All rater averages)

Variable	Factor1	Uniqueness
TLF1	0.849	0.280
TLF2	0.874	0.236
TLF3	0.872	0.240
TLF4	0.868	0.247
TLF5	0.873	0.237
TLF6	0.768	0.411
TLF7	0.827	0.316
TLF8	0.848	0.280
TLF9	0.799	0.362

Notes: N=5,121. Teacher-by-year TLF scores for 2010-11 and 2011-12. Varimax rotated factor loadings.

Table A3.3. Mean Item Scores for Five Latent Profile Solution TLF 2010-11 (data underlying Figure 3.1)

	Latent Class 1		Latent Class 2		Latent Class 3		Latent Class 4		Latent Class 5	
	n=363		n=147		n=596		n=723		n=857	
	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE
TLF1	2.447	0.066	1.993	0.071	3.56	0.025	2.868	0.052	3.21	0.029
TLF2	2.36	0.063	1.823	0.077	3.666	0.026	2.777	0.054	3.169	0.044
TLF3	1.985	0.064	1.583	0.046	3.449	0.029	2.433	0.055	2.911	0.054
TLF4	2.32	0.056	1.796	0.084	3.785	0.02	2.806	0.07	3.348	0.051
TLF5	2.405	0.065	1.859	0.074	3.689	0.025	2.84	0.051	3.239	0.044
TLF6	2.229	0.098	1.709	0.075	3.581	0.033	2.735	0.062	3.104	0.044
TLF7	1.793	0.051	1.413	0.061	3.247	0.036	2.209	0.064	2.65	0.048
TLF8	2.325	0.092	1.75	0.052	3.715	0.021	2.834	0.065	3.342	0.049
TLF9	2.609	0.085	1.973	0.061	3.682	0.022	3.066	0.057	3.418	0.03

Table A3.4. Mean Item Scores for Five Latent Profile Solution—CLASS and FFT, 2010-11 (data underlying Figure 3.2)

	Latent Class 1		Latent Class 2		Latent Class 3		Latent Class 4		Latent Class 5	
	n=122		n=264		n=266		n=360		n=582	
	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE
F_CERR	1.807	0.087	2.573	0.068	2.241	0.034	2.983	0.024	2.826	0.015
F_USDT	1.652	0.033	1.963	0.048	1.981	0.023	2.547	0.039	2.237	0.036
F_ECL	1.752	0.057	2.289	0.057	2.152	0.030	2.829	0.035	2.588	0.027
F_MCP	1.776	0.101	2.602	0.060	2.265	0.035	2.882	0.020	2.784	0.015
F_CS	2.000	0.039	2.452	0.058	2.338	0.024	2.893	0.027	2.685	0.022
F_MSB	1.871	0.114	2.743	0.053	2.282	0.039	2.993	0.023	2.896	0.012
F_ESL	1.725	0.048	2.186	0.049	2.154	0.021	2.744	0.034	2.504	0.033
F_UAI	1.686	0.055	2.043	0.049	2.032	0.025	2.583	0.031	2.352	0.030
C_PC	3.314	0.060	3.620	0.097	4.135	0.043	5.197	0.071	4.502	0.096
C_NC	4.020	0.104	5.016	0.159	4.400	0.097	5.855	0.083	5.478	0.065
C_TS	3.266	0.060	3.446	0.097	4.023	0.041	4.735	0.060	4.175	0.074
C_RSP	2.280	0.074	2.320	0.079	3.052	0.048	3.838	0.085	3.138	0.093
C_BM	4.191	0.179	5.794	0.114	5.243	0.056	6.249	0.031	6.093	0.026
C_PD	4.417	0.107	5.573	0.111	5.374	0.042	6.128	0.034	5.924	0.028
C_ILF	3.032	0.048	3.434	0.111	3.976	0.051	4.795	0.054	4.281	0.079
C_CU	2.758	0.049	3.119	0.121	3.625	0.046	4.436	0.069	3.895	0.063
C_APS	1.835	0.038	1.992	0.064	2.462	0.044	3.294	0.081	2.651	0.075
C_QF	2.522	0.047	2.707	0.100	3.358	0.048	4.325	0.091	3.576	0.088
C_ID	2.294	0.048	2.401	0.077	3.093	0.046	4.051	0.091	3.302	0.096
C_SE	3.468	0.077	4.202	0.098	4.520	0.046	5.439	0.046	4.919	0.077

Table A3.5. Mean Item Scores for Five Latent Profile Solution TLF 2010-11 (data underlying Figure A3.1)

	Latent Class 1		Latent Class 2		Latent Class 3		Latent Class 4		Latent Class 5	
	n=151		n=794		n=441		n=778		n=529	
	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE
TLF1	2.016	0.043	2.530	0.043	2.925	0.052	3.273	0.043	3.652	0.028
TLF2	1.958	0.044	2.528	0.055	2.928	0.049	3.310	0.051	3.718	0.023
TLF3	1.719	0.036	2.24	0.047	2.674	0.058	3.096	0.054	3.532	0.031
TLF4	1.950	0.048	2.562	0.052	3.015	0.060	3.425	0.048	3.789	0.023
TLF5	1.929	0.046	2.545	0.056	2.925	0.042	3.284	0.051	3.697	0.026
TLF6	1.770	0.051	2.453	0.059	2.855	0.046	3.211	0.051	3.619	0.031
TLF7	1.547	0.038	2.082	0.052	2.504	0.056	2.884	0.046	3.356	0.037
TLF8	1.907	0.046	2.511	0.052	2.997	0.064	3.397	0.052	3.767	0.018
TLF9	2.113	0.052	2.726	0.048	3.158	0.054	3.470	0.039	3.771	0.019

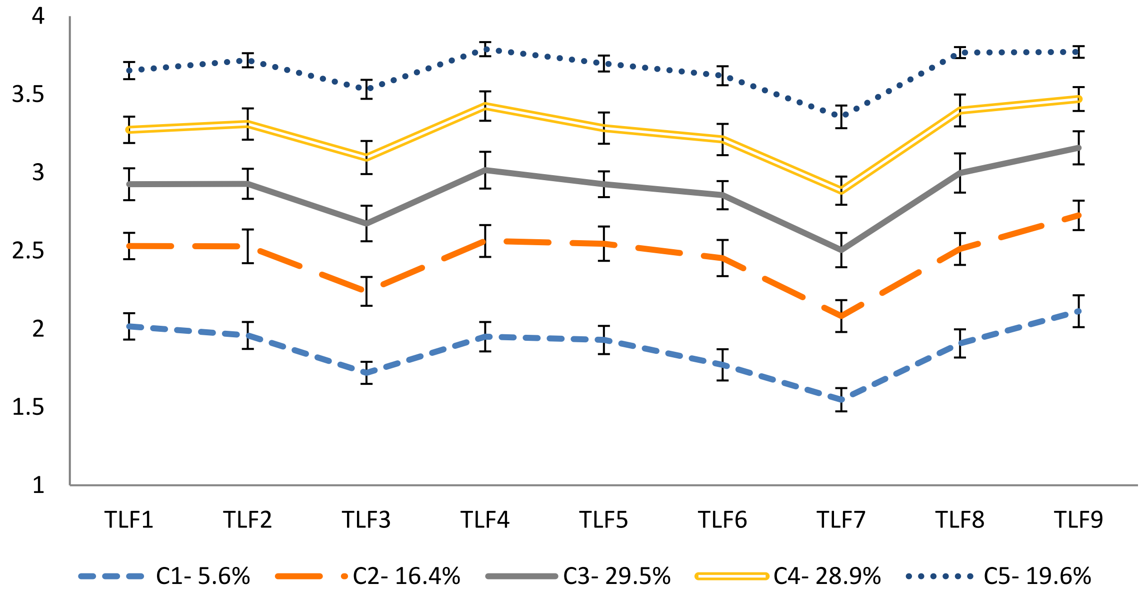
Table A3.6. Mean Item Scores for Five Latent Profile Solution CLASS 2010 (data underlying Figure A3.2)

	Latent Class 1 n=79		Latent Class 2 n=173		Latent Class 3 n=490		Latent Class 4 n=611		Latent Class 5 n=201	
	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE
C_PC	3.200	0.257	3.425	0.085	4.031	0.046	4.729	0.055	5.349	0.057
C_NC	3.886	0.098	4.889	0.295	4.839	0.076	5.558	0.069	5.892	0.086
C_TS	3.213	0.152	3.233	0.085	3.872	0.033	4.353	0.041	4.883	0.060
C_RSP	2.257	0.135	2.155	0.080	2.805	0.042	3.353	0.046	4.044	0.079
C_BM	3.778	0.510	5.65	0.264	5.643	0.062	6.092	0.027	6.267	0.034
C_PD	4.087	0.346	5.427	0.219	5.587	0.047	5.959	0.026	6.167	0.036
C_ILF	2.926	0.239	3.203	0.061	3.868	0.043	4.453	0.037	4.935	0.055
C_CU	2.731	0.264	2.850	0.090	3.527	0.038	4.059	0.033	4.602	0.068
C_APS	1.806	0.127	1.822	0.054	2.328	0.033	2.833	0.039	3.505	0.083
C_QF	2.496	0.193	2.457	0.105	3.181	0.036	3.808	0.051	4.531	0.060
C_ID	2.282	0.166	2.186	0.080	2.890	0.046	3.530	0.047	4.283	0.072
C_SE	3.254	0.351	4.029	0.077	4.474	0.048	5.097	0.038	5.548	0.045

Table A3.7. Mean Item Scores for Five Latent Profile Solution FFT 2010 (data underlying Figure A3.3)

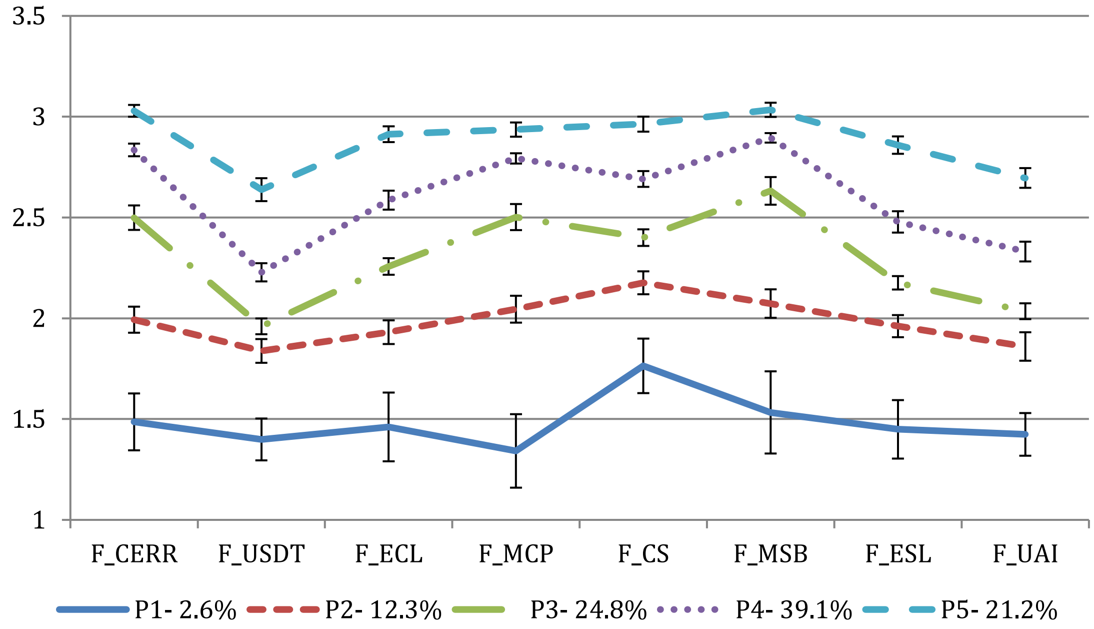
	Latent Class 1		Latent Class 2		Latent Class 3		Latent Class 4		Latent Class 5	
	n=41		n=191		n=386		n=607		n=329	
	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE
F_CERR	1.486	0.072	1.993	0.033	2.499	0.031	2.835	0.016	3.029	0.015
F_USDT	1.399	0.053	1.838	0.030	1.960	0.020	2.228	0.023	2.638	0.029
F_ECL	1.461	0.087	1.931	0.03	2.257	0.021	2.586	0.024	2.913	0.020
F_MCP	1.342	0.093	2.045	0.034	2.502	0.033	2.793	0.013	2.936	0.018
F_CS	1.764	0.069	2.176	0.029	2.400	0.021	2.691	0.020	2.963	0.019
F_MSB	1.533	0.104	2.073	0.036	2.632	0.035	2.895	0.012	3.034	0.018
F_ESL	1.449	0.074	1.961	0.028	2.176	0.017	2.478	0.027	2.859	0.022
F_UAI	1.424	0.054	1.860	0.036	2.035	0.020	2.331	0.025	2.696	0.025

Figure A3.1. Mean Item Scores for TLF (all rater averages), Five Class Solution (2010-11)



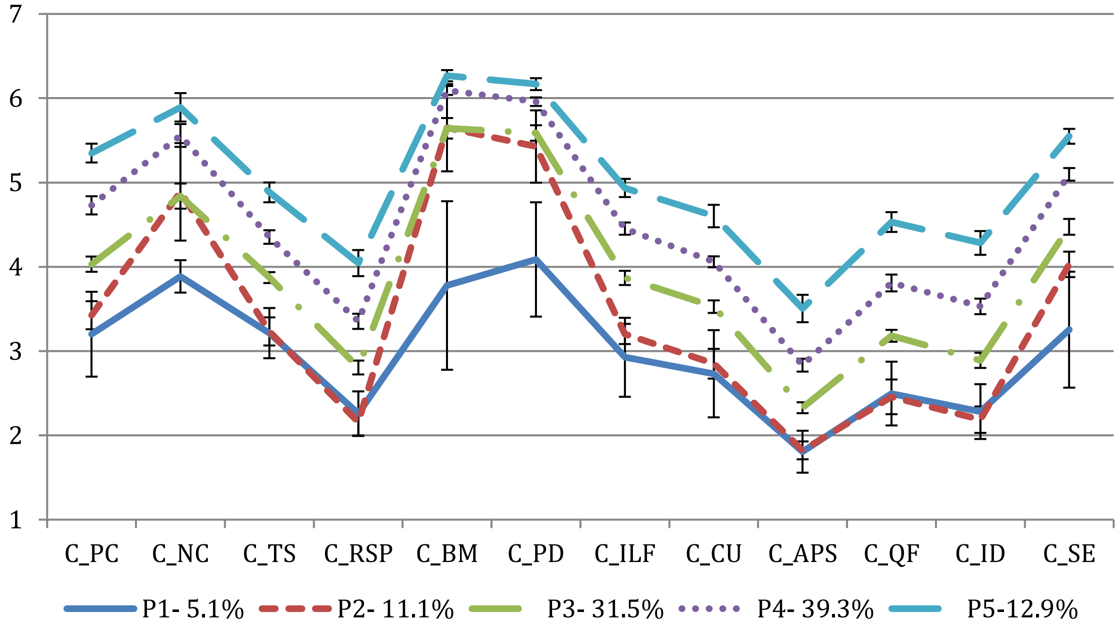
Notes: Error bars show 95% confidence intervals on mean estimates for each profile. Comparison is between height of each item; slopes are irrelevant.

Figure A3.2. LPA Mean Item Scores, FFT Five Profile Solution (2009-10)



Notes: N=1,554. Error bars show 95% confidence intervals on mean estimates for each profile. Comparison is between height of each item; slopes are irrelevant.

Figure A3.3. LPA Mean Item Scores, CLASS Five Profile Solution (2009-10)



Notes: N=1,554. Error bars show 95% confidence intervals on mean estimates for each profile. Comparison is between height of each item; slopes are irrelevant.