DEVELOPING TEACHER QUALITY:
EVIDENCE FROM THE DISTRICT OF COLUMBIA PUBLIC SCHOOLS

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

Jessalynn K. James

May 2019

Jessalynn K. James
Education Policy
Curry School of Education
University of Virginia
Charlottesville, Virginia

APPROVAL OF THE DISSERTATION

This dissertation, *Developing Teacher Quality: Evidence from the District of Columbia Public Schools*, 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)

Julia J. Cohen

Peter Youngs

Eric Taylor (Harvard Graduate School of Education)

J. Patrick Meyer (NWEA)

March 12, 2019
Date

DEDICATION

This dissertation is dedicated first to my parents—Christopher, Miriam, and Peter—who each have spent their careers in the service of the public and have motivated me to attempt to understand and alleviate systemic inequities in access to opportunity and justice.

I also dedicate this dissertation to the teachers who inspired me throughout my childhood and young adulthood—individuals who exemplified through their own intelligence, kindness, and passion for knowledge the importance of teachers for students' success both in school and beyond, which so many researchers have since confirmed.

Finally, I also dedicate this dissertation to my partner, Kevin, whose love and humor has helped make every step of this process a pleasure and whose unbounded belief in my potential has made him a loving, encouraging, and invaluable source of support.

ACKNOWLEDGEMENTS

I am grateful for the immensely helpful feedback and suggestions I have received from my dissertation committee members, Jim Wyckoff, Julie Cohen, Patrick Meyer, Eric Taylor, and Peter Youngs, who have been more generous with their time and advice than I could possibly have asked for. In particular, Jim Wyckoff, as my advisor and committee chair, must also be individually acknowledged for the immeasurably valuable advice, feedback, and support he has provided over the course of my training at UVA. My classmates and the faculty in the University of Virginia’s Education Policy program and affiliated programs have also been extraordinarily thoughtful and supportive, Katharine and Aliza in particular, and I am grateful to have had the opportunity to learn with and from such a thoughtful, passionate group of people.

I appreciate feedback on earlier versions of this research from discussants and participants at the Association of Education Finance and Policy, Association for Public Policy and Management, and American Education Research Association annual meetings, and from colleagues at the University of Virginia. I am additionally grateful to the District of Columbia Public Schools for supplying the data employed in this research and to Scott Thompson, Luke Hostetter, Lauren Norton, Chris Lewis, and Liz Kim for providing important context and background, and for answering many questions about the data and DCPS in general.

Financial support for this research comes from the Schusterman Family Foundation, the Overdeck Family Foundation, the Spencer Foundation, and the National Center for the Analysis of Longitudinal Data in Education Research (CALDER). This research was also supported by the Institute of Education Sciences, U.S. Department of Education, through Grants #R305H140002 and #R305B140026 to the Rectors and Visitors of the University of Virginia. The views expressed in the paper are solely my own and may not reflect those of the funders. Any errors or omissions are my own.

TABLE OF CONTENTS

ELEMENTS

LIST OF TABLES

LIST OF FIGURES

DISSERTATION OVERVIEW

A large and growing body of literature has established over the past couple of decades that the quality of a student's teacher can have considerable effects on that student's outcomes, both short- and long-term—and the extent to which individual teachers drive academic achievement can vary substantially (Aaronson, Barrow, & Sander, 2007; Chetty, Friedman, & Rockoff, 2014a, 2014b; Rivkin, Hanushek, & Kain, 2005; Rockoff, 2004). This expanding literature on teacher effectiveness has in part been the impetus for significant developments in the evaluation of teachers nationally, with the goal of identifying and developing quality teaching.

Until recently, teacher evaluation served primarily as a tool for teachers' development through cycles of observation and feedback conducted by their supervisors. By and large, however, evaluations remained informal and infrequent (Donaldson & Papay, 2015), until the 2000's, when reformers began to address the inadequacy of most existing evaluation programs, noting that they failed to differentiate levels of quality across teachers and rarely provided educators with actionable feedback for improving their instruction (Weisberg, Sexton, Mulhern, & Keeling, 2009). Meanwhile, the expansion of standardized testing that accompanied school-level accountability reforms at the turn of the century provided researchers with the data necessary to link student achievement data to individual teachers. These newly-available data paved the way for much of the most-influential research on teacher quality to date, providing evidence not just of the importance of teachers for student outcomes, but also the magnitude of variation in individual teachers' contributions to student learning and the extent to which teachers improve with experience. Though these accountability policies were targeted primarily at schools, its effects rippled down to teachers, as well. New data systems allowed for objective measures of individual teachers' performance, in the form of value-added scores, which could be used to differentiate teachers across the quality distribution (Donaldson & Papay, 2015). In addition, in response to the perceived coerciveness of federal accountability mandates, and encouraged by the accumulating body of evidence around variation in teacher quality and effectiveness, the federal government implemented a series of incentive programs that encouraged states to develop and adopt rigorous new evaluation programs (McGuinn, 2012).

The extent to which teacher evaluation transformed during the past two decades is significant, as are the implications of these changes for teachers' development and in-service training. One of the most prominent, high-stakes, and arguably more successful examples of the recently expanded use of evaluation is the program employed by the Washington, DC public school (DCPS) system—IMPACT, which serves as the basis for each of the three chapters in this dissertation. IMPACT consists of several measures of teacher quality, including as many as five separate observations of teaching practice per year conducted by both a principal and a trained external observer, and—for teachers in tested grades and subjects—value-added scores that reflect teachers' contributions to their students' achievement gains. In addition, DCPS uses teacher-assessed student achievement data—student performance from assessments that have been selected by the individual teacher, benchmarked against pre-specified targets; a rubric-based assessment of the teacher's commitment to the school community; and a measure of teachers' professionalism. This program ties teachers' IMPACT scores to promotion and retention decisions. Teachers can lose their job if they score low enough on the IMPACT scale or fail to make adequate improvement within a specified period. Conversely, teachers who exceed a certain performance threshold are also eligible for promotion, coupled with an increase in base pay, as well as a bonus payment that can well exceed $10,000, with larger financial rewards available to teachers in the hardest-to-staff schools.

IMPACT was intentionally designed to improve teaching and learning in DCPS and to do so through two core mechanisms: by incentivizing teachers to improve their practice, and by altering the composition of the district's teaching force. Compositionally, IMPACT contains features that could lead to differential attrition across teachers' performance levels. The threat of dismissal associated with poor performance could dis-incentivize low-performing teachers' retention, while explicitly removing the lowest performers. On the other end of the distribution, DCPS encourages high-performers' retention with generous financial incentives and professional advancement opportunities. Meanwhile, these same features might also change the composition of *entering* teachers if, for example, the risk associated with low performance discouraged lower-potential teachers from working in the district, and if the high financial rewards and professional advancement opportunities that accrue to the top teachers in the district make teaching in DCPS more attractive to high-performing teachers who would otherwise have worked elsewhere.

Beyond compositional effects, IMPACT aims to improve the performance of extant teachers through the feedback, professional supports, and incentives that are embedded in the program. IMPACT attempts to accelerate the development of effective teaching by incentivizing improvement among low-performing or inadequately-improving teachers who would otherwise lose their jobs, as well as by encouraging further performance gains and continued high performance among its best teachers with financial and professional rewards. Its system of frequent evaluation and feedback for teachers at all performance levels can facilitate this development by providing teachers with regular and explicit guidance for improvement.

Together, these features of IMPACT should not simply shift the overall distribution of teaching quality in the DCPS upward, but also narrow variation in performance across the district, thereby lessening differences in access to quality teachers across students. Each of these mechanisms are explored through the three chapters of this dissertation, as I explore how teacher evaluation can drive teachers' performance and retention decisions in high-stakes settings, and how observation-based measures of teachers' performance may glean insights into teachers' development over time.

The first chapter focusses on the effect of the transition to the Partnership for Assessment of Readiness for College and Careers (PARCC) exam, which was aligned to the new Common Core State Standards (CCSS) for student learning, on teaching quality in DCPS. The CCSS and the PARCC exam were expected to significantly shift teaching and learning in the U.S. Out of concern that teachers could not be fairly evaluated using student achievement measures until teachers had time to familiarize themselves with the new tests and adapt their instruction accordingly, most districts implementing new, CCSS-aligned assessments—including DCPS—chose to temporarily halt the use of value-added scores for teacher accountability. Meanwhile, districts struggled to prepare their teachers to adapt to the teaching required by the new CCSS-aligned exams. This chapter first explores the extent to which the new test would have mattered for teachers' value-added scores in DCPS, and then attempts to identify whether teaching practices are differentially important for student achievement on PARCC relative to the preceding exam. I find that teachers' value-added scores do not change meaningfully with the new test, and there is inconsistent evidence as to whether teachers' practices are differently important for student achievement across the exams—though math achievement on PARCC may be more responsive to the teaching practices assessed on DCPS's teacher observation protocol.

Meanwhile, there was less public concern about the effect of the new test on teachers' practice, even though proponents of the CCSS expected the new assessments to require a substantial shift in instruction (Conley, 2014; Student Achievement Partners, 2013, 2014). For this reason, teachers in tested grades and subjects might have responded in a high-stakes context like DCPS by strategically shifting their instructional focus to the practices that are theoretically better aligned to student learning on the new assessment; on the other hand, their practice could have suffered as they attempted to implement new expectations for teaching and learning in their classroom. Indeed, I find the latter to be the case in DCPS; scores on the district's observation rubric for teachers in tested grades and subjects were negatively affected by the change in assessments, suggesting that in high-stakes settings like DCPS, measures of teachers' practice may be sensitive to test changes and policymakers should consider the ramifications for more teacher quality measures than teachers' value-added scores alone. Such transitions are not uncommon, and this chapter provides insights into what other districts might expect in future transitions, particularly in an era when measures such as value-added scores and classroom observations are commonly used to evaluate teachers' performance.

The second chapter also considers the consequences of IMPACT's high stakes, though in the context of effects on teachers' retention and performance intentional to the program's design. This chapter, coauthored by Tom Dee and Jim Wyckoff, builds upon early evidence from DCPS to determine whether the threat of dismissal associated with low performance has retained its salience in the face of modifications to the evaluation program's design, alongside political, structural, and contextual changes within the district. We contrast our findings with early evidence from DCPS (Dee & Wyckoff, 2015), where teachers who performed at a level that defined them as Minimally Effective (ME), just missing the cut-off for a non-consequential Effective rating, demonstrated a higher rate of attrition and—for those who remained in DCPS—higher scores the next year, than their peers just above the threshold who were not faced with performance sanctions. Our new research demonstrates effects at least as large as in the early years, in spite of what is arguably a weaker treatment contrast; teachers who score just above the ME level are now considered Developing, and are still subject to dismissal if they do not attain a higher rating within three years.

This chapter is particularly important in the context of a recent evolution to the policy discussion around teacher evaluation. While the teacher-evaluation reforms of the past decade, IMPACT included, were touted as pivotal for improving teacher effectiveness and, in turn, student achievement, recent high-profile evaluations have yielded mixed evidence of effectiveness, and the public image of teacher evaluation has been accordingly tarnished. These discussions, however, often fail to address the quality of implementation—and competing forces that might diminish its effects—relative to a given setting's success. This evidence from DCPS is a rare second look at a program that had initially strong effects on teacher retention and performance, but that might have experienced drastically different effects as IMPACT, and the district, evolved. Rather, we find that in the face of each of these changes, IMPACT continues to increase both attrition and the performance of returning teachers when teachers are faced with imminent dismissal threats. This chapter suggests that teacher evaluation can be sustained in a manner that continues to improve the quality of teaching.

The performance effects that we find in the second chapter for incentivized teachers lend suggestive evidence to the malleability of teaching quality in response to performance incentives. The final chapter, however, takes a broader view of teacher development to better understand the ways in which teachers improve as they gain experience in the classroom. This chapter, written with Eric Taylor and Jim Wyckoff, explores patterns in teachers' early-career skill development. While it is now well documented that teachers get significantly better at influencing student achievement as they gain experience, with the steepest returns to experience in the first few years of teachers' careers, less is known about the nature of these improvements. Chapter 3 investigates the extent to which teachers also make large early-career gains in terms of the practices and skills they exhibit in the classroom. Using scores from the Teaching and Learning Framework (TLF), DCPS's observation protocol, we identify substantial improvements to teachers' overall practice in their initial years in the classroom, with wide variation in gains across practices and across teachers. These gains suggest that teachers do a considerable amount of learning in the classroom; front-loading targeted professional development—or giving teachers richer access to opportunities to practice these skills in their preparation programs—might lead to substantially more-effective cohorts of new teachers entering the district each year. Importantly, we also establish an association between teachers' improved practices and their students' learning gains, suggesting that when teachers improve their skills, their students benefit, as well.

Together, these chapters indicate that teacher evaluation in DCPS can provide useful insight into teacher's' development—and that the process of evaluation in DCPS may drive this development in and of itself. The research presented in this dissertation suggests that teachers' performance is sensitive to the contexts in which they teach, as chapter 1 demonstrates that changing standards and expectations can significantly influence teachers' practice—and can do so in potentially deleterious ways. It also demonstrates (in chapter 2) that, even in mature evaluation systems, low-performing teachers make decisions about whether to remain in the profession when faced with consequences of their inadequate practice, and these teachers improve when they choose to stay. Finally, the third chapter demonstrates that improvements do not simply occur at consequential performance thresholds, but that new teachers in general, who are in the process of learning and developing the skills that will make them better teachers, make substantial performance gains though their early years in the classroom. The variation in these gains across teachers and across practices, likewise, highlights areas where targeted professional development might facilitate steeper performance trajectories for DCPS teachers overall.

The broader relevance of evidence from IMPACT invariably and justifiably is often questioned, given that DCPS's evaluation system is uniquely high stakes, and the district serves a relatively non-representative population, with a disproportionately high share of its student body coming from underprivileged backgrounds. Yet, while the policy conversation has begun to drift away from teacher accountability for teachers, elements of more-rigorous evaluation reforms largely remain in place nationally (Doherty & Jacobs, 2015), and issues of recruitment, retention, and teacher development remain critical for school districts more widely. The unique features of IMPACT, however, are what make DCPS a useful context for learning about teacher development and the ways it might be fostered through rigorous teacher evaluations. The rich data collected by DCPS additionally provide us with the ability to ask—and answer—nuanced questions about teachers' development and the evaluation conditions that facilitate improved teaching and learning.

Such questions form the basis of this dissertation, and are particularly meaningful given the population the District serves. Understanding teacher development in settings like DCPS, which serves a low-income, high-minority student population, is vital for improving educational equity across student demographics. The literature on access to quality teaching shows that students with demographic characteristics similar to those of the student body of DCPS are more likely to have underqualified, less experienced, and possibly less-effective teachers (Boyd, Lankford, Loeb, Rockoff, & Wyckoff, 2008; Clotfelter, Ladd, & Vigdor, 2005, 2006; Goldhaber, Quince, & Theobald, 2016; Jackson, 2009; Kalogrides, Loeb, & Bèteille, 2012; Lankford, Loeb, & Wyckoff, 2002; Sass, Hannaway, Xu, Figlio, & Feng, 2012). Meanwhile, the strongest outcomes associated with teacher effectiveness and development are often concentrated among high-poverty or low-achieving schools, classrooms, and students (Adnot, Dee, Katz, & Wyckoff, 2017;

Jacob & Levitt, 2003; Jacob, 2005), and among new and low-performing teachers (Boyd, et al., 2008; Pope, 2019; Sun, Mutcheson, & Kim, 2015; Taylor & Tyler, 2012), suggesting that policies to facilitate teacher development might be more meaningful in these contexts and could serve to make a sizeable dent in the academic achievement gap.

CHAPTER 1

Teacher Quality in the Common Core Era:
Does the Test Make a Difference?

Abstract – The recent transition to the Common Core State Standards (CCSS), with a corresponding shift to new assessments such as the PARCC exam, was expected to have significant implications for teaching and learning in the U.S. Out of concern that teachers could not be fairly evaluated using student achievement measures until teachers had time to familiarize themselves with the new tests and adapt their instruction, most districts implementing new, CCSS-aligned assessments chose to temporarily halt the use of value-added scores for teacher accountability. Meanwhile, districts struggled to prepare their teachers to adapt to the teaching required by the new CCSS-aligned exams. This paper: 1) explores the extent to which the new test would have mattered for teachers' value-added scores in the District of Columbia Public Schools (DCPS); 2) attempts to identify the differential importance of core teaching practices for PARCC achievement relative to the preceding exam in DCPS; and 3) evaluates the effect of the transition to the new assessment on DCPS teachers' practice. I find that teachers' value-added does not change meaningfully with the new test, nor is there consistent evidence as to whether teachers' practices are differently important for student achievement on the new exam relative to the prior assessment. Teachers' instructional practice, however, was negatively affected by the change in assessments, suggesting that policymakers should consider the ramifications of testing changes on more teacher quality measures than teachers' value-added scores alone.

INTRODUCTION

In June 2010, the National Governors Association (NGA) and the Council of Chief State School Officers (CCSSO) unveiled the Common Core State Standards (CCSS). By the end of 2011, the standards were officially adopted by all but five states ("Common Core or Something Else?", 2015) and new assessments aligned to the standards were rolled out by the 2014-15 academic year (AY).[1] A key goal motivating both the new standards and the accompanying new tests was to raise expectations and increase the rigor of material learned by U.S. students (Conley, 2014), representing a marked shift for most states and districts. The type of teaching required for students to gain proficiency on these standards was expected to differ from the type of teaching required to meet other, less-rigorous standards (Conley, 2014), and school districts scrambled to better equip teachers with the skills and practices necessary for student success on CCSS-aligned exams (Jochim & McGuinn, 2016). Meanwhile, concerns about the fairness of using student achievement on new tests for teacher evaluation led the U.S. Department of Education to allow states and districts to delay or put on hold the use of these exams for teacher accountability while teachers and districts adapted to the new assessments (Duncan, 2014). Many states and districts exercised this option and omitted value-added scores from their teacher evaluations during the transition to CCSS-aligned exams, even while continuing to evaluate teachers using measures of their practice.

In this paper, I use evidence from the District of Columbia Public Schools (DCPS) to explore how measures of effective teaching may differ across the transition to

[1] Membership in the Common Core has evolved since this time, as several states withdrew or revised the standards. As of September 2017, eleven states had announced that they would replace or revise the standards.

a CCSS-aligned exam and what might be learned about teachers' differential teaching skills. DCPS started using the Partnership for Assessment of Readiness for College and Careers (PARCC) exam in AY 2014-15 (see figure 1.1 for a timeline) in lieu of its predecessor in DCPS, the Comprehensive Assessment System (CAS). DCPS is uniquely suited to understanding this transition. The district has used a rigorous, multiple-measure evaluation system, IMPACT, to assess the performance of all of its educators since AY2009-10. As part of IMPACT, DCPS maintained a consistent teacher observation rubric through the transition to new standardized student assessments. The observation rubric used in DCPS is the "Teach" section of the Teaching and Learning Framework (TLF), which consists of nine key teaching practices, rated on a scale of 1 through 4. These practices are defined in appendix table A.1.1. Since the advent of IMPACT, DCPS has also linked students in tested grades and subjects to their respective teachers, allowing for the estimation of teachers' value added to student achievement, even in years when value-added scores were not formally included in teachers' evaluations.

Using data from DCPS, I investigate whether more-rigorous student achievement exams prioritize different teaching skills and whether the transition to these new exams causes teachers to alter their practice. Specifically, I ask: first, whether PARCC and CAS rank teachers differently according to their value added; second, whether students have stronger learning gains on the PARCC exam than on the CAS exam when they have teachers who are relatively stronger on certain skills; and third, whether teachers change their practice in response to the transition to the PARCC exam.

These questions address issues of significant interest across multiple disciplines. From a policy perspective, each of the research questions examined here can provide important insight to teacher evaluation systems and the contexts in which they operate. For example, if teachers are differently ranked across the two exams in terms of value-added scores, it may lend support to states' and districts' decisions to take a hiatus from using student achievement to evaluate teachers during major transitions such as the shift to the CCSS-aligned exams. It could also raise questions as to whether either or both of the exams are sufficiently well suited for use as a value-added outcome measure. Meanwhile—relevant to both policy and instructional researchers—if student achievement is more strongly associated with different teaching practices across the exams, it could provide guidance to districts on which practices and skills they should focus professional development (PD) under the new standards and testing regime. Similarly, the effects of the new assessment on teachers' practice can highlight areas where teachers' instructional skills may be sensitive to assessment changes.

This research also explores multiple approaches to linking test scores across the distinct examinations to facilitate comparisons. The measurement field typically equates scores using raw test scores from populations that are randomly equivalent. This approach, however, is not feasible for examinations that operated in different years, and so this paper employs and evaluates the quality of three non-conventional linking methods in order to compare student achievement across CAS and PARCC. This paper is the first I am aware of to assess the quality of these linking approaches using scores from distinct exams.

BACKGROUND

A Push for More Rigor in American Education

The CCSS were developed by a group of governors and state education officials, with input from education researchers, teachers, and content experts, whose shared intent was to create a common set of coherent, rigorous, and evidence-based standards for what students should know and be able to do at the end of each grade (Conley, 2014; CCSS Initiative, 2010a, 2010b). Developed in response to evidence that the U.S. was trailing other economically-developed nations, the CCSS were intended to improve the depth of U.S. students' understanding such that they might be globally competitive (Conley, 2014; NGA, 2008).

Prior to the CCSS, the scope, rigor, and clarity of learning standards varied considerably across states. In some states, standards were opaque, while in others they were excessively granular. In general, states had a plethora of standards that touched an excess of topics but did not address deeper expectations for knowledge, leading to the claim in many locations that their standards, particularly in mathematics, were "a mile wide and an inch deep" (Chang, 2013; Schmidt, McKnight, & Raizen, 2007). In addition, expectations for learning were inconsistent from one state to the other; a skill that was expected for a first-grader to learn in one state might not be broached until the third grade in another. These inconsistent expectations were also evident in large and widening achievement gaps across states and districts, which were illuminated by a database recently developed to facilitate comparisons across states using different assessments (Reardon, 2013; Reardon, Kalogrides, & Ho, 2019).

Among the goals driving development of the CCSS were that the new standards would be “fewer, clearer, and higher”. These standards were designed, first, in response to the evidence of a core set of skills (“fewer”) required for success in two-year college, regardless of program path. The CCSS developers also aimed to present these standards coherently and without redundancy (“clearer”) such that each standard could be clearly linked to learning materials (e.g., curricula and assessments). They also focused on deeper, conceptual learning (“higher”) from which students could more easily transfer knowledge and skills across contexts and disciplines. Both the math and English Language Arts (ELA) learning standards implied an increase in expectations, encouraging students to learn content on a deeper level than what most states specified before the CCSS (Student Achievement Partners, 2013 & 2014). Studies of the standards and assessments across the transition demonstrate that in most cases these goals have been attained (Conley, 2014; Peterson, Barrows, & Gift, 2016; Doorey & Polikoff, 2016; Yuan & Le, 2012). Actual effects on the depth of student learning, however, have yet to be ascertained; this is a difficult question to assess, as there are few if any appropriate comparison groups from which to contrast achievement in CCSS-adopting states and districts (Polikoff, 2017).

New Standards, New Tests

In conjunction with the standards, the creators of the CCSS also aimed to develop assessments that could provide formative information about students’ knowledge and abilities (Bill and Melinda Gates Foundation, 2010; McDonnell & Weatherford, 2013). Two national consortia of states, PARCC and the Smarter Balanced Assessment

Consortium (SBAC), were convened to address this goal, each developing its own CCSS-aligned assessment to be used across participating states.

External evaluations of assessment rigor and alignment to the CCSS found the PARCC and SBAC assessments to be well aligned to math and ELA content standards across grades, and good matches to the depth of learning prescribed by the CCSS (Doorey & Polikoff, 2016; Schultz, Michaels, Dvorak, & Wiley, 2016). Although states establish their own proficiency levels even across common assessments, an analysis that compared proficiency standards—before and after the transition to CCSS-aligned assessments—to a rigorous, nationally-recognized benchmark found that most states, including Washington, D.C., significantly raised their expectations for students' proficiency (Peterson, Barrows, & Gift, 2016).

Perhaps recognizing that states and districts might only superficially adopt the CCSS, key leaders in the development of the new standards explicitly acknowledged the importance of the assessment consortia for clarifying the standards' definitions and associated expectations (Bill and Melinda Gates Foundation, 2010), which would be done through consortia- and partner-developed resources, as well as the assessments themselves, which—once available—would provide insight into how to interpret and apply these standards. Importantly, PARCC and SBAC make use of innovative new technologies to include test items that are meant to more accurately measure the complex skills delineated by the CCSS than traditional item formats (i.e., multiple-choice and constructed-response).[2]

[2] The extent to which these items were as innovative as initially envisioned is up for debate. Regardless, the test format is inarguably different from the exams students were typically taking before PARCC and SBAC, raising the possibility that differences in achievement across the two exams may in part reflect familiarity (or lack thereof) with the new exams' item formats and the technology with which the new

The Common Core in DCPS

I focus on the contrast between the PARCC assessment, which is currently used by DCPS, and the District-specific assessment used in preceding years, the DC CAS. While the District formally adopted the CCSS in AY2011-12, there is reason to believe that teachers may not have adapted their teaching to the new standards before they began using the national CCSS-aligned assessments.[3] First, while DCPS had been recognized in at least one review as having relatively high standards in advance of their transition to the CCSS (Carmichael, Martino, Porter-Magee, & Wilson, 2010), the CCSS still represented a significant shift for DCPS. An informal review of District-developed crosswalks (Office of the State Superintendent of Education [OSSE], 2011a, 2011b) reveals key differences even in the criteria that DCPS considered comparable across the two sets of learning standards. For example, the grade 3 ELA CCSS point to use of more complex, diverse texts, and place more emphasis not just on locating and identifying relevant information, but explaining the relevance of that information. In Grade 3 math, the CCSS explicitly expect students to "develop understanding" in order to apply mathematical concepts and demonstrate fluency, rather than simply showing that they can perform certain operations and procedures.

Blueprints for the two tests reveal stark differences in test structure and format. CAS consisted primarily of selected-response items (94%), while the PARCC exam contains a higher share of complex item types, including performance tasks, which place

exams are delivered (i.e., computers versus paper and pencil). The issue of comparability between the PARCC and CAS exam formats, and implications of these differences, is discussed in more detail under Assumption 1 in section 4 of appendix B.

[3] As a robustness check, I test for changes in teachers' practice at the point of transition to the Common Core State Standards but before the transition to the PARCC exam to test whether this assertion holds empirically.

a higher emphasis on more cognitively complex skills, and items for which scoring procedures can make it more difficult to earn points (Darling-Hammond & Adamson, 2014; Doorey & Polikoff, 2016; Schultz, Michaels, Dvorak, & Wiley, 2016). In math, PARCC relies heavily on scaffolded, interdependent items that require a mix of written and non-written responses, with formats that include open- (e.g., typing in a calculated number) and selected-response, as well as technology-enhanced formats, such as drop-down, multiple-select, plotting, and graph-building. At least 80 percent of math PARCC items are categorized as Type II or Type III—multi-part questions that typically cannot be scored fully by machine. The ELA exam is similarly comprised of multi-part items that use a mix of selected-response, constructed-response, and technology-enhanced items.

These differences across assessments are key to the expectation that effective teaching might differ under PARCC, but it is less clear precisely which teaching practices might be more important for student learning on PARCC than on CAS. I theorize that the practices that are defined in terms of deeper learning and conceptual understanding may be more highly associated with student learning on the PARCC exam, given the new assessments' focus on higher-order skills and conceptual fluency over procedural knowledge. I base this theory in part on the body of evidence that shows moderate but consistent correlations between teachers' practice and their students' learning, as measured by standardized assessments.

The Test Matters

Few studies to date have looked at transitions in learning assessments and standards in order to assess differences in how teachers rank in terms of value added.

Backes, Cowan, Goldhaber, Koedel, Miller, and Xu (2018) compared teachers' performance across changes in assessments and standards in five states or cities, with a focus primarily on the year of transition, to test the stability of value-added estimates across "regime" changes, questioning the assumption that it would be unfair to impose accountability policies on teachers during such transitions. Two of these locations included transitions to the CCSS and associated assessments: Kentucky (which created its own CCSS-aligned assessment), and Massachusetts (which allowed its school districts to choose between their local CCSS-aligned assessment and PARCC). The remaining regime changes occurred prior to states' adoption of the CCSS. Backes et al. found a small decline in correlations between current- and prior-year value-added scores for some of the states, particularly in reading, though there was not a consistent pattern in terms of whether these transitions were related to the CCSS. Looking at the stability of teachers' place in the distribution, only one state (Kentucky) saw significant changes in the likelihood of bottom-decile reading teachers remaining in the same decile across the transition. Generally, across these locations, instability during the transition year was low and not statistically significant.

Districts' concerns about using teachers' value-added scores for accountability purposes when new tests are introduced are not, however, unfounded. There are a number of potential differences in any two tests that could potentially drive differences in teachers' value-added rankings, such as floor or ceiling effects. Koedel and Betts (2010), for example, found that some tests may have ceiling properties which could disproportionately affect teachers with certain types of students. In the DCPS context, the student population typically has low achievement levels, and so there may more likely be a floor effect—about which there is little evidence in the value-added literature—than a ceiling effect. PARCC 8th grade math, for example, has a low share of low-cognitive-demand (DOK 1) items (Doorey & Polikoff, 2016), in particular relative to the share of DOK 1 standards, which may make the assessment less effective at capturing the full range of students' abilities across the standards—and this may disproportionately affect students at the lower end of the performance distribution. If there were in fact a PARCC floor effect (or a CAS floor or ceiling effect), there would be clusters of students at the bounds of the observed score ranges. The distribution of scores across the assessments (not shown) indicates floor effects on both exams, and some ceiling effects on CAS. In their analysis of ceiling effects on value-added distributions, however, Koedel and Betts found no significant change in teachers' rankings in the presence of test ceilings except when the ceilings were severe.

Teachers might also rate differently across tests when there are differences in score weighting across learning standards. Even when assessments are aligned to the same construct, they may apportion different weights to each learning standard in a way that could affect teachers' value-added rankings. Lockwood, McCaffrey, Hamilton, Stecher, Le, and Martinez (2007) found that this was true of the Stanford 9 mathematics assessment in one large school district, where value-added estimates were somewhat sensitive to the value-added model and covariates used, but far more sensitive to the different subcomponents of the assessment. Lockwood et al. experimented with different weighting schemes for these subcomponents, and found that—on this test and sample—teacher effects differed substantially according to the relative weights assigned to each construct within the overall assessment.

Other factors that may contribute to instability in value added across assessments include the timing of the testing window, measurement error, the stakes associated with the assessment (Cohen, 2015; Corcoran, Jennings, & Beveridge, 2013; Papay, 2011), and possibly the item types (e.g., multiple choice, constructed response, etc.; Grossman, Cohen, Ronfeldt, & Brown, 2014). All of this leads to the hypothesis that teachers' value added will differ across the PARCC and CAS exams in DC, as the assessments vary in difficulty, content, and other key properties that have been shown to influence teachers' value-added rankings.

Other Factors that May Affect Teachers' Value-Added Scores

Even within a consistent standards and testing regime (and within a consistent value-added modeling approach), we would expect to see some variation in teachers' value-added rankings over time. Studies examining the stability of teachers' value-added scores have found year-to-year correlations ranging between 0.18 and 0.63 (Koedel, Mihaly, & Rockoff, 2015). A key driver of this instability is measurement error, often arising from the assessments or from the number of students who can be linked to a given teacher (Goldhaber & Hansen, 2013; McCaffrey, Sass, Lockwood, & Mihaly, 2009; Papay, 2011).

There are, however, several additional factors that may affect a teacher's place in the value-added distribution, conditional on her underlying effectiveness. These include shocks akin to the assessment change I analyze here, as well as changes in the grade or subject taught (Cook & Mansfield, 2017; Ost, 2014). Systematic improvements can also be expected among novice teachers, who on average make large gains as they acquire experience in the classroom (Boyd, Lankford, Loeb, Rockoff, & Wyckoff, 2008; Harris & Sass, 2011; Papay & Kraft, 2015; Rockoff, 2004; Wiswall, 2013).

Another such factor is a teacher's underlying ability level. Teachers at different places in the performance distribution exhibit different levels of stability in their value-added scores over time. Goldhaber and Hansen (2013) find that the stability of value-added scores is slightly higher for teachers on the upper end of the value-added distribution, with monotonic increases in stability at each decile of performance; this pattern holds even among a restricted sample of teachers with at least five years of experience. Teachers at different skill levels may also be differently able to adapt to new standards and assessments, though I am aware of no empirical evidence documenting such a relationship.

The consistency of teachers' year-to-year value-added also varies according to the type of students taught. Stacey, Guarino, and Wooldridge (2018) find that the stability of value-added estimates is higher for teachers with higher-performing students. Much of the instability experienced by these teachers can be explained by transitory shocks (e.g., changes in classroom makeup or a teacher's peers from year to year), though the student achievement tests used in their study may also have higher measurement error for lower-scoring students. It is unclear the extent to which this finding is specific to the exam and context where Stacy et al. (2018) conducted their study, nor their applicability to the exams used to calculate value-added in DCPS. In terms of measurement error contributions, technical specifications for the PARCC and CAS exams do not report reliabilities or standard errors of measurement by performance level, but there are no meaningful differences in measurement error across other reported student characteristics (race on either exam, and disability, poverty, or English language learner [ELL] status on the PARCC exam).

Finally, a particularly important potential factor in the context of DCPS is the stakes under which students and teachers are assessed. In DCPS, teachers' performance on the overall evaluation measure has high stakes for their ability to retain their jobs or to earn large financial rewards.[4] The high-stakes nature of DCPS's teacher evaluation policy may induce teachers whose current-year performance places them near a key score threshold to improve their next-year value-added scores. However, changes to the incentive structure went temporarily into effect during the transition to PARCC. In the years immediately preceding the implementation of the PARCC exam, teachers in tested grades and subjects were evaluated in part by their contributions to student learning on the CAS exam, with 35% of their IMPACT scores coming from these value-added scores. For the first two years of PARCC, however, value-added scores were omitted from the calculation of teachers' IMPACT scores. Instead, higher weight was placed on teachers' TLF scores.

This shift in incentives may drive changes in the relationship between teachers' performance across years, along with the shift in tests. Such an association has been suggested by research on value-added in other contexts. While possibly due to differences in content and design, Corcoran, Jennings, and Beveridge (2013) observed greater variation in teachers' value-added on high-stakes exams used in Houston, Texas for math and reading than in low-stakes exams given in the same years and for the same subjects. McCaffrey et al. (2009) likewise found different levels of stability between

[4] See Dee and Wyckoff, 2015, and Toch, 2018, for details.

concurrently-administered exams in Florida; however, there was no consistent pattern across districts in terms of whether the high- or low-stakes exams had higher adjacent-year stability. Additional evidence that teachers may improve their value-added scores in response to incentives comes from a working paper by Dinerstein and Opper (2017), who found that New York City teachers for whom school administrators considered their value-added performance when making tenure decisions had higher value-added scores than when their performance on the measure was not formally considered; this effect was small, however, and did not persist to future years of students' achievement.

It is not clear how the limited evidence on teachers' value-added rankings across tests with different stakes might apply to the DCPS context. First, the high-stakes exam explored by McCaffrey et al. (2009) was used for school-, rather than teacher-level, accountability. Second, the Dinerstein and Opper (2017) analysis showed only small incentive benefits to student outcomes and only in the treatment year. The policy in New York City also represented the *introduction* of a test-based incentive, which may be more salient than the temporary removal of such an incentive. Finally, other research from DCPS indicates that the TLF was already a major focus of teachers' efforts even when value-added scores were formally included in teachers' evaluations. In an analysis of the effects of performance incentives on teachers' practice, Adnot (2016) found, using pre-PARCC data, that DCPS teachers near key performance thresholds (i.e., those that placed teachers either at risk of dismissal or gave them potential for large financial rewards) made significant gains on the TLF the following year. Given that the TLF already accounted for a plurality of a teacher's overall evaluation score (with the exception of AY 2009-10, when it accounted for 50 percent of scores for teachers in tested grades and subjects), and the TLF also serves as a formative measure, providing both written descriptions of exemplary teaching and in-person feedback from evaluators, teachers may be better equipped to adapt their teaching in response to their classroom observation scores than relatively opaque value-added scores.

Student Achievement and Teaching Practice

While student learning gains, as measured by value added, are generally moderately correlated with measures of teachers' practice (Ho & Kane, 2013; Kane, Taylor, Tyler, & Wooten, 2011; Whitehurst, Chingos, & Lindquist, 2014), relatively little is known about which practices are more important for student learning or how that relative importance might differ across assessments. Recent research on the new CCSS-aligned assessments, however, suggests that these new tests may be differently sensitive to instructional differences between teachers (Kane, Owens, Marinell, Thal, & Staiger, 2016), though the specific practices that might explain these differences were not measured.

Kane et al. (2011) found that the teaching practices identified by Charlotte Danielson's (1996) Framework for Teaching—an observation rubric with which the TLF shares many similarities—are important overall for student achievement in both reading and math, though certain practices may have relatively more consequence for student achievement gains. Specifically, classroom management may be more important for math achievement than the instructional skills defined by the Framework for Teaching, and achievement in reading may be more responsive to teachers' questioning skills than to their instructional planning.

Gill, Shoji, Cohen, & Place (2016) examined the relationship between value-added and seven dimensions of teaching practice that are commonly identified across five frequently-used observation instruments. Each dimension was at least somewhat predictive of value added, with correlations ranging between 0.13 and 0.28—similar correlation values to those observed for the instruments as a whole. They found that classroom management was the most strongly and consistently correlated with value-added scores, but it was only modestly better correlated with value added than other dimensions of teaching practice.

While most of the research on relationships between teaching practice and student learning focuses on a single assessment, two studies (Cohen, 2015; Grossman et al., 2014) have explored this relationship across distinct exams. In one of these studies, Cohen (2015), using data from the Measures of Effective Teaching (MET) project, found that teachers' scores on two domains of the PLATO observation rubric were predictive of students' achievement on a high-stakes math test: procedural strategy instruction (e.g., rules, algorithms, and formulas to use when approaching academic tasks) and modeling. When Cohen examined these findings alongside a qualitative review of these teachers' practices, she found that the relationship between teachers' performance on these two domains and their value-added may have been driven by educators teaching test-taking strategies in ways that fit into the rubric's definition of *procedural strategy instruction* and *modeling*, rather than using these two practices to encourage students' deeper understanding of the content. This could support previous findings (e.g., Corcoran et al., 2013) that there is greater persistence of teacher effects for low-stakes tests than high-stakes tests, a result that may reflect teachers focusing on developing short-term (e.g., test-taking) skills in order to induce student performance gains under high-stakes conditions.

In a similar analysis, Grossman et al. (2014) also used MET data to examine the relationship between teachers' PLATO scores and their value-added from two different assessments, this time in ELA. In this analysis, value-added scores that were calculated using the lower-stakes, mixed-format assessment were more highly correlated with PLATO scores than the all-multiple-choice state assessment. This difference was associated with teachers' ratings on the *cognitive and disciplinary demand of classroom talk and tasks* domain of the observation rubric, which may reflect better ability of that assessment—which includes problem-solving within the test construct and is comprised of constructed-response items—to capture deeper learning skills.

These studies together yield mixed evidence about what one might expect to see in DCPS, and on PARCC relative to CAS. Differences in the practices found to be more meaningful for student achievement may reflect differences in study methodologies, though possibly also variance in the relationship between practice and achievement across observation rubrics and assessments. Certainly, the Cohen (2015) and the Grossman et al. (2014) findings suggest that, even with a common teacher evaluation rubric, the relative importance of certain teaching practices can vary according to the assessment used to evaluate teachers' effectiveness.

Teaching Under the Common Core Exams

Regardless of whether teachers' practice may be differentially related to student learning across the exams, a separate question remains as to whether the transition has caused teachers to change their practice in response to new expectations for their teaching. Proponents of the Common Core expected the shift in focus from the new standards and assessments to require a substantial change in instruction (Conley, 2014; Student Achievement Partners, 2013 & 2014), with an increased emphasis on the complexity and depth of children's learning. Given that assessments are themselves important drivers of teachers' practice decisions (Cunningham, 2014; Jennings & Lauen, 2016), the transition to the PARCC exam may have catalyzed a shift in instructional emphasis in DCPS toward the teaching practices that focus on conceptual understanding and depth of learning. Indeed, in a survey of educators across five states that adopted the PARCC and SBAC assessments, large majorities of teachers reported changing their instruction or more than half of their instructional materials in response at least in part to the new assessments (Kane, Owens, Marinell, Thal, & Staiger, 2016); in another national survey administered as states were transitioning to the new standards, a large majority of teachers expected the CCSS to require them to change their instruction by teaching more conceptually, and more than 90% expected the new CCSS-aligned assessments to influence their instruction (McDuffie et al., 2017).

While a small handful of qualitative studies have recently begun to explore the effect of this transition on teachers' practice (Ajayi, 2016; Stosich, 2018), there has been to date no empirical evidence as to whether these exams have actually caused teachers to alter their instructional emphasis. The qualitative literature, however, suggests that teachers had difficulty adapting to the new standards and exams. Localized surveys and interviews with teachers (Ajayi, 2016; Stosich, 2018) and with their students (Kolluri, 2018) provide suggestive evidence that, while there may be variability in the extent to which teachers succeeded in shifting their practice, many teachers struggled during the transition to alter their teaching to effectively emphasize conceptual learning in the classroom.

The limited literature that has looked at this question so far suggests that there may be differences in performance across TLF-defined practices during the transition. However, the probable direction of that effect is not clear. Educators in DCPS who teach in CCSS subjects (i.e., math and ELA) may exhibit a drop in TLF performance as they transition and adapt their teaching to the new exam, but there may also be performance gains for CCSS-aligned practices as teachers shift from an emphasis on procedural instruction to conceptual instruction.

Hypotheses

Each of the three questions I pose in this paper will provide context for the others. Understanding how—and whether—PARCC caused changes in teachers' practice will explain potential shifts in teachers' value-added rankings and illuminate the practices where teachers either struggled to excel during the transition or those where they succeeded in shifting their instructional emphasis. Likewise, if certain practices are differentially associated with student achievement on the PARCC exam relative to CAS, it could explain observed movement across the value-added distribution during the transition.

Once student achievement scores are linked across exams, there are several associations between student achievement and teachers' practice that one might reasonably expect. For example, if the CCSS-aligned tests are sensitive to instructional practice, they may better capture effective teaching—as defined by generating deeper learning—than traditional standardized tests, and the type of learning emphasized by the

CCSS may prioritize certain teaching practices over others. Teachers in tested grades and subjects might also shift their teaching emphasis toward those practices where they expect higher returns on the new exam, or the quality of their teaching might suffer if they struggle to adapt to the new standards and assessments.

Given the CCSS's focus on high standards, deeper, conceptual understanding, and critical thinking—and the PARCC exam's development around these goals, I hypothesize that the Teach standards that specifically reference deeper and conceptual understanding will be more associated with student learning under PARCC than CAS, and therefore may also be where teachers focus their instruction following the transition to PARCC. This is not simply because of potentially different content emphases across the exams, but also differences in structure and item format. Traditional, selected-response assessments (e.g., CAS) are generally less sensitive to students' conceptual understanding relative to their procedural skills than assessments that rely heavily on more-complex item formats (e.g., PARCC); on a traditional assessment, students can fit a response to the limited set of answers provided—or redo the problem until they find a solution that fits one of the available responses. In traditional standardized assessments, students do not need to demonstrate how or why they have arrived at a given answer; the new CCSS-exams, in contrast, often ask students to provide justification for answers or ask scaffolded questions to demonstrate a student's thinking.

There are four Teach standards on the TLF that I identify a priori as being potentially more important for achievement on PARCC than on CAS. The first of these, Teach 4 (*provide students multiple ways toward mastery*), emphasizes that students develop deep understanding through multiple ways of engaging with the content. These "multiple ways" can include engaging students "through a variety of learning styles, modalities (auditory, visual, kinesthetic/tactile), and intelligences (spatial, linguistic, logical-mathematical […])," but must serve to develop students' deep understanding of the content. PARCC correspondingly attempts to assess students' understanding across multiple ways of representing and engaging with a specific learning standard, where appropriate, in order to evaluate students' depth of understanding. Students who have been taught material in different formats that allow them to engage repeatedly with content in different ways should theoretically better understand how or why an answer is what it is; likewise, they should more fluently be able to answer questions that measure content in different formats. For example, the grade 3 math standard 3.OA.A.1, which requires that students be able to *interpret products of whole numbers*, could be assessed logically, as in the top panel of figure 1.2, or visually, as in the second panel. While both items assess the same CCSS standard, they do so in different ways. In ELA, the new assessments might incorporate images or other media along-side written texts to assess how well students can analyze the ways in which different media can support or develop the meaning of a text; an example of this is conveyed in the sample item for grade 5 ELA standard RL.5.7 in figure 1.3.

The remaining three standards which might be relatively more important for CAS achievement—Teach 5, Teach 6, and Teach 7—each highlight teaching that emphasizes students' depth of understanding. Specifically, Teach 5 (*check for student understanding*) explicitly defines checks for understanding in terms of ascertaining the *depth* of students' understanding. Teach 6 (*respond to student understanding*) describes not just whether teachers catch and correct misunderstandings, but also whether they probe correct responses to ensure that students understand the content. Finally, Teach 7 (*develop higher-level understanding through effective questioning*) may be more important for achievement under PARCC than CAS, given that the TLF defines effective teaching under this standard as posing increasingly complex questions, following up with strategies to support understanding, and eliciting meaningful responses from students. If evaluators interpret these standards with fealty to their intent, as opposed to a more superficial or less rigorous understanding of these terms (e.g., Hill, 2001), then students whose teachers perform relatively higher on these standards should also perform relatively higher on the PARCC exam.

DATA

To answer these questions, I use an administrative dataset from DCPS which contains student-level data from AY2009-10 through AY2015-16, including the students' demographic characteristics (e.g., race/ethnicity, gender, age, grade level, ELL status, and special education status) and academic achievement—with CAS scores from AY2006-07 through AY2013-14 and PARCC scores for AY2014-15 and AY2015-16. DCPS also provides linking and dosage rosters which connect individual students to their teachers in tested grades and subjects, and allows for the adjustment of teacher effects by the amount of time each student spends with his or her teacher. The teacher-level data include teachers' race/ethnicity, gender, age, and teaching experience. See table 1.1 for descriptive statistics on the analytic database.

In addition, I have data on teachers' performance across several evaluation measures, including the TLF and its nine subcomponents. Because the average teachers' TLF performance has shifted somewhat over the past several years—which may be attributable to shifts in teacher quality within the district, but also possibly changes in how the rubric is operationalized over time—I standardized TLF scores within year when estimating the differential associations between teaching practices and student outcomes across the two exams.

Given, however, that there are a large number of sub-scores and these teaching standards are correlated with one another (see appendix table A.1.2), I use a principal component factor analysis to reduce the number of TLF dimensions from nine to two. Though commonly-accepted thresholds for factor loadings yield just one factor (see Adnot, 2016), I force the data to load onto a second factor (appendix table A.1.3). This produces a dominant factor that is highly correlated with the first seven, *instruction*-oriented, TLF practices; the secondary factor captures the TLF components (Teach 8 and Teach 9) that address the *classroom environment*. Though I rely primarily on the two-dimensional TLF factors for my analyses, I include results from the full set of TLF sub-scores in the appendices to this paper.

In addition, while teachers are evaluated by a combination of internal (administrator) and external ("Master Educator") evaluators, I limit the analysis of TLF performance to the scores assigned by external evaluators. I do this because, while school administrators typically assign more reliable scores, external evaluators generally assign scores that are more strongly associated with objective measures of teacher quality, even when adjusting for reliability (Ho & Kane, 2013; Gill et al., 2016; Meyer, 2016; Whitehurst et al., 2014); the external evaluators' scores are likewise less subject to ceiling effects, as administrators in this sample tend to rate teachers' performance more highly than the master educators (MEs).

RESEARCH QUESTION 1

Methods

The first question I explore in this paper is whether PARCC and CAS rank teachers differently according to their value added. While the data include value-added scores calculated by DCPS, the district took a hiatus from using value-added scores during the PARCC transition, and official scores were therefore not provided for AY2014-15 or AY2015-16. Instead, I estimate new value-added scores across all years of the analysis. To do so, I first standardize student achievement scores by subject, year, and grade. I then use the *tfxreg* Stata program (Cowan, 2017) to estimate value-added scores separately by year and by subject.[5] I include in the value-added model controls for students' same-subject and opposite-subject lagged test scores, gender, an indicator for whether the student changed schools, ELL status, free- or reduced-price-lunch eligibility, special education status, prior-year absences, and grade level. I then use Mathematica's publicly-available *eb_shrinkage*[6] Stata program for empirical Bayes shrinkage to shrink less-precise teacher effect estimates toward the mean.

I take three approaches to understand whether PARCC and CAS rank teachers differently according to their value added. First, I correlate year-by-year value-added scores to determine if there are significant differences in these correlations across the transition to the new assessment. While the instability typically observed in value-added scores precludes the possibility of teachers being identically ranked each year, a

[5] This program is adapted from Mihaly, McCaffrey, Lockwood, and Sass's (2010) *felsdvregdm* Stata program, which improves on standard fixed-effects estimation programs that typically estimate effects relative to a left-out reference teacher, which can affect standard error estimates—by using sum-to-zero constraints. Cowan's *tfxreg* program is nearly identical to *felsdvregdm*, except that it more-efficiently inverts $X'X$ to estimate teacher effects.

[6] https://www.edimpactlab.com/programmer-resources/free-program-code

reduction in the Spearman rank correlation between adjacent-year scores at the transition would provide support for DCPS's decision to temporarily omit value-added scores as an accountability measure.

This adjacent-year-correlations approach allows me to use the full sample of teachers with value-added scores across this period, but also introduces two drawbacks. The first is that a single year of value-added scores has substantially higher sampling error than a multi-year average (McCaffrey et al., 2009), and the second drawback is that these correlations may be capturing changes in scores that are acquired through experience, given that novice teachers' value added improves substantially in their first few years of teaching. Such confounding would be of particular concern if teachers were differently assigned to teach in tested grades and subjects during the PARCC transition.[7]

I can also explore this relationship in terms of teachers' value-added rank. I do this first by defining the corresponding percentile rank of each teacher's value-added score for a given subject and year, and estimating (1), where I regress the absolute change in teacher k's year t percentile rank relative to the prior year $(t-1)$, ΔPR_{kt}, on an indicator for the transition year (AY2014-15), as in Backes et al. (2018).

$$|\Delta PR_{kt}| = \beta_0 + \beta_1 PARCC_{kt} + \epsilon_{kt} \tag{1}$$

This approach allows me to identify whether there are differences in the relationship between value-added scores across the exams for certain types of teachers. To detect whether this is the case, I add an interaction term for characteristics of the teacher (i.e., experience, year $t-1$ IMPACT score quintile, and year $t-1$ value-added score

[7] Regressions of an indicator for teaching in a tested grade and subject on teacher characteristics (experience, prior-year IMPACT rating, race, gender, and educational attainment) and interactions between teacher characteristics and whether the year is a PARCC exam year suggest that there was, at most, negligible sorting of this kind in DCPS.

quintile) and her students (i.e., share minority, ELL, or low-income).[8] I also test for robustness to teacher and school fixed effects to reduce bias from unobserved confounders. This model, however, is still susceptible to bias from changes in teachers' experience.

As a third approach, I identify teachers with at least three years of experience and at least two years of value-added scores on each assessment, average their value added within-exam, and regress average value-added scores on the PARCC exam on average value-added scores on the CAS exam—also plotting this relationship on a scatterplot—to understand how teachers' effectiveness, as measured by value-added scores, compares across the two assessments:

$$\overline{VA}_k^{PARCC} = \overline{VA}_k^{CAS} + \epsilon_k \qquad (2)$$

By restricting the sample to teachers with at least three years of experience—after which teachers' returns to experience on value-added scores begin to diminish (Boyd et al., 2008; Harris & Sass, 2011; Papay & Kraft, 2015; Rockoff, 2004; Wiswall, 2013)—I mitigate bias from concurrent changes in experience. By averaging teachers' value-added scores within exam, I also reduce the measurement error associated with value-added scores (McCaffrey et al., 2009). Both of these steps, however, reduce the sample size to

[8] Due to the small number of teachers at key incentive thresholds under teach testing regime, I am insufficiently powered to isolate an incentive effect on teachers' value-added to determine the extent to which DCPS's unique incentive structure may affect my results. I suspect, however, that any difference in incentive effects on value-added during this period would be negligible given that the largest weight in IMPACT is assigned to observation scores. DCPS's observation rubric describes exemplary practice across each of the nine TLF teaching standards, which could provide teachers with guidance on how they might improve their practice; in addition, formal observations in DCPS are followed within two weeks with written and verbal feedback on their practice. Value-added scores, meanwhile, provide only summative evidence of teachers' performance in contrast to the in-part formative nature of classroom observations. Indeed, in an analysis of the effects of performance incentives in DCPS on teachers' practice, Adnot (2016) found that teachers near key thresholds (i.e., those that place teachers either at risk of dismissal or with the potential for large financial rewards) make significant gains on the TLF the following year.

about 70 teachers in each subject, potentially limiting statistical power and external validity. On the other hand, this method allows for a better understanding of the functional form of the relationship between value-added scores across these assessments; for example, it may reveal non-linearity in the association between teachers' measured effectiveness across the exams.

Results

Figure 1.4 illustrates the changes in Spearman rank correlations between adjacent-year value-added scores in math and ELA across the years of analysis. In both subjects, these correlations are imprecisely estimated in each year, generally hovering around 0.40. There is no apparent jump or drop in correlations at the transition to the PARCC exam for either subject, while there appears to be an increased correlation in adjacent-year value-added scores for math teachers in 2014—before the adoption of PARCC—relative to earlier years.[9] This higher correlation coefficient, however, may be a function of differences in which teachers have adjacent-year value-added scores from year to year; a test of the equivalence of these correlation coefficients in 2013 and 2014 that is restricted to teachers with both current and prior-year value-added scores in those years indicates no statistically discernable difference in these correlations.

The next analytic approach that I employ to understand the relative stability of value-added scores under the new testing regime uses changes in teachers' ranking. Table 1.2 shows the additional instability—as defined by absolute change in percentile rank from one year to the next—associated with testing under PARCC relative to CAS. Overall, instability effects are small, generally non-significant, and robust to the inclusion

[9] Spearman rank correlations estimated with dis-attenuation for measurement error in value-added scores produce qualitatively similar correlation coefficients to those not adjusted for measurement error.

of time-variant teacher and student controls, as well as school and teacher fixed effects. Only one model in math (the third column) yields a statistically significant effect, though this effect is substantively small (approximately three percentile points) and could be confounded by differences in the population of teachers across the two exams; within-teacher models, on the other hand, demonstrate a level of instability under PARCC that is statistically no different from instability under CAS.

These overall effects might, however, mask heterogeneity given that the literature demonstrates that certain types of teachers (e.g., lower-performing teachers, or those teaching lower-achieving students) may experience greater instability in value-added scores from year to year. Tables 1.3a and 1.3b, which display heterogeneity of instability across teacher and class characteristics, show little evidence of changes to the stability of value-added with the new assessment. Results are sensitive to model choice, but differences are generally small. In math, instability effects are statistically no different for zero when estimated within teacher and school; in ELA, novice teachers may experience an increase in instability (16.62 percentile points; $p < 0.05$), as might teachers with higher-shares of black students (36.16 percentile points, $p < 0.10$), while teachers with high-FRPL classrooms may experience declines in instability (27.04 percentile points; $p < 0.10$). Even the largest point estimates, however, are not robust to adjustments for multiple hypothesis testing, after which none of the subgroups demonstrate statistically significant differences in value-added stability across the exams.

It is not just the test itself that was of concern for teachers and school administrators, however. Rather, they were also concerned about the fairness of using value-added scores to evaluate teachers when these scores were estimated using a new assessment—and before students and teachers had an opportunity to familiarize themselves with the new test. In fact, models where I estimate whether the *transition* (i.e., the first year of the PARCC exam) was associated with changes in year-to-year stability reveal similarly non-significant effects, either overall (appendix table A.1.4) or for subgroups of teachers (appendix tables A.1.5a and A.1.5b).

While these results are consistent with the lack of meaningful change in adjacent-year correlations following the adoption of the new assessment, this approach to estimating effects on value-added is susceptible to bias from changes in teachers' experience. Effects may also be attenuated because of the measurement error associated with value-added scores from one year as opposed to multiple years. My third approach to answering this research question attempts to mitigate both of these problems, given that it averages value-added scores across multiple years and uses a restricted sample of more-experienced teachers who are less likely to make meaningful gains from year to year. In figure 1.5, I plot the relationship between value-added scores across the two exams. For both subjects, this relationship is noisy, but linear.[10] If there were no difference in teachers' relative value-added scores across the tests, we might observe a slope equal to 1 (plotted as a dashed line in figure 1.5); however, the OLS regression estimates illustrated by solid lines in figure 1.5 demonstrate a shallower relationship, with regression estimates of 0.51 in math and 0.55 in ELA.[11] In both subjects, teachers with more extreme value-added scores on the CAS exam have less extreme scores on PARCC.

[10] I also run models with quadratics of mean CAS scores; coefficients on these higher-order polynomials are non-significant and add no explanatory power to regression models for either subject.

[11] Tests of equality of the slope coefficients to 1 also indicate that the slopes are statistically different from 1 at $\alpha = 0.05$.

This result is robust to restrictions on the sample by experience and the number of years of value-added scores that are averaged across exams (see appendix table A.1.6).[12]

To put these results in context, a math teacher in this restricted sample who averages one standard deviation above the average on CAS in terms of value-added scores would have an average percentile rank of 78 on that exam, but achieve a value-added score only 51 percent of a standard deviation above the average on PARCC; this PARCC value-added score is equivalent to an average percentile rank of 65, a decline of 12 percentile points across the exams. Similarly, an ELA teacher who averages one standard deviation above the mean on CAS would have an average percentile rank of 79 on that exam, but achieve a value-added score only 55 percent of a standard deviation above the mean on PARCC, equivalent to an average percentile rank of 65 on that exam—likewise a decline of 12 percentile points across the exams.

This result supports the findings from the preceding analyses in that it demonstrates that teachers' value-added scores on one exam (i.e., CAS) are related to their value-added scores on another exam (i.e., PARCC). The size of the slope is consistent with the adjacent-year correlations between PARCC and CAS value-added scores shown in figure 4, and the linearity of this relationship across levels of CAS value-added likewise supports the lack of heterogeneity by performance level in inter-year stability of value-added scores across the assessments.

[12] To explore sensitivity to value-added modeling decisions, I test for robustness across each of the three methods described to alternative value-added estimates. Specifically, I estimate: 1) a version of my value-added model that explicitly incorporates teachers' experience; 2) a two-step aggregated residuals approach (see Koedel et al., 2015 for a description of this method); 3) a value-added model that accounts for "drift" in teacher quality over time (see Chetty, Friedman, & Rockoff, 2014a); and 4) the method developed by Mathematica Policy Research for use in DCPS before the PARCC transition (Isenberg & Walsh, 2014). While some methods produce somewhat more precise value-added estimates than others, none yields qualitatively different findings from those described here.

However, these scatterplots and corresponding regression coefficients demonstrate potentially meaningfully different magnitudes of differences in value-added scores across the exams. These differences could have multiple different sources. First, this analysis relies on a smaller and more-experienced sample of teachers than the preceding approaches. These teachers may be more effective, given that more-experienced teachers tend to have stronger effects on student achievement (Boyd et al., 2008; Harris & Sass, 2011; Papay & Kraft, 2015), but they may have more difficulty on average adjusting their practice to new expectations if they have had more time to become ingrained in their ways of teaching. This would make even seasoned teachers less effective on the new exam until or unless they adapted their teaching to the new assessment. Second, the variation might also reflect sizeable remaining measurement error or experience effects that preclude a one-to-one relationship in value-added scores across the assessments. For example, if I instead estimate the relationship between average value-added scores for teachers who taught for multiple years on the same exam (e.g., comparing average CAS value-added scores in one set of years to average CAS value-added scores in another set of years), I find similar and only marginally higher slopes, equal to 0.57 ($SE = 0.12; n = 80$) in math and 0.63 ($SE = 0.10;\ n = 87$) in ELA; these estimates are similarly statistically different from a slope equal to 1; this may reflect that other factors outside of the test, such as confounding characteristics like teachers' experience, as well as measurement error in value-added scores, may be an important source of differences in teachers' performance across the CAS and PARCC exams.

All together, these three approaches demonstrate that the transition to the new exam produced little if any discernable changes to the ways in which teachers were ranked according to their value-added scores relative to changes that we would have observed under a consistent testing regime. In spite of great consternation across stakeholders about the fairness of evaluating teachers with value-added scores on a new exam, the PARCC exam appears not to have led to a statistically discernible difference in value-added rankings for the average teacher in DCPS beyond what we would have seen had DCPS continued using CAS.

The scatterplots nevertheless reveal that many teachers perform differently on one assessment than they had on the other; meanwhile, other teachers may achieve similar levels of performance in terms of their effects on student achievement if they are skilled at adapting their teaching to new contexts such as the transition to PARCC. This analysis does not attempt to understand what may explain teachers' relative effectiveness across the two assessments, but it is possible that teachers' skills are differentially important for student achievement across the two exams. This is the focus of the next research question: whether students have stronger relative learning gains on the PARCC exam than on the CAS exam when they have teachers who are relatively stronger on certain skills.

RESEARCH QUESTION 2

Methods

If one were estimating the association between teaching practices and achievement scores for each test separately, one might estimate something like the following two equations individually, for student i with teacher k in grade (cohort) g and subject s in year t, where the superscripts P and C represent the PARCC and CAS exams, respectively:

$$A^{P}_{ksgti} = \beta^{P}_{0} + \beta^{P}_{1} A^{P}_{ksg[t-1]i} + \beta^{P}_{2} TLF^{P}_{ksg[t-1]i} + \boldsymbol{X}_{ksgti}\beta^{P}_{3} + \gamma_g + \tau_i + \varepsilon_{ksgti} \qquad (3)$$

$$A^{C}_{ksgti} = \beta^{C}_{0} + \beta^{C}_{1} A^{C}_{ksg[t-1]i} + \beta^{C}_{2} TLF^{C}_{ksg[t-1]i} + \boldsymbol{X}_{ksgti}\beta^{C}_{0} + \gamma_g + \tau_i + \varepsilon_{ksgti} \qquad (4)$$

Using this method, β^{P}_{2} is the estimated effect of a one standard deviation (SD) increase in a teacher's TLF score on a student's PARCC achievement (also in standardized units), while β^{C}_{2} is the estimated effect on CAS achievement. Each model includes grade fixed effects (γ_g), student fixed effects (τ_i), and a vector of time-variant student characteristics ($\boldsymbol{X}_{ksgti}$), such as free or reduced-price lunch (FRPL) status, ELL status, and special education status. I run these regressions separately for math and ELA, and TLF scores are standardized within year.

To answer this research question, however, one needs to understand the difference in the slopes on TLF scores across each assessment (i.e., $\beta^{P}_{2} - \beta^{C}_{2}$), which can be more-efficiently estimated using equation (5), where the coefficient on $PARCC * TLF_{ksgti}$ (β_6) is equal to the difference in TLF slopes across the two assessments. The estimate for β_6 could then be interpreted as the additional effect of each SD unit of TLF scores on student achievement on PARCC relative to CAS.

$$\begin{aligned} A_{ksgti} = \beta_0 + \beta_1 A_{ksg[t-1]i} + \beta_2 TLF_{ksgti} + \beta_3 PARCC_{ksgti} + \boldsymbol{X}_{ksgti}\beta_4 \qquad (5) \\ + \gamma_g + \tau_i + \beta_5 PARCC * A_{ksg[t-1]i} + \beta_6 PARCC * TLF_{ksgti} \\ + PARCC * \boldsymbol{X}_{ksgti}\beta_7 + PARCC * \gamma_g + PARCC * \tau_i + \varepsilon_{ksgti} \end{aligned}$$

I estimate the regression model for the TLF overall, and for the *instruction* and *classroom environment* dimensions of the TLF identified through factor analysis, adjusting for multiple-hypothesis testing with a Bonferroni correction.

Using the overall TLF and the distinct teaching dimensions (*instruction* and *classroom environment*) as outcomes, I employ a series of specifications of equation 5 which address different trade-offs between precision and bias. The ability to generate precise estimates of β_6 is a concern across all specifications, because while the overall TLF has relatively high reliability for a classroom observation measure, the subdomains will have lower reliability (i.e., higher measurement error) which could attenuate estimates (Meyer, 2016).[13] At the same time, there are a number of potentially endogenous traits—such as differences in teacher quality and student demographics across the two testing regimes—that I wish to control for in order to mitigate bias.

In the preferred specification, I include student fixed effects (τ_i). By estimating within-student variation in PARCC and TLF interaction effects, I can rule out differences in the testing population across the transition. This specification will limit the sample size and impede external validity given that it requires students with at least four consecutive years of test scores—a year t and year $t - 1$ score on each exam—so I also run this specification without student fixed effects.

I also explore the inclusion of teacher fixed effects (τ_k) in lieu of student fixed effects to leverage within-teacher variation in TLF scores to estimate these differences. For several reasons, however, this is not the preferred specification. First, given that there are relatively few teachers who remain teaching in tested grades and subjects over time in DCPS, this specification may further lower my power to detect meaningful differences. Second, I do not know whether the variation I detect is attributable to measurement error,

[13] Reliability for the overall TLF averages about .72 across years, based on a generalizability study conducted by Meyer (2016) for the first five years of IMPACT. Reliability has not been estimated for the nine individual subdomains.

which would not affect the bias of estimates, or whether evaluators were changing their operationalization of the TLF in response to the new tests and standards, which could bias estimates and understate differences in the relative importance of teaching practices across the exams.

My primary analysis for this research question omits data from AY2014-15 to mitigate potential differences in how lagged and current achievement are associated with each other between and across exams as well as differences that might be attributable to disruption effects from the first year of the exam. As a robustness check, I also estimate results using the full panel of PARCC-year data.

I still face, however, a nontrivial barrier to making comparisons across these assessments; as PARCC and CAS are unique exams, I cannot directly compare the distribution of student achievement across the two tests, even with standardization, without meeting strict assumptions about the interval properties of the two exams, among other requirements (Ost, Gangopadhyaya, & Schiman, 2017; Penney, 2007).

Ideally, I would have item-level data from similarly- and concurrently-administered exams which either contain a common set of items which I would use as anchors from which to equate scores across the two tests, or the tests would have been administered to the same students at the same time. Such a method adjusts for both differences in examinees (ability) and in the tests (difficulty). Without common items, however, and with different examinee populations across the two tests, I cannot assume randomly equivalent groups and therefore am unable to distinguish differences in exam difficulty across the transition from examinees' abilities. I am also limited in that I have access only to students' overall scaled scores, and not item-level data.

There are, however, options that may allow me to approximate[14] equated scores in order to answer this research question, three of which I explore in this paper. The first is to use propensity score matching (PSM) to create pseudo-equivalent groups; the second is to translate scores on each test to a common distribution using the National Assessment of Educational Progress (NAEP) as a benchmark; and the third is to use within-student performance to predict achievement had the test not changed, and transform 2015 PARCC scores to the distribution of predicted CAS achievement. Each of these methods is described briefly below and in more detail—with a discussion of the benefits and limitations of each—in appendix B.

Linking scores through propensity score matching

A common problem with comparing non-anchored exam scores is that one cannot assume that the underlying distributions are identical, in part because the groups completing each exam are not identically distributed across test questions. Indeed, in DC that may be a concern because of a steady shift in the demographics of enrolled (and tested) students in the city's public schools. There may also be trends in other student characteristics not captured in DCPS's administrative data occurring during this time. Meanwhile, these demographic shifts transpired during the years when DCPS transitioned to PARCC, yet there is no common set of items from which I can equate the performance distribution across exams. One potential method for tackling this issue is to use propensity scores to match students tested on one exam to the other, and link score

[14] Because the tests differ in more than just the individual items that make up the exam (i.e., format, content weighting, scoring, etc.), this process is technically referred to as linking rather than equating (Kolen & Brennan, 2014).

distributions across matched groups (Haberman, 2015; Kim & Lu, 2018; Longford, 2015).

This process follows three general steps, done separately for each tested grade and subject. First, I create a sample of matched pairs using propensity scores from pre-treatment characteristics, matching as many PARCC students to CAS students as possible. I limit this process to adjacent years (2014 CAS v. 2015 PARCC) to further minimize differences in the testing populations. This step establishes pseudo-equivalent groups across the exams. I then define an equipercentile linking function, which transforms the distribution of the matched sample's PARCC scores from 2015 to that of the CAS scores from 2014. This function is then applied to the full PARCC sample, transforming their scores to the CAS scale—adjusting for differences in difficulty and testing populations across the two exams.

Translating scores on each test to the NAEP distribution

A second approach to linking scores across the exams is to use NAEP as a statistical moderator, similar to the method used by Reardon et al. (2019) to develop their national database of district- and subgroup-level academic achievement. NAEP is uniquely suited for this purpose because a) it was delivered to DCPS students in both CAS- and PARCC-tested subjects, grades, and years; and b) NAEP intentionally samples its Trial Urban District Assessment (TUDA) examinees, which include DCPS students, to be representative of students in their home district, such that the examinees comprising DCPS's NAEP results should be comparable to the fourth- and eighth-grade testing population within DCPS in the same year.

As with Reardon et al. (2019), this approach first requires linearly interpolating NAEP means and standard deviations from grades 4 and 8, which are commonly-tested grades across NAEP, PARCC, and CAS, to grades 5 through 7, and extrapolating to grade 3. I then interpolate NAEP scores for each grade and subject in even (non-tested) years using score distributions (means and standard deviations) from odd years, as the NAEP exam is only administered in odd-numbered years. Finally, I linearly transform each subject-year-grade standardized score to its corresponding score in the NAEP distribution, adjusting for measurement error in the CAS and PARCC exams using subject-grade stratified-alpha reliability estimates for the given test.

Linking scores across exams using predicted CAS achievement

The third approach attempts to leverage within-student variation in achievement on CAS to estimate the distribution of expected PARCC performance. Specifically, I regress students' CAS achievement in a given grade and year on their lagged performance on that exam with controls for a vector of observed student characteristics. Using coefficients from this regression, I then estimate predicted CAS scores for students taking PARCC in 2015 who have lagged CAS scores from 2014. I then apply the linking function for these students to the full sample of grade 4 through 8 PARCC examinees in 2015. This method may provide the weakest linkages (see appendix B), given that the CAS-linked (i.e., predicted) scores will be attenuated by measurement error, violating the symmetry assumption for equating (see Dorans & Holland, 2000; Holland & Dorans, 2006). Additionally, score expectations can only be defined for students with pre-treatment characteristics (e.g., lagged achievement).

Each method is technically feasible given the data available, but each introduces measurement error into the newly linked scores. However, these alternative approaches are preferable to simply standardizing scores by subject, grade, and exam/year, as each contributes to our ability to translate distributions across the exams without conflating differences in test difficulty and examinee ability. By using multiple methods to link scores, I may be better able to triangulate true differences in effects associated with the teaching practices measured by the TLF. I briefly discuss the quality of these linkages in the results section of this paper, and in more detail in appendix 1.B.

Results

Selecting a linking method

None of the linking methods I use for this analysis approaches a level of performance one would expect for a true equating, but the PSM approach generally outperforms the other two techniques when evaluated against the five assumptions laid out by Dorans and Holland (2000), suggesting that this is the most appropriate of the three choices for linking scores. For this reason, my preferred estimations are those that rely on the PSM method; however, I also estimate my models using NAEP- and regression-linked scores as robustness checks, as well as un-linked standardized scores. A full discussion of how each approach performs is included in appendix 1.B.

Findings

Table 4 shows results from PSM-linked scores. This table omits data from AY2014-15 to mitigate differences in how lagged and current achievement are associated with each other between and across exams as well as differences that might be attributable to disruption effects from the first year of the exam.[15] In math, overall performance on the TLF is associated with approximately equivalent gains on PARCC relative to CAS in the model that controls for neither student nor teacher fixed effects. The effect grows to 0.34 standard deviations higher achievement in PARCC relative to CAS when controlling for teacher fixed effects, though this point estimate is imprecisely estimated. Controlling for student fixed effects instead of teacher fixed effects, however, produces a similar but statistically precise result: students whose teachers score a SD higher on the TLF perform 0.039 SDs higher on PARCC relative to CAS ($p < 0.001$). While numerically small, this point estimate is not necessarily small in substantive magnitude; given that students in these grades typically experience gains of approximately 0.40 standard deviations across a year of learning (Hill, Bloom, Black, & Lipsey, 2008), this effect is equivalent to ten percent of a year of additional learning, or roughly three and a half weeks.

In ELA, the least-restrictive model for the overall TLF effects produces a small and marginally significant effect, where a SD increase in teachers' TLF scores is associated with a 1.7 percent of a SD increase in student achievement ($p > -.10$). In context, such an effect is roughly equivalent to six percent of annual learning in reading at these grade levels, or close to two additional weeks of schooling. Controlling for teacher fixed effects increases the magnitude of the point estimate nearly ten-fold, to 10.9 percent of a standard deviation ($p < 0.001$). The student-fixed-effects model, on the

[15] Appendix table A7 presents results from the preferred specification (controlling for time-varying student and teacher characteristics and time-invariant student characteristics) using data from all testing years. These results differ meaningfully from those that exclude the transition year in math, suggesting that the relationship between changes in student achievement and teachers' TLF scores was different during the transition to the new exam.

other hand, suggests no difference in the relationship between teachers' practice and student achievement across the two exams.

Across the TLF sub-domains produced by my factor analysis, evidence of the relationship between teachers' performance on the *instruction* domain and their students' math achievement across the exams is mixed. The least-restrictive model suggests that teachers' instructional skills are associated with higher CAS achievement than PARCC achievement (-0.029 standard deviations; $p < 0.05$), but this relationship is not robust to the inclusion of teacher or student fixed effects, where point estimates are statistically no different from zero. In ELA, teachers' performance on the *instruction* domain is likewise not differentially important for student achievement across the exams, regardless of the specification used.

Classroom environment effects in ELA are likewise null in the least-restrictive model, as well as in the model that includes student fixed effects, but large and positive ($\beta_6 = 0.145, p < 0.001$) when estimated with teacher fixed effects. In math, however, *classroom environment* appears more important for student achievement gains on PARCC than on CAS; teachers who score a SD higher on *classroom environment* have students scoring 0.039 standard deviations higher on the PARCC exam than on CAS when estimated only with controls for observed student and teacher characteristics. This effect remains positive, but imprecisely estimated with the inclusion of teacher fixed effects, and is large and statistically significant with the inclusion of student fixed effects ($\beta_6 = 0.064, p < 0.001$). Analyses at the original TLF sub-score level (appendix table A.1.8) suggest that these effects in math are being driven entirely by Teach 9 (*build a supportive learning-focused community*) rather than the more classroom-management-oriented component of the *classroom environment* factor (Teach 8: m*aximize instructional time*). Notably, Teach 9 is defined in the context of student behaviors rather than the teacher's activities and behaviors, as is the case for other domains defined in the TLF. For example, the first descriptor for this practice refers to students' investment in their work and valuing of academic success, as well as their investment in their peers. This sort of investment may be pivotal for students' motivation to learn and master more-difficult material in math. Similarly, the rubric identifies students' risk-taking—their willingness to take on challenges, eagerness to ask questions, and comfort engaging in constructive feedback with their peers. It may be that when teachers are able to facilitate learning environments in math where their students embrace learning challenges and support each other through rigorous learning they are more able to influence their students' conceptual growth than they are through such practices as questioning and checking for student understanding.

These patterns in the association between teachers' practice and student achievement across the exams are generally robust to the linking method used (appendix tables A.1.9a and A.1.9b). Teachers' practice, as defined by the *instruction* and *classroom environment* domains, is generally consistently signed within models and across linking methods (with the exception of results from the predicted-score approach, which tend to deviate a bit more from other linking methods' results), suggesting that model specification is more important for this research question than the method used to link scores across the two exams. Importantly, these results are not robust to the inclusion of school-administrator-assigned TLF scores (appendix table A.1.10), suggesting that ratings by external evaluators and school leaders may differently capture the relationship between student achievement and teachers' practice. Results that include all evaluators' scores produce estimates that vary in magnitude and direction across specifications; in ELA, the teacher-fixed-effect models imply that both factors are more meaningful for student achievement on PARCC, but with stronger effects from *classroom environment*. Meanwhile, the models that use just student and teacher controls and those that add student fixed effects produce results that are nearly the inverse of the teacher-fixed-effect models.

The question of which model best reduces bias is unclear. It is possible that teachers were differently sorted to students across the transition, though tests for sorting on observable characteristics give no indication of this occurring. My baseline models include controls for teacher characteristics, but there might also be sorting on unobservable teacher traits which could bias estimates. If this were true, the inclusion of teacher fixed effects would reduce such bias. Differential sorting could bias results in either direction. For example, if principals strategically shifted more experienced teachers into classrooms in tested grades and subjects, these teachers might be better at the skills not captured by—but correlated with—the TLF that also drive student learning (e.g., providing students with emotional support, engaging effectively with parents, and generating enthusiasm for subject matter). On the other hand, the teachers with greater skills in these areas may have chosen to shift from teaching into coaching and mentoring roles at the transition, possibly being replaced in the classroom with less experienced teachers.

Meanwhile, the population of students in DCPS has, without question, shifted over the period of my analysis.[16] The students in DCPS are less likely to be black and more likely to be white or Hispanic in PARCC years, and students were also shifting between traditional and charter public schools over this period.[17] These population shifts could very well correlate with changes in student achievement over time, as well as changes in teachers' TLF scores given that teaching evaluations can be biased by the students in the classroom (Campbell & Ronfeldt, 2018), and so the within-student model will address this likely cause of endogeneity. Given that there were documented shifts in the student population over time, the student fixed effects model likely better adjusts for bias than the teacher fixed effects model. However, given that there is not exploitable variation within teachers *and* students across the CAS-to-PARCC transition, the student fixed-effect model likely still suffers from omitted variable bias.

RESEARCH QUESTION 3

Methods

While it is not clear with certainty whether teachers' practice was differently associated with student learning on PARCC, the introduction of the new assessment may nevertheless have affected the quality of teachers' practice, given the general understanding at the time that the new CCSS-aligned assessments would require a substantial change in instruction (Conley, 2014; Kane et al., 2016; Student Achievement Partners, 2013 & 2014). The emphasis among educators, school leaders, and reformers on the dissimilarity of the new exams to the tests they were replacing could have affected

[16] See, for example https://dcps.dc.gov/page/dcps-glance-enrollment.

[17] Trends in charter versus traditional public school enrollment are available at https://www.dcpcsb.org/facts-and-figures-market-share.

teachers' practice in many ways. For example, teachers may have shifted their instructional emphasis to the practices for which they would expect relatively higher returns to student achievement on PARCC; on the other hand, they might have seen their practice suffer as they learned to adapt their instruction in real time. To understand the effects of the PARCC exam on teachers' practice, I employ a standard difference-in-differences (DiD) model, leveraging the transition to the new assessment for teachers in PARCC-tested subjects and grades versus other general education teachers in DCPS—who experienced no such transition—to estimate effects on teachers' practice. The model takes the following form:

$$Y_{ktms} = \beta_0 + \beta_1 G_{ktms} + \beta_2 PARCC_{ktms} + \beta_3 G_{ktms} * PARCC_{ktms} + \boldsymbol{X}_{ktms}\beta_4 + \tau_k + \varepsilon_{ktms} \tag{6}$$

In this model, Y_{ktms} is teacher k's TLF score in year t at school m while teaching subject s; G_{ktms} is an indicator for the teacher's status as a math or ELA teacher in tested grades; and $PARCC_{ktms}$ is an indicator for whether the teacher is teaching in a PARCC-exam year. β_3 represents the treatment effect. Preferred models control for a vector of pre-treatment teacher characteristics ($\boldsymbol{X}_{ktms}$) and teacher fixed effects (τ_k), though I also test for robustness to a school's estimated level of departmentalization and school fixed effects, as well as for the individual assigning a given teacher's TLF rating.

Treatment in this context is not simply the shift to the new exam for teachers in tested grades and subjects (referred to in DCPS as "group 1" teachers), but also a shift in the components making up those teachers' evaluation scores. From AY2009-10 when DCPS first rolled out its teacher evaluation program, IMPACT, through AY2013-14, group 1 teachers' performance in DCPS was evaluated in part by their contributions to student learning, as measured by individual value-added scores, in addition to their TLF scores, while their general education peers (group 2) were evaluated primarily on their TLF performance (see table 1.5). Overall IMPACT scores are used to make high-stakes decisions in DCPS; low-performing teachers can be terminated, while high-performing teachers are eligible for large financial rewards (Dee & Wyckoff, 2015). Because of concerns about teachers making the transition under such high-stakes circumstances, however, DCPS chose to omit value-added scores from teachers' overall performance ratings in the first two years in which PARCC was administered. The weight of student achievement on standardized exams for group 1 teachers was shifted to the TLF, effectively increasing the incentive for these teachers to perform well on the TLF while decreasing their incentive to improve student achievement. This co-occurring change in stakes for group 1 teachers is captured by β_3 along with any potential disruption effects or intentional changes in practice for the PARCC exam.

There are two key assumptions for internally valid estimates of the causal effect of the PARCC transition on teachers' observed practice. The first is that the discontinuity of treatment associated with the transition to the PARCC exam for teachers in tested grades and subjects is exogenous—that is, that changes in the probability of being a group 1 teacher in a PARCC exam year are as good as random. This assumption would be violated if, for example, teachers with higher teaching ability were disproportionately assigned to tested grades and subjects during the PARCC transition relative to the rates at which they taught such classes in prior years. I test this assumption through a series of covariate balance tests in which I replace the left-hand variable in (6) with pre-treatment teacher characteristics. Table 1.6 demonstrates good balance on teachers' gender, race, and education level, but potential imbalance by experience; for this reason, I include these teacher controls in my main model. The individuals rating teachers could also affect estimates of PARCC effects, biasing estimates, for instance, if the individual observers assigning TLF scores are differentially assigned to group 1 teachers at the transition and if raters differ in how they operationalize the rubric. To avoid confounding rater effects with PARCC effects, I estimate alternative specifications that control for individual raters.

The second assumption, which I first test graphically, is that of common trends in TLF scores for group 1 teachers and their general education peers (group 2) before treatment; there should be a parallel relationship in TLF scores over time across the two groups of teachers. If trends in TLF scores for the two groups from AY2010 (the start of IMPACT) through AY2014 (the last year of the CAS exam) were perfectly parallel, I would have strong evidence of conditional independence. This relationship is critical to the identification of PARCC effects. Figure 1.6 demonstrates that these trends are parallel, albeit noisy. Overall TLF scores moved in the same direction for group 1 as group 2 teachers before DCPS adopted the PARCC exam. The same is true for the *instruction* and *classroom environment* subdomains, where the graphs demonstrate TLF scores across the groups rising and falling roughly in tandem.

While visual evidence indicates generally parallel trends, a potential threat to the internal validity of these estimates is the trend toward departmentalization in DCPS, given that DCPS expected the specialization associated with departmentalization to make it such that teachers would be “better able to craft rigorous and engaging lessons for students” (DCPS, 2016). Neither data on which teachers were departmentalized nor departmentalization rates within schools were collected or maintained by the district, so to control for departmentalization effects I instead estimate the share of teachers in tested grades and subjects who have value-added scores in both math and ELA. School-year observations with high shares of teachers with value-added scores in both subjects can be assumed to have lower rates of departmentalization. This proxy variable suggests a trend toward departmentalization in the district over the period of my analyses. To avoid confounding PARCC effects with departmentalization effects, I run models that include this measure as a control. In addition, I test models that control for the possibility that teachers at different ability levels may have been differentially assigned to tested subjects and grades at the transition by including teacher fixed effects. I also test for robustness to rater fixed effects in case raters are differentially assigned to group 1 teachers at the transition.[18] Beyond the shift toward departmentalization, there were no apparent substantive policy changes in DCPS in 2015—outside of those that were bundled with the transition to the PARCC exam—which would have differentially affected one group of teachers over the other.

Other factors occurring before the transition, however, might influence group 1 teachers' scores differently from their group 2 peers. One such factor is a decrease in emphasis on value-added scores that occurred with a series of other structural changes to IMPACT in AY2012-13, where the weight of value-added scores on group 1 teachers' overall IMPACT scores was decreased from 50 to 35 percent, with some of that reallocation going toward TLF scores; the TLF weight for these teachers shifted from

[18] DCPS uses a rater training system, *Align*, to calibrate classroom observers. If the rater alignment process were working perfectly, it should not matter who rates an individual teacher. However, it is still possible that different raters might operationalize the TLF differently from each other.

35% in AY2011-12 to 40% in AY2012-13.[19] In case this reweighting of incentives toward the TLF or other structural shifts that occurred with the 2013 changes to IMPACT led to differential performance trends for group 1 and 2 teachers relative to the preceding years, I run specifications that omit the first three years of IMPACT.

Similarly, figure 1.6 suggests that teachers in both groups scored differently in *instruction* and *classroom environment* in the first year of IMPACT, AY2009-10, than in the years following. There are two readily apparent reasons why the first year of the panel might differ from subsequent years. First, two of the TLF domains, Teach 5 and Teach 9, originally consisted of three additional sub-scores; the rubric was then streamlined in the second year of IMPACT with those sub-scores collapsed to one score each for Teach 5 and Teach 9. This adjustment may have altered how the TLF was operationalized for each group of teachers relative to subsequent years. Second, evidence from the early years of IMPACT suggests that teachers responded differently in the first year of the program in part because they did not expect IMPACT to persist under political pressures at the time (Dee & Wyckoff, 2015). In case anomalous scores from the first year of IMPACT are distorting the slopes of group 1 and 2 trends, I run an additional set of specifications that omit only the first year of IMPACT.

In addition to visual and theoretical inspection for parallel trends, I also test for common trends empirically, regressing teachers' TLF scores on interactions between each year and teachers' group 1 status, omitting the last pre-treatment year (2014), and including year fixed effects as well as the other covariates used in my primary specification, as below:

[19] See chapter 2 of this dissertation for a summary of the changes that went into place in AY2012-13.

$$Y_{ktms} = \sum_{t \neq 2014} \beta_t G_{ktms} * PARCC_{ktms} + \boldsymbol{X}_{ktms}\beta + \tau_k + \theta_t + \varepsilon_{ktms} \tag{7}$$

If trends are parallel, the estimates for each value of β_t from $t = 2010$ through $t = 2013$ should be statistically no different from zero, as each coefficient represents the difference in TLF scores between group 1 and group 2 teachers relative to the difference in scores just before the implementation of the PARCC exam. Figure 1.7 plots the results of these tests, and suggests that, while the trends in *instruction* scores did not appear to deviate before PARCC, there may not have been parallel trends in TLF scores for *classroom environment;* potential violation of this assumption would bias estimates of classroom environment effects. For this reason, I rely primarily on results from *instruction* specifications and cannot say with confidence that *classroom environment* effects are internally valid.

In addition to the analyses described above, I test for robustness to alternative identification strategies by estimating a comparative interrupted time series (CITS) model; the CITS adds to equation 6 a set of controls and interactions for the year in which a TLF score is assigned, centered at the transition to PARCC. The CITS approach relaxes the conditional independence assumption, requiring at a minimum that the change in level and trend in group 2 teachers is the change in level and trend in TLF scores we would expect to observe had group 1 teachers not transitioned to PARCC. A key threat to this assumption is the possibility of a confounding instrumentation effect—that is, CITS results could be biased if raters were systematically changing how they operationalized the TLF over time for group 1 teachers, but not for group 2 teachers (e.g., raters might use changing benchmarks for what constitutes "depth of understanding" when implementing the rubric during the testing transition).

Results

Table 1.7 presents the results from my main DiD specifications, the first column of which displays estimates from a model that controls only for the level of departmentalization within a teacher's school.[20] This model indicates a decline in teachers' overall practice of roughly fifteen percent of a standard deviation, much of which appears to be driven by the *instruction* domain of the TLF, where teachers' performance declines by almost 17 percent of a standard deviation when they switch to PARCC. These overall-TLF effects are robust to the inclusion of teacher controls and school fixed effects, but are reduced to a statistical zero with the inclusion of teacher fixed effects—which is the preferred specification. Estimates from alternative specifications that control for rater fixed effects, rater and teacher fixed effects, and a combination of school, teacher, and rater fixed effects also indicate null results, as do models that include administrator-assigned TLF scores or exclude earlier years of IMPACT from the analysis (appendix table A.1.11). CITS estimates produce similar results to the DiD for overall TLF scores (appendix table A.1.12); results from a linear CITS specification with teacher fixed effects produce mixed effects directionally across the first two years of PARCC, though neither year's estimate is statistically different from zero. Estimates from a specification that allow for nonlinear (i.e., quadratic) trends in the pre-PARCC years suggest large negative effects on teachers' overall practice,

[20] A naïve model is shown in the first column of appendix table A11. Interestingly, although proponents of departmentalization would argue that departmentalization might improve teaching by allowing educators to specialize in a single subject area, point estimates are similar with and without controlling for departmentalization, suggesting that the level of departmentalization may have little effect on this relationship; instead this covariate only serves to improve the precision of estimated treatment effects.

predominantly in the second year of the exam. These results are consistent with the direction of effects from the primary (DiD) specification with teacher fixed effects.

At the sub-dimension level, the negative overall performance effects appear to be concentrated within the *instruction* factor. DiD estimates are consistent and precisely estimated across specifications, ranging from 13 to 19 percent of a standard deviation decline in teachers' instructional performance. Similar magnitude declines are estimated with the CITS approach. The consistency of these effects across specifications and the two methodological approaches provides strong evidence that the quality of *instruction* suffered for teachers in tested grades and subjects when the new exam was introduced. This result is consistent with teachers' concerns about their preparedness to teach to the new exam; if teachers had insufficient or poorly aligned instructional materials, they may have struggled to define and enact quality instruction in the context of the PARCC exam.

Meanwhile, results for the second dimension of the TLF suggest modest improvements in the quality of teachers' *classroom environments* when they transitioned to PARCC. While estimates from models that do not fully account for changes in teacher characteristics across the transition are null, the results from the specification with teacher fixed effects are of modest magnitude (0.08 SD, $p < 0.10$). Effects are of similar size and statistical significance when accounting for rater effects and school effects and when estimated with both internal and external evaluators' TLF scores, and are large and estimated with high precision when using restricted definitions of pre-treatment years (appendix table A.1.11). A CITS approach (appendix table A.1.12) similarly produces positive effects in the first year of PARCC, though the CITS point estimates are sensitive to model specification, where a quadratic approach suggests a large decline in *classroom environment* in the second year of the new exam.

It is not immediately apparent what might be driving the different PARCC effects across the two TLF dimensions, though a DiD analysis that separately explores each of the original nine Teach standards suggests that Teach 8—which is weighted more heavily in the *classroom environment* factor—is among the only TLF scores where group 1 teachers generally exhibit positive effects across specifications. It is possible that, in the absence of aligned curricular materials in the early years of the exam which teachers could use to guide their practice, these teachers struggled to align their instruction with their perceived expectations under PARCC. They may have instead focused their efforts on the practices where expectations and definitions were unlikely to change in the context of a new assessment—i.e., classroom management.

DISCUSSION AND CONCLUSION

The transition from traditional achievement tests to more rigorous, Common-Core aligned exams provides an opportunity to explore a variety of questions that could illuminate several aspects of teaching skills. In this paper, I explore three complimentary questions, each of which has implications for policy.

First, concerns were raised about the fairness of using value-added scores in teacher evaluations as states and districts transitioned from traditional standardized assessments to Common-Core-aligned exams. I find that as DCPS transitioned from CAS to PARCC, few teachers' value-added scores were affected by this transition, with only weak indication of effects for particular subgroups of teachers. This is consistent with findings from other settings in which changes in standardized assessments have generally had no meaningful effect on teachers' value-added scores (Backes et al., 2018). It is unclear why this transition did not make more of a difference. It may be that PARCC is not as different from its predecessor as expected, or that teachers who are skilled at raising student achievement on one assessment are good at raising student achievement more generally. It's also possible that random measurement error in the tests or the error attributable to value-added sampling may make it difficult to identify differences across assessments. If so, this error would protect against the differential ranking potential of distinct assessments, but might also mask meaningful differences in teachers' skills—already a concern of critics of value-added measures (Baker et al., 2010).

Given the seemingly large differences in skills students must demonstrate to succeed on PARCC relative to traditional exams, many believed that effective teaching would emphasize different skills. Here, too, the evidence is mixed. This may also in part reflect measurement error—both in the student assessments and in the observation rubric (Hill, Charalambous, & Kraft, 2012), which could attenuate estimates and make it more difficult to detect differences. It may also be that there are practices that teachers engage in that are differentially important for achievement across these exams, but which are not measured by the TLF. A recent study by Kane et al. (2016) suggests, for example, that the new CCSS-aligned tests may be more sensitive to instructional differences between teachers, though the authors of this study were unable to compare student learning gains to specific teaching practices.

These results highlight the importance of design and adoption of assessments and standards that are aligned to good teaching. When observation rubrics capture practices that generate learning gains, teachers can use these rubrics to guide and improve their practice. Assuming that PARCC captures higher-order skills that are important for college and career readiness, educators should be assessed according to teaching practices that are related to generating these skills among their students. As districts and schools move to improve achievement gains under the new learning standards, they may need to target their PD toward practices they may not have previously prioritized. Given the limited resources schools often have at their disposal, it is in their interest to shift their feedback and coaching efforts away from practices that may be less important for students' achievement and toward those that are more strongly associated with student learning under the current assessment regime. Unfortunately, however, this research produces conflicting evidence as to which practices those might be, although analysis at the TLF sub-score level (appendix table A.1.8) suggests that schools should target more of their PD resources for math teachers toward those practices and skills identified under Teach 9 (*building a learning-focused classroom community*).

An arguably more meaningful finding is that the transition to the new exam may have altered the quality of teachers' practice. For example, teachers improved their classroom environments, but at the same time experienced large declines in instruction. At minimum, the relative decline in group 1 teachers' *instruction* skills points to potential gaps in curricular preparedness. The tests upon which students are assessed often provide important information for teachers on how to operationalize the standards and expectations to which the assessments are aligned (Cunningham, 2014; Jennings & Lauen, 2016; McDuffie et al., 2017). Textbooks and other curricular materials are also key resources for teachers during major shifts like that of the transition to the CCSS (Kane et al., 2016; Polikoff, 2012) and for their practice in general (Charalambous, Hill,

& Mitchell, 2012), yet materials that claim alignment to the CCSS and PARCC do not always adhere well to the scope and intent of the new standards; they often overemphasize procedural over conceptual understanding relative to the proportional emphasis defined by the CCSS and PARCC (Polikoff, 2015). The poor quality of instructional materials initially available to teachers may have limited their ability to effectively design or otherwise implement new curricula. While teachers felt considerable pressure to adapt their instructional materials to their new testing environment (Kane et al., 2011), few teachers felt well prepared to help their students perform well on new exams like PARCC (Kane et al., 2016). An anonymous teacher was quoted in *Education Next* (Jochim & McGuinn, 2016), for example, lamenting that "We start testing on standards we're not teaching with curriculum we don't have on computers that don't exist."

Teachers' difficulties during the transition were not unknown to DCPS. In AY2016-17, following the receipt of results from the first year of PARCC testing, DCPS announced a major reform to its approach to PD. In part out of concern that its teachers were struggling to align their teaching with the CCSS, DCPS launched "LEarning together to Advance our Practice" (LEAP), an intensive PD program that provides grade- and subject-specific coaching and content support to all its teachers on a weekly basis. In other districts that have made the shift to PARCC or similar exams, Kane et al. (2016) have found that the schools that saw greater achievement on the new math assessments engaged their teachers in more frequent content-specific observations and feedback, held more days of PD, and included scores on CCSS-aligned tests in their teacher evaluations—all strategies that DCPS is using today. While Kane et al. (2016) are unable to control for all potential confounders in this relationship, these findings suggest PD such as LEAP may help teachers develop strategies to recover those practices upon which they floundered during the transition, as well as better align their teaching to the type of instruction that will enable students to excel on the standards laid out by the CCSS and assessed by PARCC.

Transitions in standards and assessments are not uncommon (Backes et al., 2018), and the research presented in this paper provides insights into what other districts might expect in future transitions, particularly in an era when measures such as value-added scores and classroom observations are commonly used to evaluate teachers' performance. This research adds to evidence that, in spite of great apprehension, such changes do not always matter for teachers' value-added. On the other hand, districts may want to more deeply consider the fairness of other measures—even those not directly linked to student achievement—when significant changes are made to standards and assessments. DCPS, like many districts during the transition to CCSS-aligned tests, shifted the weight of its evaluation measures away from student-achievement-based outcomes and toward classroom observation. Yet this research suggests that some of teachers' practice may have suffered during this transition. Teachers in tested grades and subjects may need more time and additional supports to adapt to new assessments in order for their performance on classroom observation measures to remain unhurt by the change.

TABLE 1.1

Analytic Sample of DCPS Teachers

		Group 1		Group 2	
Teacher Characteristic	All	CAS	PARCC	CAS	PARCC
TLF Score					
All observers	3.11	3.1	3.09	3.11	3.13
	(0.46)	(0.49)	(0.49)	(0.47)	(0.44)
Administrators only	3.19	3.17	3.2	3.19	3.21
	(0.52)	(0.53)	(0.51)	(0.53)	(0.48)
Master educators only	3.01	3.01	2.96	3.00	3.03
	(0.51)	(0.53)	(0.57)	(0.51)	(0.48)
Gender					
Female	0.72	0.75	0.74	0.71	0.72
Missing	0.03	0.03	0.01	0.03	0.02
Race/Ethnicity					
Black	0.51	0.52	0.52	0.51	0.49
White	0.31	0.30	0.32	0.31	0.32
Hispanic	0.04	0.03	0.04	0.04	0.06
Missing	0.10	0.12	0.09	0.10	0.08
Education					
Graduate degree	0.67	0.67	0.69	0.66	0.67
Missing	0.03	0.02	0.02	0.03	0.02
Experience					
0 - 3 years of experience	0.29	0.34	0.27	0.29	0.28
4 - 9 years of experience	0.28	0.03	0.36	0.25	0.31
10+ years of experience	0.40	0.37	0.32	0.45	0.35
Missing	0.02	0.00	0.06	0.01	0.06
Count	*15,808*	*2,003*	*1,278*	*8,838*	*3,689*

Notes. Statistics from analytic sample of teachers in DCPS between the 2009-10 and 2015-16 academic years (AY). Group 1 consists of teachers in tested grades and subjects; group 2 consists of all other general education teachers. The DC Comprehensive Assessment System (CAS) was in place through AY2013-14, after which DCPS switched to the Partnership for Assessment of College and Career Readiness (PARCC) exam. The Teaching and Learning Framework (TLF) score is a rubric-based classroom observation score, and possible scores range from 1 to 4. Standard deviations are in parentheses.

TABLE 1.2
Stability of Value-Added Percentile Ranks Across Exams

	MATH				ELA			
PARCC	-0.51	0.98	2.87^{*}	1.73	-1.79	-1.21	-1.24	-0.35
	(1.33)	(1.42)	(1.46)	(2.42)	(1.27)	(1.37)	(1.45)	(2.79)
Classroom and teacher controls		X	X	X		X	X	X
School FE			X	X			X	X
Teacher FE				X				X
Constant	24.33^{***}	14.14^{**}	11.33	12.12	24.53^{***}	-0.83	20.05^{*}	16.66
	(0.70)	(5.59)	(11.29)	(16.69)	(0.69)	(6.00)	(9.66)	(17.04)
n	1,111	1,111	1,111	1,111	1,168	1,168	1,168	1,168
R-squared	0.00	0.04	0.19	0.64	0.00	0.08	0.19	0.63

Notes. Classroom controls include the proportion of students who are male, black, Hispanic, another non-white race, eligible for free or reduced-price lunch, with limited English proficiency, or in special education, as well as the average lagged match test score and average lagged ELA test score. Teacher controls include experience level, prior IMPACT rating, and quintile of lagged value-added scores in the subject. PARCC is an indicator for years in which the PARCC exam was administered (i.e., AY2014-15 and AY2015-16).
*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; and $^{+}$ $p < 0.10$

TABLE 1.3a

Stability of Value-Added Percentile Ranks Across Exams, by Teacher Characteristics

	Math			ELA		
Teacher characteristics						
Experience < 3 years * PARCC	-8.44 **	-9.29 **	1.98	-4.27	-6.31 $^{+}$	16.62 *
	(3.51)	(3.79)	(6.58)	(3.53)	(3.81)	(8.38)
Experience >= 10 years * PARCC	-7.49 **	-6.52 *	-6.89	-5.25 $^{+}$	-4.37	-2.34
	(2.91)	(2.93)	(4.98)	(2.92)	(3.01)	(6.17)
HE in year *t* - 1 * PARCC	3.96	1.79	3.14	-1.30	-3.26	6.36
	(3.36)	(3.46)	(5.28)	(3.47)	(3.59)	(5.42)
D in year *t* - 1 * PARCC	0.30	-3.56	-6.50	-6.95 $^{+}$	-8.56 $^{+}$	-6.78
	(4.02)	(4.27)	(7.19)	(4.20)	(4.47)	(7.47)
ME in year *t* - 1 * PARCC	-5.14	-7.78	-11.16	-2.43	-4.04	-8.98
	(7.31)	(6.52)	(9.75)	(5.54)	(6.18)	(12.46)
Top quintile of IVA in year *t* - 1 PARCC	-6.10 $^{+}$	-4.71	-4.76	-3.56	-4.14	1.88
	(3.46)	(3.50)	(5.15)	(4.20)	(4.24)	(6.34)
Bottom quintile of IVA in year *t* - 1 * PARCC	7.63	8.63 $^{+}$	7.61	-5.43	-1.82	-4.49
	(5.11)	(4.93)	(9.03)	(4.31)	(4.39)	(9.27)
School FE		X	X		X	X
Teacher FE			X			X
n	1,111	1,111	1,111	1,168	1,168	1,168
R-squared	0.07	0.21	0.65	0.09	0.20	0.64

Note. All models include teacher (experience level, prior IMPACT rating, and quintile of lagged value-added scores in the subject) and class (proportion of students who are male, black, Hispanic, another non-white race, eligible for free or reduced-price lunch, with limited English proficiency, or in special education, as well as average lagged math and ELA test scores) controls. PARCC is an indicator for years in which the PARCC exam was administered (i.e., AY2014-15 and AY2015-16).

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; and $^{+}$ $p < 0.10$ before correcting for multiple hypothesis testing. Bonferroni corrections for multiple hypothesis testing show that none of the interacted teacher and classroom characteristics are significant at conventional levels.

TABLE 1.3b

Stability of Value-Added Percentile Ranks Across Exams, by Class Characteristics

	Math			ELA		
Student characteristics						
% Male * PARCC	-0.86	-1.09	-10.36	10.91	7.25	-0.90
	(15.13)	(15.28)	(22.70)	(14.41)	(15.56)	(26.76)
% Black * PARCC	-9.94	12.37	14.35	-10.12	7.67	36.16 +
	(14.18)	(13.94)	(18.06)	(12.63)	(13.12)	(20.87)
% Hispanic * PARCC	-8.35	9.12	3.60	-4.71	14.50	38.08
	(16.77)	(16.27)	(18.94)	(13.59)	(13.99)	(23.66)
% Other race * PARCC	7.22	17.04	19.27	-4.63	10.18	38.78
	(31.45)	(30.83)	(42.03)	(33.66)	(31.87)	(49.45)
% FRPL * PARCC	-2.87	-23.10 **	-20.83	-2.34	-14.83	-27.04 +
	(10.18)	(9.24)	(13.35)	(8.39)	(9.58)	(14.64)
% Limited English proficiency * PARCC	-5.47	7.07	13.85	8.75	4.50	-0.13
	(23.27)	(22.67)	(32.96)	(21.80)	(21.46)	(32.40)
% Special education * PARCC	39.46 +	31.07	-0.99	21.50	-3.21	27.41
	(21.12)	(19.40)	(33.11)	(18.59)	(18.04)	(29.39)
Mean lagged math score * PARCC	3.74	4.11	-1.38	0.99	0.81	-6.25
	(7.05)	(6.78)	(10.58)	(6.52)	(7.10)	(11.91)
Mean lagged ELA score * PARCC	-6.37	-13.05 +	-10.92	-0.23	-2.67	7.00
	(7.59)	(7.57)	(11.85)	(7.58)	(8.12)	(13.94)
School FE		X	X		X	X
Teacher FE			X			X
n	1,111	1,111	1,111	1,168	1,168	1,168
R-squared	0.07	0.21	0.65	0.09	0.20	0.64

Note. All models include teacher (experience, prior IMPACT rating, and quintile of lagged value-added scores) and class (proportion of students who are male, black, Hispanic, another non-white race, eligible for free or reduced-price lunch, with limited English proficiency, or in special education, as well as average lagged math and ELA test scores) controls. PARCC is an indicator for years in which the PARCC exam was administered (i.e., AY2014-15 and AY2015-16).

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; and $^{+}$ $p < 0.10$ before correcting for multiple hypothesis testing. Bonferroni corrections for multiple hypothesis testing show that none of the interacted teacher and classroom characteristics are significant at conventional levels.

TABLE 1.4

The Relative Association Between Teachers' Practice and Student Achievement Across Exams

	Math			ELA		
TLF Overall	0.003	0.034	0.039***	0.017^{+}	0.109***	0.000
	(0.013)	(0.060)	(0.014)	(0.009)	(0.041)	(0.010)
Factor 1: *Instruction*	-0.029*	0.025	-0.008	0.012	-0.004	0.009
	(0.013)	(0.045)	(0.014)	(0.010)	(0.048)	(0.011)
Factor 2: *Classroom Environment*	0.039**	0.023	0.064***	0.013	0.145***	-0.005
	(0.013)	(0.050)	(0.016)	(0.008)	(0.034)	(0.009)
Student and teacher controls	X	X	X	X	X	X
Teacher FE		X			X	
Student FE			X			X
n	15,765	15,765	15,765	19,723	19,723	19,723

Notes. Student controls include gender, race, eligibility for free or reduced-price lunch, limited English proficiency status, and special education status; teacher controls include race, gender, education, and experience. This table shows results from two regressions within each estimation model: the first rows show the interacted effects of overall TLF scores, standardized within year, and the PARCC exam on student achievement; the following rows show interacted effects between the PARCC exam and the *instruction* and *classroom environment* domains. PARCC exam scores used for this analysis are linked to the CAS scale and distribution using propensity-score matching followed by an equipercentile transformation (Approach 1 in appendix 1B); scores from each test are then standardized within subject and grade relative to the distribution of CAS scores in the final year of the CAS exam. TLF scores are standardized within year and use scores assigned only by external (i.e., master educators) evaluators. Data from AY2014-15 are omitted from this analysis.

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; and $^{+}$ $p < 0.10$ after Bonferroni adjustment for multiple hypothesis testing.

TABLE 1.5
IMPACT Score Components and Weights, AY2009-10 to AY2015-16

	CAS				PARCC
IMPACT Components	2009-10 to 2011-12		2012-13 to 2013-14		2014-15 to 2015-16
	Group 1	Group 2	Group 1	Group 2	Groups 1 & 2
Individual Value Added (IVA)	50%	0%	35%	0%	0%
Teaching and Learning Framework (TLF)	35%	75%	40%	75%	75%
Teacher-Assessed Student Achievement Data (TAS)	0%	10%	15%	15%	15%
Commitment to the School Community (CSC)	10%	10%	10%	10%	10%
School Value-Added	5%	5%	0%	0%	0%

Note. Group 1 consists only of reading and mathematics teachers in grades for which it is possible to define value added with the available assessment data. IMPACT scores can also be adjusted downwards for "Core Professionalism" (CP) violations reported by principals. Group 1 teachers did not have IVA calculated during the first two years of the PARCC exam (AY2015 & AY2016); in those years, group 1 teachers had the same score components and weights as group 2 teachers.

TABLE 1.6
Difference-in-Differences Covariate Balance

Female	0.011	
	(0.021)	
Black	0.027	
	(0.023)	
White	0.009	
	(0.021)	
Hispanic	-0.007	
	(0.010)	
Graduate Degree	0.021	
	(0.022)	
Experience: 0-1 years	-0.035	
	(0.019)	
Experience: 2-4 years	-0.040	*
	(0.020)	
Experience: 5-9 years	0.012	
	(0.020)	
Experience: 10-14 years	0.043	**
	(0.017)	
Experience: 15-19 years	0.009	
	(0.013)	
Experience: Missing	-0.001	
	(0.009)	

$^{***}\,p < 0.001$; $^{**}\,p < 0.01$; $^{*}\,p < 0.05$; and $^{+}\,p < 0.10$

TABLE 1.7

Difference-in-Differences Estimates of PARCC Effects on Teachers' Practice

Overall TLF	-0.147***	-0.163***	-0.166***	-0.054
	(0.042)	(0.041)	(0.038)	(0.039)
Factor 1: *Instruction*	-0.167***	-0.175***	-0.187***	-0.133***
	(0.038)	(0.037)	(0.036)	(0.042)
Factor 2: *Classroom Environment*	-0.025	-0.040	-0.032	0.078+
	(0.040)	(0.039)	(0.038)	(0.041)
Control for Level of Departmentalization	X	X	X	X
Teacher Controls		X	X	X
School FE			X	
Teacher FE				X
n	22,785	22,785	22,785	22,785

Note. The outcome variable is the TLF score assigned by master educators (MEs), standardized relative to the overall mean and standard deviation of ME-assigned TLF scores across the years of analysis (AY2010-AY2016). Teacher controls include education level, race, gender, and experience. Robust standard errors, clustered at the teacher level, are in parentheses.

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; and $^{+}p < 0.10$

FIGURE 1.1. *Timeline of DCPS Adoption of the Common Core State Standards and the PARCC Exam*

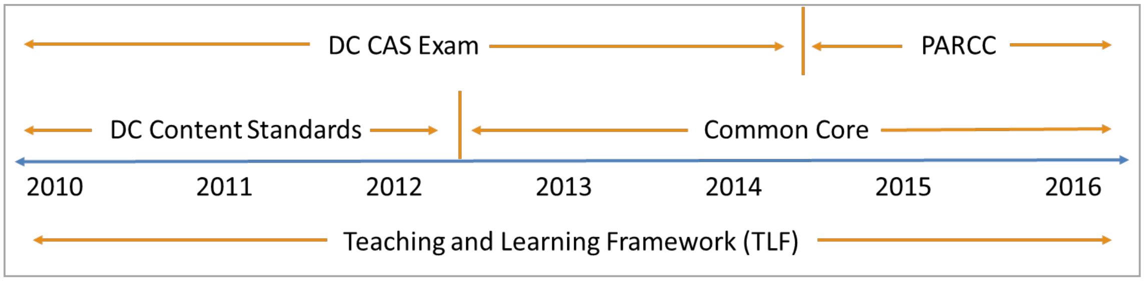
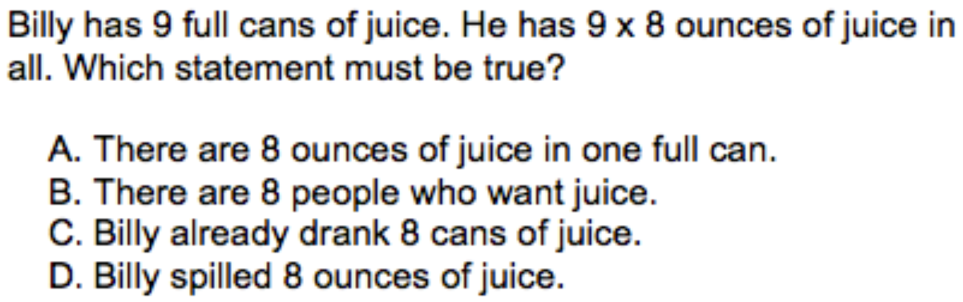


FIGURE 1.2. *Sample CCSS-Aligned Items, Grade 3 Math Standard 3.OA.A.1*

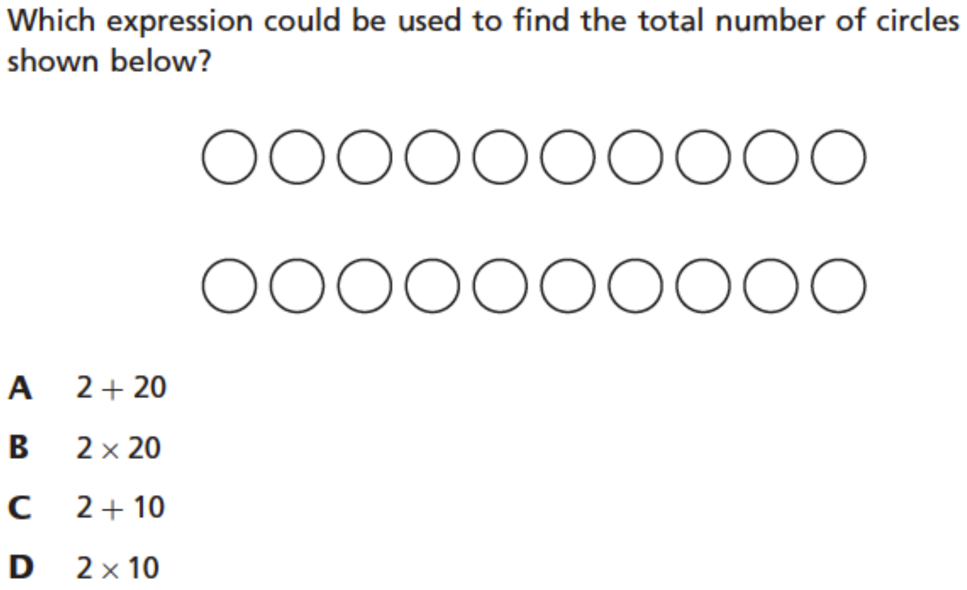
Example A:



Billy has 9 full cans of juice. He has 9 x 8 ounces of juice in all. Which statement must be true?

A. There are 8 ounces of juice in one full can.
B. There are 8 people who want juice.
C. Billy already drank 8 cans of juice.
D. Billy spilled 8 ounces of juice.

Example B:



Which expression could be used to find the total number of circles shown below?

A $2 + 20$

B 2×20

C $2 + 10$

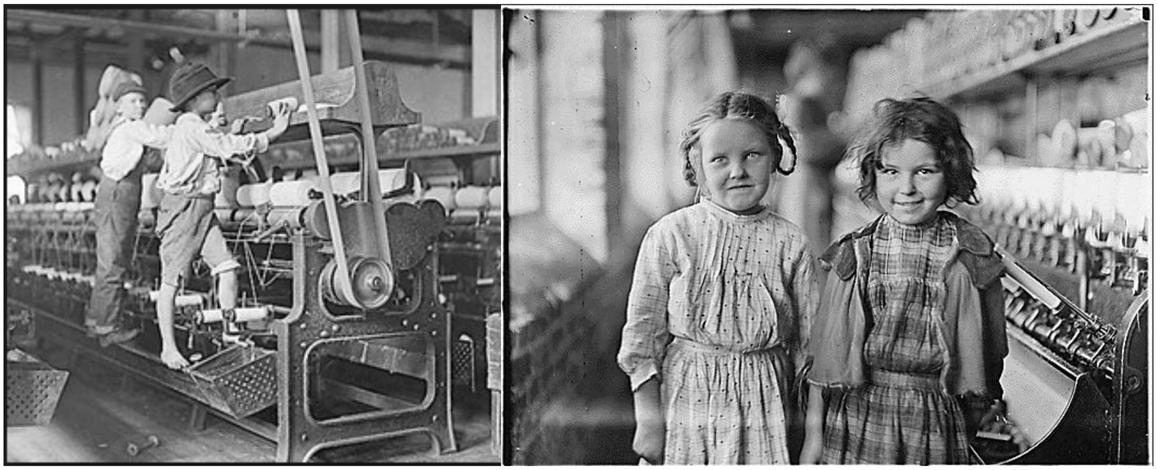
D 2×10

Note. Items were adapted from the SmarterBalanced item specifications (example A) and EngageNY.org released items (example B). Original source details are provided in the Student Achievement Partners (2017a) PowerPoint notes.

FIGURE 1.3. *Sample CCSS-Aligned Item, Grade 5 ELA Standard RL.5.7*

Stimulus 1:
Excerpt from *Counting On Grace* by Elizabeth Winthrop[a]

Stimuli 2 & 3:



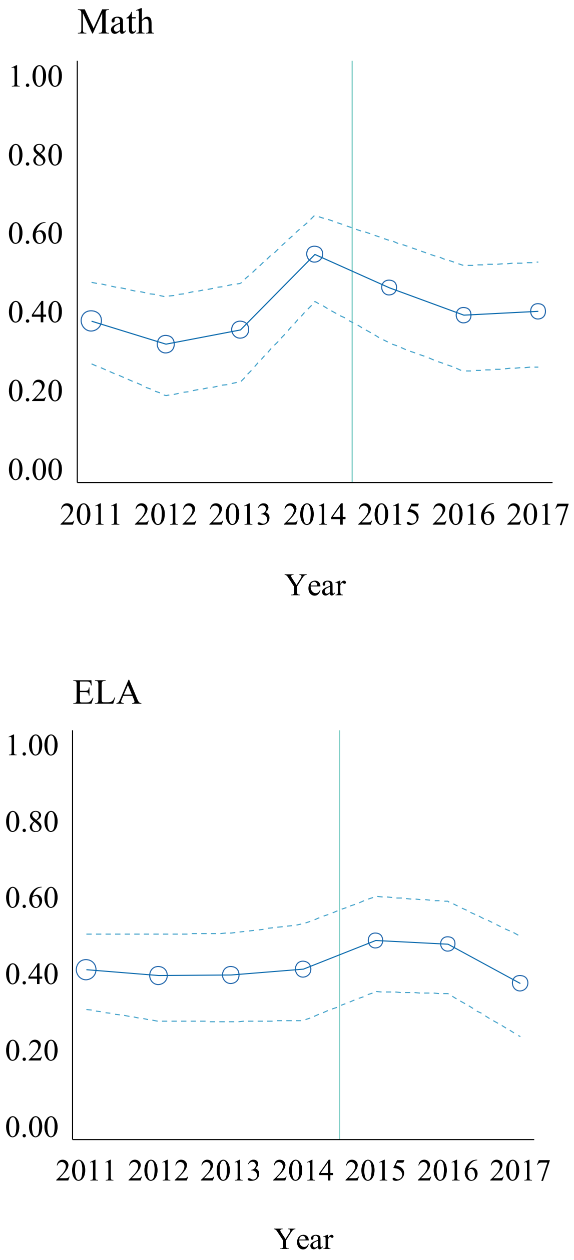
What two aspects of the story are further developed by the pictures of children in the cotton mill?

A. Children were working around machines that might be dangerous.
B. Teachers felt like they should check on their students as the children worked in the mill.
C. Children wrote letters that called for the investigation of child labor in the mills.
D. Very young children were hired to work in the mills.
E. Laws existed that protected children from being forced to work.
F. Children often chose to work instead of attend school.

[a] Excerpt text available at https://achievethecore.org/page/496/counting-on-grace-by-elizabeth-winthrop-mini-assessment

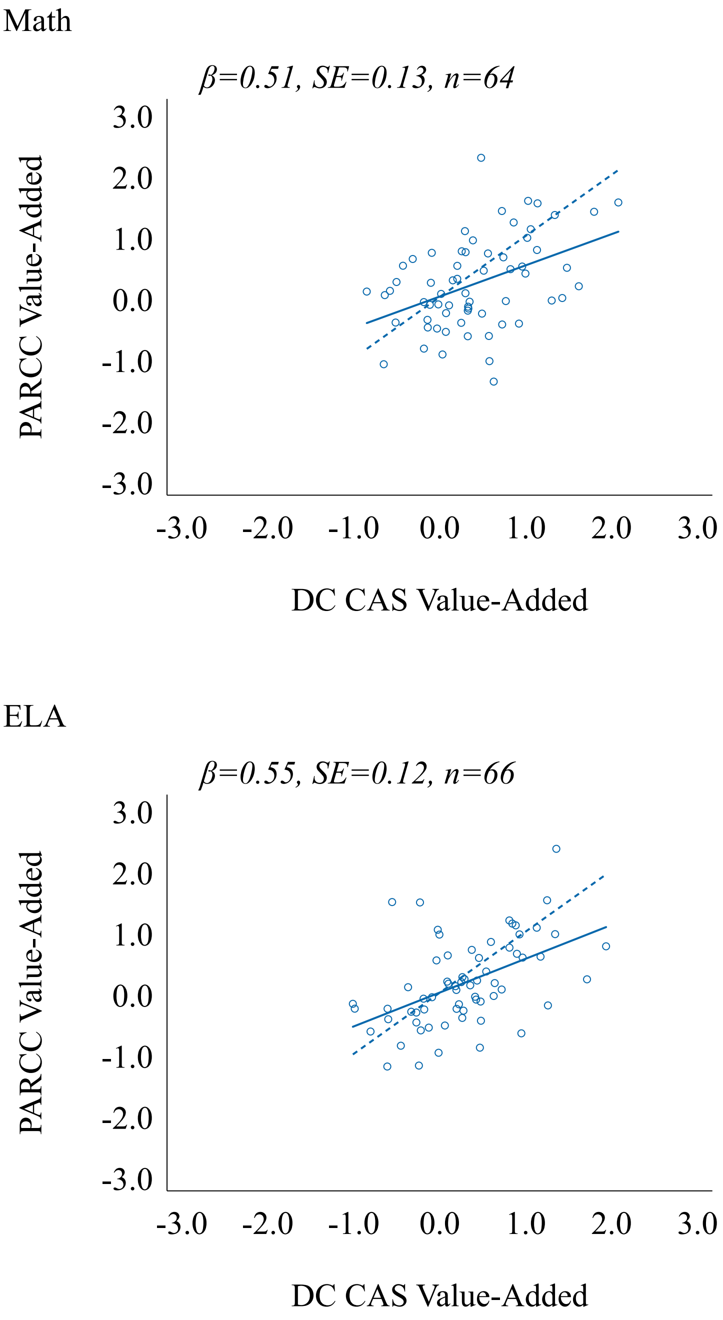
Source. Student Achievement Partners (2017b).

FIGURE 1.4. *Adjacent-Year Correlations of Value-Added Scores*



Notes. The solid blue line represents the correlation between teachers' current and prior-year value-added scores; dashed lines represent confidence intervals for each correlation coefficient. The vertical line identifies the year when DCPS transitioned to the PARCC exam.

FIGURE 1.5. *Relationship Between Teachers' Value-Added Scores on PARCC v. CAS*

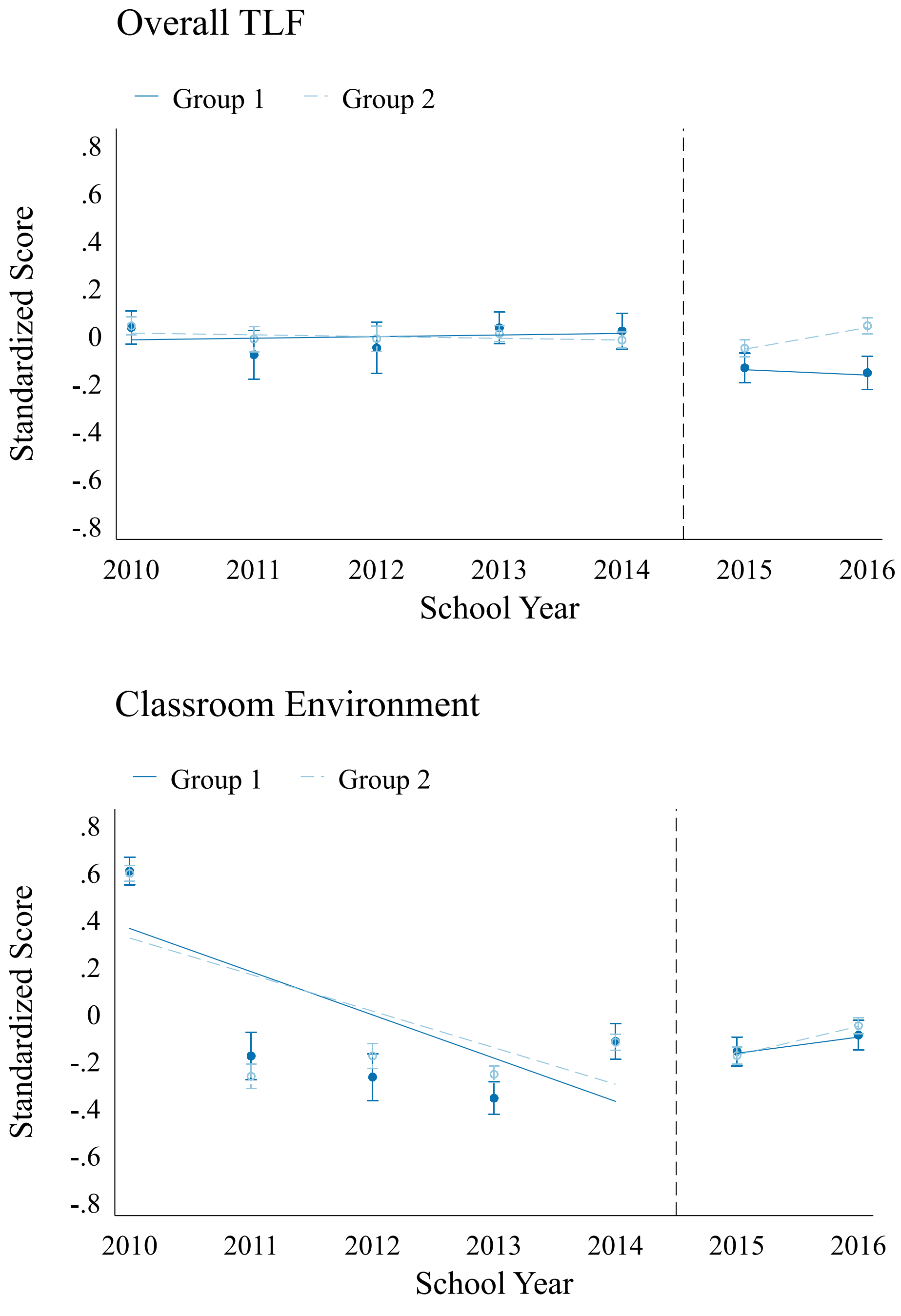


Math

ELA

Notes. The solid line represents the fitted regression line, while the dashed line represents a one-to-one relationship between value-added scores across the tests. Each plotted value-added estimate is the average of two or more years of value-added scores on the given assessment for an individual teacher in DCPS.

FIGURE 1.6. *Difference-in-Differences of Teachers' Practice Across the Transition to PARCC*


Overall TLF

Classroom Environment

Instruction

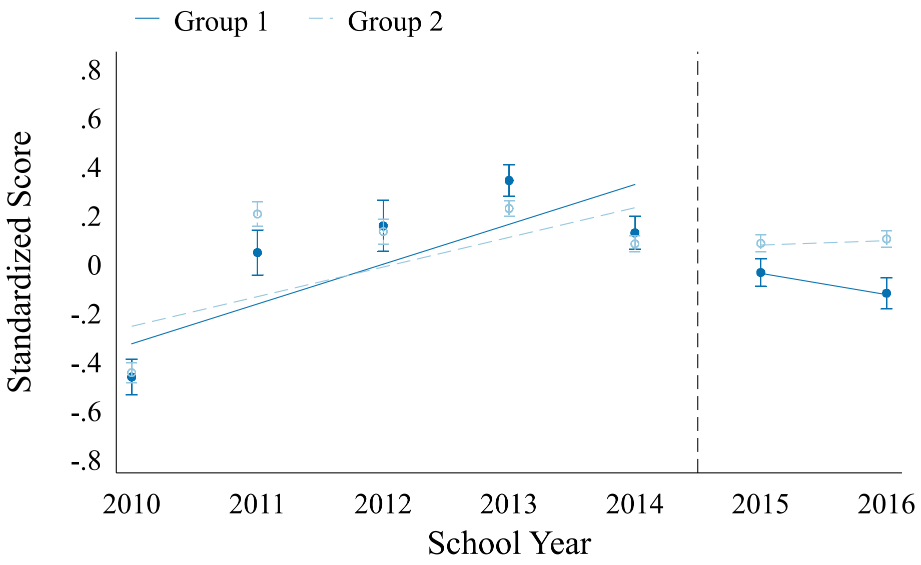
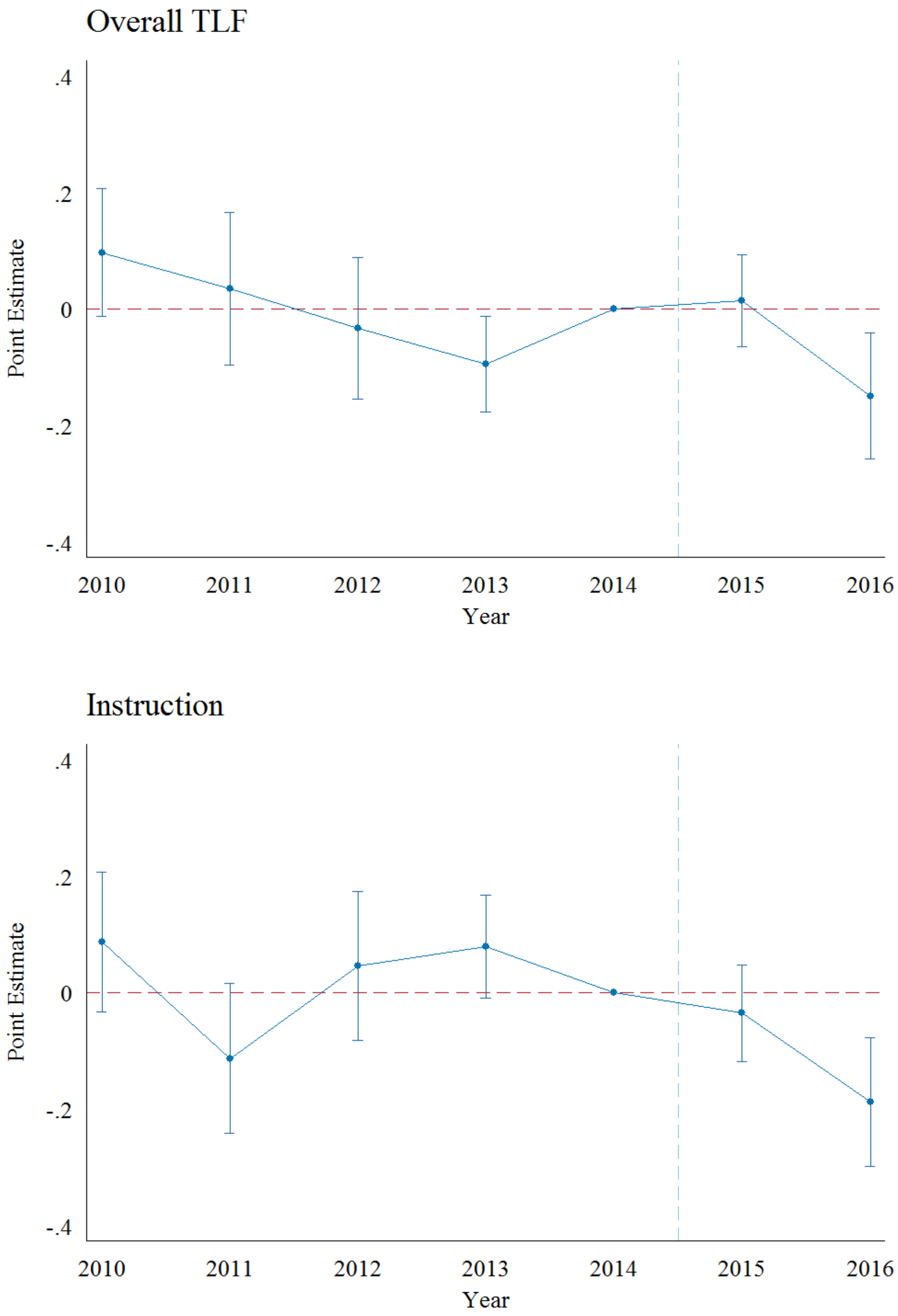
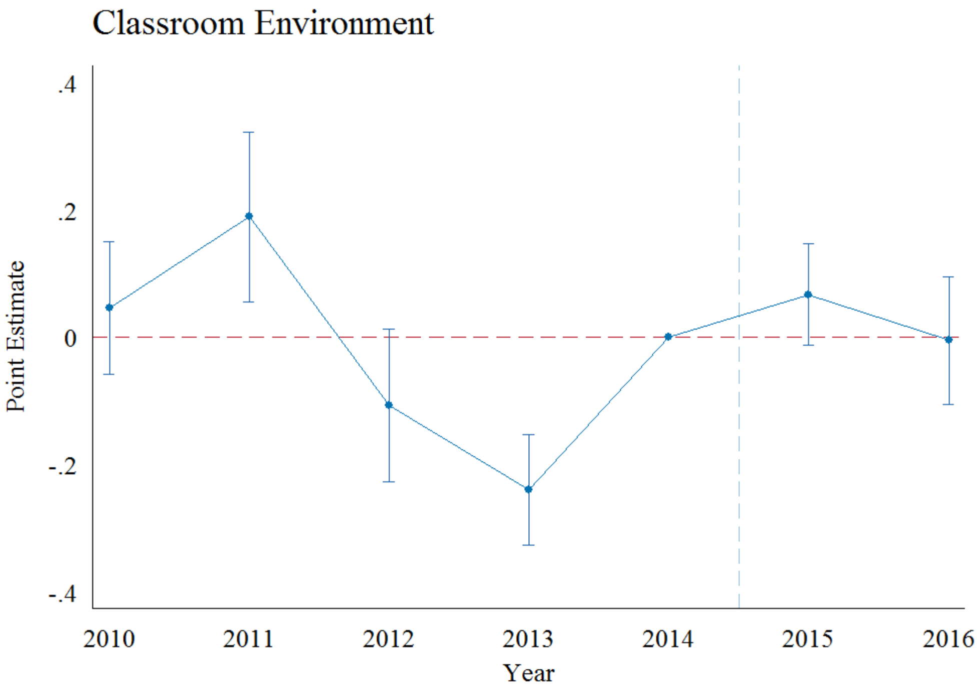


FIGURE 1.7. *Test for Common Trends in Teachers' Practice Across the Transition to PARCC*



Overall TLF

Instruction



Notes. Point estimates from a regression of TLF scores on interactions between group status and year of evaluation, with the last pre-treatment year (2014) serving as the reference point. Group 1 (treatment) consists of teachers in tested grades and subjects; group 2 (control) consists of all other general education teachers. Confidence intervals are at the 95% level.

CHAPTER 2

Is Effective Teacher Evaluation Sustainable?
Evidence from DCPS

(with Thomas Dee and James Wyckoff)

Abstract – Ten years ago, many policymakers viewed the reform of teacher evaluation as a highly promising mechanism to improve teacher effectiveness and student achievement. Recently, that enthusiasm has dimmed as the available evidence suggests the subsequent reforms had a mixed record of implementation and efficacy. Even in districts where there was evidence of efficacy, the early promise of teacher evaluation may not sustain as these systems mature and change. This study provides evidence on this question by examining the evolving design of IMPACT, the teacher-evaluation system in the District of Columbia Public Schools (DCPS). We describe the recent changes to IMPACT which include higher performance standards for lower-performing teachers and a reduced emphasis on value-added test scores. Descriptive evidence on the dynamics of teacher retention and performance under this redesigned system indicate that lower-performing teachers are particularly likely to either leave or improve. Corresponding causal evidence similarly indicates that imminent dismissal threats for persistently low-performing teachers increased both teacher attrition and the performance of returning teachers. These findings suggest teacher evaluation can provide a sustained mechanism for improving the quality of teaching.

INTRODUCTION

Ten years ago, many education reformers championed rigorous and consequential teacher evaluation as an intervention that would improve the effectiveness of the teacher workforce and, in turn, increase student outcomes. In particular, both the federal government and prominent philanthropies encouraged such reforms through a variety of high-profile initiatives (e.g., Race to the Top, Teacher Incentive Fund, the Measures of Effective Teaching Project, NCLB waivers and Intensive Partnerships for Effective Teaching). In response, most states and school districts designed and implemented new teacher evaluation systems (Steinberg & Donaldson, 2016).

As reports on the effects of these teacher-evaluation reforms have begun to accumulate, the corresponding public discussion has arguably become muddled. At a high level, states and school districts designed very similar systems. They all contained a teacher observation component and most included some form of student-achievement outcomes for which the teacher is responsible (Steinberg & Donaldson, 2016; Kraft & Gilmour, 2017; NCTQ, 2017). However, some evidence suggests that rigorous teacher evaluation improved teaching and student outcomes in Washington, DC (Adnot, Dee, Katz & Wyckoff, 2017; Dee & Wyckoff, 2015), Chicago (Steinberg & Sartain, 2015), and Cincinnati (Taylor & Tyler, 2012). Nonetheless, there is a growing public narrative that teacher evaluation reform has been a costly failure (Bill Gates and Melinda Gates 2018 Annual Letter; Strauss, 2015; Iasevoli, 2018) and waste of resources (Dynarski, 2016; NCTQ, 2017). For example, a recent RAND study (Stecher, Holtzman, Garet, Hamilton, Engberg, & Steiner, 2018) of three school districts and four charter management organizations found that teacher evaluation did not improve student achievement, but also suffered from "incomplete implementation."

The logistical and political challenges to implementing meaningful and informative teacher evaluation appear to be widespread. Kraft and Gilmour (2017) surveyed 24 states with teacher-evaluation reforms and found that, in most states, roughly 95 percent of teachers are still rated as effective or better. This finding is strikingly similar to those reported in *The Widget Effect*, a report from The New Teacher Project (TNTP) that precipitated much of the discussion regarding teacher evaluation reform (Weisberg, Sexton, Mulhern, & Keeling, 2009). Currently, we know relatively little about why the implementation of teacher-evaluation practices differs across contexts. And, more generally, we know relatively little about whether and under what circumstances teacher-evaluation reforms have produced systematic changes in teaching and learning.

Even if teacher-evaluation reforms produced meaningful *early* effects during the surge of enthusiasm and initial focus, the implementation literature offers ample cautions that such effects might not be maintained (Fixsen, Naoom, Blase, Friedman, & Wallace, 2005). Unless reforms altered school-level organizational cultures, effectively creating buy-in from principals and teachers, the forces that maintained the *status quo* pre-reform are likely to diminish the effects of these efforts. From this perspective, teacher evaluation is particularly vulnerable. The catalysts for teacher evaluation initiatives were typically "top-down" and the design and implementation of teacher evaluation was often hurried to meet federal grant-eligibility deadlines. Moreover, implementation often minimized or ignored the concerns of principals, teachers, and teacher unions (Chuong, 2014; McNeil, 2014). To become sustainable, the implementation literature suggests, such reforms would need to be implemented robustly and adapted over time to feedback and changing circumstances. Administrators need to provide continuing support and leadership, and teachers and principals must find teacher evaluations practical and useful (Fixsen et al., 2005).

It is against this backdrop that we provide new evidence on IMPACT, the controversial teacher evaluation system in the District of Columbia Public Schools (DCPS). Prior research has documented that aspects of IMPACT initially improved teacher performance (Dee & Wyckoff, 2015) and student achievement (Adnot et al., 2017). In this paper, we examine the evolving design features of IMPACT and the corresponding effects of its incentives on teacher attrition and performance under this mature and redesigned system. Notably, the design changes to IMPACT include a de-emphasis on evaluating teachers with conventional value-added test scores and an increase in the performance standards. The higher expectations for teacher performance include a new rating category (i.e., “Developing”) for lower-performing teachers who would have previously been considered “Effective.” Even in the absence of these design changes, the longer-term effects of IMPACT’s incentives are an open empirical question. For example, these reforms might be sustained if they remained well-implemented and if they catalyzed positive changes in school culture and performance. Alternatively, their effects might be attenuated in the context of a changed teacher workforce as well as in response to the presence of leadership turnover, shifts in organizational focus, and internal pressure to limit their most binding consequences.

We begin by describing the key design features and their evolution into the “IMPACT 3.0”, system which was in place beginning with the 2012-13 school year. We then examine descriptively the dynamics of teacher retention and performance under IMPACT during the period from 2012 to 2016. Overall, we find lower-performing teachers are substantially more likely to either leave DCPS or to improve their performance relative to higher-performing teachers. We also provide corresponding causal evidence on this relationship through a regression-discontinuity (RD) design that focuses on IMPACT's high-powered dismissal threat. Specifically, we examine the effects on teacher retention and performance of being rated as "Minimally Effective" (ME) instead of "Developing" (D). This treatment contrast effectively compares the credible and immediate dismissal threat for ME teachers who do not improve immediately to the incentives faced by D-rated teachers who instead have *two* years to achieve an E rating.[21] Consistent with the descriptive evidence, we find that facing an immediate, performance-based dismissal threat increased the voluntary attrition of lower-performing teachers. We also find qualified evidence that such threats increased the performance of teachers who returned. Our study concludes with a discussion of the implications of these findings for the ongoing research and policy agenda on teacher evaluation.

Incentives and Evaluation in Washington, DC

In 2007, following his election on a reformist agenda, Mayor Adrian Fenty secured approval for mayoral control of the District of Columbia Pubic Schools (DCPS). The low-income, largely-minority district suffered from chronically low academic achievement and persistently struggled to make meaningful improvements. For example,

[21] IMPACT's other rating thresholds imply additional opportunities to examine IMPACT's incentives in RD design. However, the changes to IMPACT, which we describe in detail below, made the incentive contrasts at other thresholds less stark.

DCPS's scores on the NAEP math tests in 2007 were lower than any other state or participating urban district in the country. The District was also among the lowest in reading performance (USDOE, 2007). Before long, the quality of DCPS's teaching force became a focal point for these reforms. Evidence of the importance of teachers for driving student outcomes (e.g., Gordon, Kane, & Staiger, 2006; Rivkin, Hanushek, & Kain, 2005; Rockoff, 2004) provided a motivation for this focus. Students in high-poverty schools are the least likely to have high-quality teachers, and poor schools attract less experienced teachers and have higher rates of teacher attrition (Clotfelter, Ladd, & Vigdor, 2005). Additionally, evidence suggests that the largest impacts of teacher quality occur for less-advantaged students, specifically African-American students and those whose performance is in the low and middle ranges of the achievement distribution (Aaronson, Barrow, & Sander, 2007).

It was in this context that, in 2009 under the direction of then-Chancellor Michelle Rhee, DCPS implemented IMPACT, a teacher performance-assessment system.[22] A fundamental intent of IMPACT was to reward high quality teaching, while removing low-performing teachers who failed to make adequate improvements. In the 2012-13 school year, DCPS changed several design features of IMPACT. Four features define much of IMPACT's structure: a) the components of the multi-measure evaluation system, b) the rating categories that distinguish teacher performance levels, c) the thresholds that determine rating categories, and d) the stakes associated with rating categories. Each of these has changed since IMPACT's inception to address feedback from teachers and evolving goals for improving student performance. Taken together, these changes became

[22] For a thorough and insightful discussion of the design and implementation of IMPACT, see Toch (2018).

known in the district as IMPACT 3.0.[23] We discuss each of these design features and their changes in turn.

Multi-measure Components. The components that make up teachers' IMPACT scores, and their weighting, depend on the grades and subjects taught. The majority of general-education teachers (i.e., 80 percent) teach in grades and subjects for which value-added scores based on standardized tests cannot be defined. For these "Group 2" teachers, 75 percent of overall IMPACT scores are based on a classroom observation measure, the Teaching and Learning Framework (TLF). TLF scores reflect average performance across 9 domains, measured as many as five times during the school year by a combination of in-school and external evaluators. When introduced, a teacher's overall IMPACT score included several other measures: the principal's assessment of their contributions to the professional life of the school (Commitment to School Community; CSC), student performance on a measure chosen by the teacher and approved by the principal (TAS), and school-level value-added measure (SVA). For teachers in tested grades and subjects ("Group 1"), the largest contributor to their IMPACT scores was based on student achievement, as measured by individual value-added scores (IVA). IVA was calculated employing a typical state achievement test, the DC-CAS until 2014-15, when the DCPS adopted the PARCC exam. For the first two years of the PARCC exam (2014-15 and 2015-16) IMPACT for Group 1 teachers did not include IVA over concerns that teachers needed time to adjust to the PARCC. Group 1 teachers were also evaluated based on TLF, CSC, and SVA, but not TAS. These components and their weights remained unchanged during the first three years (AY2009-10 to AY2011-12) of

[23] A more modest change occurred after IMPACT's first year when the teacher observation rubric we describe below switched from 13 to 9 domains. This change defined IMPACT 2.0.

IMPACT. Table 2.1 describes the weights for these components for both Group 1 and 2 teachers.

Under IMPACT 3.0, the weights applied to these components changed substantially. In particular, the emphasis put on test-based value-added measures fell. DCPS eliminated SVA entirely in response to teachers' concerns that they had virtually no control over their scores on this school-level measure. And, for Group 1 teachers, the weight applied to IVA fell from 50% to 35% in 2012-13 and then again to 0% in 2014-15. The stated intent of these changes was to reduce anxiety for Group 1 teachers, who expressed concern that such a large part of their IMPACT score was based on high-stakes value-added measures. DCPS correspondingly increased the weight applied to the more flexible TAS measure and, for Group 1 teachers, the TLF measure (table 2.1).

Teacher Performance Categories. During its first three years, teachers were assigned to one of four rating categories—Highly Effective (HE), Effective (E), Minimally Effective (ME), and Ineffective (I)—based on their overall IMPACT score, which ranged from 100 to 400. In AY2012-13, DCPS created a new performance category-- Developing (D)—by effectively dividing the Effective category in half with the lower portion becoming the Developing category. The motivations for this change included evidence that the prior Effective range reflected considerable variability in teacher performance as well as a desire to signal increased urgency to improve teaching skills and student outcomes. Teachers whose teaching performances were previously viewed as acceptable, were now under pressure to improve with stakes we describe below.

Thresholds for Performance Ratings. As noted above, the system evolved from four performance categories to five. Initial and revised thresholds are shown in table 2.2. Initially, teachers scoring 175 points or less were rated I; those scoring 175 through 249 were rated ME; teachers with scores from 250 through 349 were rated E; and teachers with scores from 350 through 400 were rated HE. In 2012-13, the E category was divided in half so that teachers whose IMPACT score was between 250 and 299 were rated D and those scoring 300 to 349 were rated E. In addition, the I category was expanded from 174 to 199, thus capturing some teachers previously labeled ME. The intent of the increased performance standards embedded in these threshold changes was to encourage teachers to strengthen their teaching skills.

Performance Stakes. Teachers identified as I by IMPACT have always faced dismissal at the end of the school year in which the rating was earned, as have teachers who scored twice consecutively as ME (table 2.2). Similarly, teachers rated HE received substantial one-time bonus payments, with amounts varying by the subject and grade level taught and the proportion of students in the teachers' schools receiving free and reduced-price lunch. In addition, before IMPACT 3.0, teachers who attained a HE rating for two consecutive years were eligible to receive a considerable base pay increase. The bonus and base-pay increases varied depending on whether teachers were teaching a subject with value-added scores, were teaching in high-poverty schools and/or were teaching a high-need subject (table 2.2).

Beginning in AY 2012-13, IMPACT 3.0 modified the stakes associated with different rating categories. As before, teachers would be dismissed with one I or two consecutive ME ratings. However, with the introduction of the D category, teachers would be separated with three consecutive D ratings (or one D and a subsequent ME or I rating). DCPS also introduced a performance-based career ladder for teachers: the Leadership Initiative for Teachers (LIFT). LIFT was intended to provide teachers with additional recognition and professional opportunities.[24] Importantly, LIFT also became the mechanism by which teacher performance base-pay increases were determined. These base-pay increases became a function of the level and persistence of performance measured by IMPACT. The incentives for HE teachers also differ somewhat from those offered under the prior design of IMPACT (table 2.2). DCPS altered bonuses to create stronger incentives to teach in the 40 most demanding schools in DCPS and substantially reduced incentives for teachers in low-poverty schools (i.e., those with less than 60 percent free and reduced-price lunch). Beginning in AY2012-13, base pay increases are only available to teachers in high-poverty schools and are based on a career ladder that is a function of IMPACT ratings.[25] These changes in stakes instantiated a focus on attracting and retaining HE teachers in high-poverty schools.

These design changes and the ongoing evolution of DCPS teachers coincided with changes in the distribution of teacher effectiveness, as shown by the two graphs in figure 2.1. As is evident, the measured performance of teachers has meaningfully increased over time. For example, between 2010 and 2016, the median IMPACT score increased from 303 to 332 (i.e., a gain equivalent to 0.58 SD). Before we examine teacher retention and

[24] The opportunities associated with advancing through LIFT stages include developing curricular materials, mentoring colleagues, and being eligible for certain fellowship opportunities. More information about the LIFT program is available on the DCPS website at https://dcps.dc.gov/page/leadership-initiative-teachers-lift.

[25] More information on this career ladder can be found at https://dcps.dc.gov/page/leadership-initiative-teachers-lift.

performance under IMPACT 3.0, we address concerns recently raised about the manipulation of measured student outcomes in DCPS.

In general, the intended goals of accountability reforms in education are to provide teachers and school leaders with actionable information that can guide their improvement as well as with incentives that encourage those changes. IMPACT seeks to improve the effectiveness of the teaching workforce through the attrition of teachers with unacceptably poor performance and has adopted dismissal policies that toward that end. A notable concern with output-based reforms is that they may also cause some individuals to engage in unintended, counterproductive (and, in some cases, illegal) activities. For example, DCPS has recently come under scrutiny for inappropriately graduating students who had not met graduation requirements, in an effort to improve graduation rates, a widely-cited measure of educational success, and one that can play a small role in DCPS principal evaluations. School leaders were also caught manipulating—or pressuring their teachers to manipulate—student attendance and course credit data to meet school-level performance targets (Balingit & Tran, 2018; McGee, 2018; Brown, Strauss, & Stein, 2018).

These allegations, while notable and troubling, are not directly salient for IMPACT. Graduation rates, attendance rates, and credit accumulation are not a component of teachers' IMPACT scores. Instead, IMPACT heavily weights classroom observations to induce teachers to improve diverse pedagogical skills and behaviors. In theory, the emphasis on TLF could encourage manipulation by principals who want to support teachers' ratings. However, the presence of additional TLF ratings by external evaluators and the corresponding system of principal accountability suggest that such manipulation is unlikely.[26] We are aware of no assertions of manipulation to improve teacher IMPACT ratings.[27] Though manipulation of IMPACT scores seems unlikely, we explicitly examine the density of observations near the relevant thresholds as part of our analysis and find no evidence of such manipulation.

In sum, IMPACT 3.0 signals the intent by DCPS to make additional improvements in student academic performance by increasing the performance of teachers. Under IMPACT 2.0, about 70 percent of teachers earned an Effective rating and this performance range was quite broad. Creating the Developing category by dividing the Effective range in half and broadening the range for Ineffective teachers sent a strong signal that DCPS believed they could meaningfully improve teacher effectiveness. DCPS also signaled an intent to increasingly focus on its lowest-performing schools. Financial incentives for high performing teachers were dramatically reduced in low-poverty schools, where base-pay incentives were eliminated and bonuses for high performance cut by 75 percent. IMPACT 3.0 also included important elements to address concerns raised by teachers. The weight applied to IVA was reduced from 50 to 35 percent for applicable teachers. The elimination of SVA also addressed a long-standing concern by teachers that this measure was beyond their direct control. The career ladder, LIFT, added formal recognition and rewards to teachers as they realized professional-development milestones.

[26] The variability in principals' TLF ratings is also inconsistent with widespread manipulation. Dee and Wyckoff (2015) also find that IMPACT incentives generated similar increases in the TLF ratings by principals and external evaluators.

[27] Allegations of cheating on the high-stakes test in DCPS received extensive coverage in the press prior to 2012-13; we are unaware of any allegations since. Dee and Wyckoff (2015) address the allegations of cheating for this earlier period and find cheating was very limited and had no effect on their estimates of the effect of IMPACT.

LITERATURE REVIEW

The conceptual foundations for teacher-evaluation policies focus on two broad mechanisms. One mechanism involves how incentives may shape the development and performance of extant teachers in ways that are beneficial to students. For example, programs that provide teachers with clear and actionable feedback on the character of their classroom performance can provide targeted support to their professional development. The presence of sanctions or rewards based on their performance can also encourage teachers both to increase their effort and to reallocate their instructional focus toward effective practices.

The empirical literature examining the effects of performance assessment and incentives on teacher performance is mixed. In particular, several small-scale and experimental attempts to use financial incentives to improve teachers' performance find limited or null effects (Fryer, 2013; Marsh, Springer, McCaffrey, et al., 2011; Springer et al., 2010; Springer et al., 2012.[28] However, there are some studies in which teachers have responded to such incentives with improved performance (e.g., Balch & Springer, 2015; Chiang et al., 2017). Furthermore, some studies (e.g., Taylor and Tyler 2012, Steinberg and Sartain 2015) provide evidence that evaluations do not necessarily need to be linked to rewards or sanctions enhance teachers' practice. A potentially important unintended consequence is that high-stakes evaluations might encourage unintended behaviors such as cheating, particularly when a single performance outcome is emphasized (Apperson, Bueno, & Sass, 2016). While such responses have been observed where stakes are tied to

[28] The incentives examined in these studies may be weak for a variety of reasons: low dollar amounts, group rather individual incentives, a focus on cash for test scores rather than more direct measures of teacher performance, and the expectation that the incentives are temporary rather than an enduring policy change.

school- or student-level performance (e.g., Jacob & Levitt, 2003; Dee, Dobbie, Jacob, & Rockoff, forthcoming), we do not know of such evidence in the context of teacher-level accountability systems.

The second mechanism that motivates teacher evaluation reforms concerns the composition of the teacher workforce. That is, another key motivation for these reforms is the expectation that they will the recruitment and retention of high-performing teachers while also encouraging the attrition of low-performing teachers (Goldhaber, 2015). Existing empirical studies generally suggest that incentives do result in improved retention. While the evidence linking incentives on retention is by no means universally positive, incentive policies have generally been associated with improved retention. Fulbeck (2013), for example, found that Denver Public School district’s ProComp program, which awards additional financial compensation for a variety of performance criteria, extra credentials, and teaching in high-poverty schools, is associated with significantly improved teacher retention within a school, though these retention effects are substantially smaller for high-poverty schools. North Carolina had similar success with a briefly-implemented program that awarded bonuses to teachers of high-need subjects who taught in low-income and low-performing schools (Clotfelter, Glennie, Ladd, & Vigdor, 2008). Chicago’s Teacher Advancement Program, which awarded bonuses according to value-added and classroom observation scores as well as to teachers who took on leadership and mentorship roles within their schools, was also associated with improved school-level retention (Glazerman & Seifullah, 2012). In Tennessee, teachers in low-performing schools who earned performance bonuses were more likely to be retained than their peers who scored just below the threshold of bonus eligibility, but this effect was concentrated only among teachers in tested grades and subjects (Springer, Swain, & Rodriguez, 2016).

Incentives and evaluation can also influence the quality of the teaching composition by encouraging higher-performing teachers to enter the profession. Such effects are less well documented in the literature, but simulations of incentive-based evaluation on entry into the teacher labor market (Rothstein, 2015) suggest that performance-based contracts can alter the performance distribution of the teaching workforce by enticing higher-ability teachers while dis-incentivizing the entry or retention of lower-ability teachers. These effects, however, may be extremely small, given that those who are new to teaching generally have little confirmation of their performance ability from which to assess their probability of earning incentives. The most compelling evidence of selection-into-teaching effects comes from California, which briefly offered the $20,000 Governor's Teaching Fellowship to the most-competitive students from accredited post-baccalaureate teacher licensure programs in return for teaching in low-performing schools. Steele, Murnane, and Willett (2010) found that these novice teachers were significantly more likely to begin their teaching careers in low-performing schools than they would have in the absence of the Fellowship program.

Evidence from DCPS. This prior literature provides an important context for understanding the mechanisms through which IMPACT might improve DCPS's teaching quality (i.e., performance, recruitment, retention, and attrition). Recent empirical studies based on the earliest years of IMPACT suggest that the District's reforms had positive impact on most of these fronts.[29] For example, there is evidence that IMPACT influenced

[29] The one exception is teacher recruitment and selection into DCPS. We know little about the causal effects of IMPACT because the policy went to scale simultaneously. However, Jacob, Rockoff, Taylor,

the composition of the DCPS teaching workforce in a manner that improved teacher effectiveness and student achievement. Using a regression discontinuity design, Dee and Wyckoff (2015) found that a dismissal threat for low-performing teachers led to a 50-percent increase in the attrition of those teachers, indicating that the program successfully induces the voluntary departure of its weaker teachers. Such teacher turnover could actually harm student learning through the disruption of teacher teams and through hiring less qualified teachers. However, Adnot et al. (2017) find that performance-based dismissals and attrition in DCPS led to replacements who were substantially more effective at raising student achievement. These achievement effects were particularly strong for students in high-poverty schools.

The early effects of IMPACT were not purely compositional, however. Dee and Wyckoff (2015) also examined the effect of strong incentive contrasts at consequential performance thresholds on retained teachers' next-year performance. They found positive performance effects for high-performing teachers facing potentially large financial rewards, as well as for low-performing teachers who faced potential dismissal but remained teaching in DCPS. Among those who returned teaching the next year, both ME and HE teachers improved by approximately 25 percent of a standard deviation of IMPACT points. Adnot (2016) built on these findings to explore the effects of these sharp incentive thresholds on teachers' practice, as measured by specific constructs in the TLF. She found that dismissal threats improved the observed classroom performance of

Lindy & Rosen, (2016) examine the screening of DCPS teacher applicants under IMPACT. Their description indicates that, under IMPACT, DCPS has a larger number of teacher applicants and a more multi-faceted screening process than exists in most districts.

teachers, particularly with regard to the teaching standards that were concrete and specific.

A more recent study by Katz and Wiseman (2018) examined the effects of the incentive changes that occurred with IMPACT 3.0 on teachers' retention and mobility in hard-to-staff schools. IMPACT explicitly attempts to attract and retain high-skill teachers in DCPS by offering considerably larger financial benefits to high-performing educators who teach in schools and subjects where recruitment and retention are most difficult. IMPACT 3.0 emphasized the incentive to teach in the district's lowest-performing and highest-poverty[30] schools by making the base-pay incentive available only to HE teachers in these schools, where previously this incentive was available to all teachers earning an HE rating. Using multiple difference-in-difference specifications, Katz and Wiseman find improved retention in high-poverty schools under IMPACT 3.0, as well as a slight increase in transfers from low- to high-poverty schools in the district. These effects, however, did not occur for HE teachers in the lowest-performing schools.

In summary, the high-fidelity implementation and sustained impact of large-scale educational reforms have proven difficult to achieve (Fixsen et al., 2005; Chiang et al., 2017; Stecher et al., 2018). Indeed, as described above, the evidence from rigorous assessments of teacher evaluation is mixed, raising important questions regarding the sustainability of this reform even in the contexts where it met with initial success. We turn to an examination of whether IMPACT was able to sustain its initial substantial

[30] Seventy-five percent of DCPS schools are identified as high-poverty as they have at least 60 percent of their students eligible for free or reduced price lunch. Of these schools, the district identified 40 schools who were the lowest performing. As shown in Table 2.2, highly effective teachers in the lowest-performing schools received an additional $10,000 bonus in addition to the bonus they were eligible to receive for teaching in a high-poverty school.

improvements in teacher effectiveness and student achievement both as the program matured and as its design evolved in important ways.

DATA AND SAMPLE

We base our analysis on a panel of teacher-level administrative data spanning from the start of IMPACT in AY2009-10 through AY2015-16. These data include, for all teachers in DCPS, information on teachers' IMPACT scores, ratings, and consequences, as well as demographic characteristics (e.g., race and gender), background (i.e., education and experience), and information about the schools in which they work and the students they teach (table 2.3). The IMPACT data include initial scores, as well as final scores that reflect the very small number of cases where scores were revised or successfully appealed. We use these data to create our two outcome variables: retention and next-year IMPACT score.

Our analysis focuses on what is arguably IMPACT's most potent incentive: the risk of dismissal for teachers who received a ME rating in the preceding year. We don't explore incentives at other thresholds for several reasons. Because treatment at the E/HE threshold is variable and relies upon different criteria over time, and because the sample sizes are quite small across many of these treatment conditions, we do not explore treatment effects for high-performing teachers incentivized by bonus pay or salary increases.[31] The criteria for treatment at the D/E cut-off in IMPACT 3.0 is complicated by several factors. First, there are several potential treatment groups depending on teachers' sequence of ratings over time. For example, teachers who are rated ME followed the next year by a D rating have one additional year to be rated higher than D.

[31] Katz & Wiseman (2018) explore the effects of the AY2012-13 changes to IMPACT's financial incentives for HE teachers employing a difference-in-differences approach.

Teachers who are rated D following an E rating (or are new to the system) have two more years to score above D. Second, for teachers who are rated D, the salience of the dismissal threat varies based on the timing of the D rating since most teachers have two years to improve. Finally, data currently available to us do not allow us to examine the ultimate disposition of teachers assigned a D in 2012-13 without conflating IMPACT effects with the effects of an intensive professional development program introduced by DCPS in AY2015-16.

The full sample consists of 17,465 teacher-by-year observations who received IMPACT ratings between AY2010-11 and AY2014-15, with approximately 3,500 teacher ratings per year. Of these observations, 13,192 (76%) are general education teachers—roughly 2,600 teachers per year. To create our analytic dataset, we construct samples which include general education teachers whose rating in year t places them on either side of the ME/E cut-off in IMPACT 2.0 (AY2010-11 to AY2011-12) and the ME/D cut-off in IMPACT 3.0 (AY2012-13 to 2014-15). In both cases, teachers who are rated ME face involuntary separation if they receive a second consecutive ME rating. This reduces our analytic sample to 4,300 teachers in IMPACT 2.0 and 1,980 teachers in IMPACT 3.0. We omit teachers from IMPACT 1.0 from our analysis because of anecdotal evidence that teachers initially did not expect IMPACT to persist beyond its first year, which is further supported by null results in Dee and Wyckoff's analysis of IMPACT's initial years.

Teachers are assigned to the ME treatment group if their score (pre-appeals) placed them in the ME score range. Under IMPACT 2.0, ME scores ranged from 175 through 249, and under IMPACT 3.0 ME scores ranged from 200 through 249. Teachers who have scored their first ME rating must improve by the following year if they wish to retain their teaching positions. The teachers scoring at the next highest rating level do not face this threat. Before the 2012-13 changes, this was teachers earning an E rating (scoring between 250 and 349); following program revisions, this group consisted of teachers earning a Developing rating (those scoring between 250 and 299).

Any teachers not assigned to the ME treatment and the rating category just above it are removed from the analytic sample. To avoid conflation of voluntary and involuntary separation outcomes, the treatment sample is then restricted to teachers who did not have an ME or D rating in the prior year—ratings which result in involuntary dismissal when immediately followed by an ME rating. The final analytic sample consists of 3,888 teachers in IMPACT 2.0, 528 (14%) of whom are rated ME, and 1,809 teachers in IMPACT 3.0, of whom 370 (20%) are rated ME.

METHODS

We first explore patterns in teachers' performance and retention descriptively by following teachers' retention decisions under IMPACT 3.0. We then turn to examining the effects of IMPACT's dismissal threat on teacher retention and performance. Specifically, we rely on a regression discontinuity (RD) design to estimate the effects of an ME rating. This approach effectively exploits the plausibly random variation in teachers' initial IMPACT ratings around the ME threshold to estimate local treatment effects. Our specifications take the following general form:

$$Y_{it} = \beta_0 + \delta(D_{it}) + f(S_{it}) + X_{it}\lambda + \tau_t + \varepsilon_{it}$$

For each threshold, Y_{it}, represents teacher i's retention or performance following year t (as measured by next-year IMPACT scores); δ represents the effect of the teachers' IMPACT rating (D_{it})—specifically, the effect of falling on the consequential side of the relevant cut point (i.e., scoring ≤249 for the ME/D threshold); $f(S_{it})$ is a flexible function of the assignment variable (i.e., the initial IMPACT score centered on the ME threshold); X_{it} is a vector of teacher covariates; τ_t represents year fixed effects to account for differences in the relationship between IMPACT assignment and baseline characteristics across years; and ε_{it} is an individual- and year-specific error term. In addition, we also explore models of the RD that include school fixed effects.

We employ several methods to test the internal validity of our estimates following best practice for RD analyses (Cattaneo, Idrobo, & Titiunik, 2018a, 2018b; Lee & Lemieux, 2009; WWC, 2017), including tests for robustness of results to assumptions about the functional form of the relationship between teachers' IMPACT scores and their retention or future performance. More specifically, our baseline specification controls for linear splines of the assignment variable above and below the ME threshold. However, we explore local linear regressions (LLR) that use increasingly small bandwidths of scores around the consequential cut point. We also examine specifications that include higher-order polynomials of the assignment variable and that apply triangular kernel weights to regressions such that greater weight is placed on scores closer to the threshold than those further away. These are discussed in our results section and presented in the appendices to this paper.

In addition to functional form, a key assumption for RD analysis is the exogeneity of treatment. We test for non-random assignment to treatment empirically, by estimating our regression specification with teachers' pre-treatment characteristics on the left-hand side in lieu of retention and performance outcomes. If treatment at the threshold is randomly determined, we should find no significant effects on δ for any of these teacher covariates. Results from these regressions are presented in appendix table A.2.1 and indicate no significant sorting of teachers to the treatment or control condition by observable characteristics at conventional significance levels. The probability of being assigned to treatment for teachers with five through nine years of experience is significant at $\alpha = 0.10$. We observe no additional indication of potential covariate imbalance. Regardless, we condition on these observable characteristics to limit potential endogeneity. Systematic score manipulation is quite unlikely in this context. This would be a concern, for example, if certain types of teachers were able to improve their initial scores to avoid assignment to the treatment, potentially confounding our treatment estimates. There are several reasons we believe this is not a concern in the case of IMPACT.

First, while it is conceivable that observation (TLF) scores could be manipulated if a school administrator were concerned about a teacher who faced separation based on prior-year IMPACT scores, giving that teacher a more generous TLF score as a result, this would be difficult to do in practice. While TLF scores are comprised in part of ratings from administrators—who might manipulate scores given their contextual knowledge of teachers' performance and personalities—external Master Educators also rate teachers and would not be privy to information about a given teacher's prior performance. We explicitly test for this by comparing treatment estimates from our regression models (not shown) where the outcome is the principals' TLF score to models where the outcome is the TLF score assigned by Master Educators; the difference in treatment estimates by type of rater is statistically indistinguishable from zero. In addition, observations are only partial contributors to IMPACT scores, so principals' potential for influence is limited. What's more, school administrators do not have specific stakes associated with their teachers' performance. While school leaders are evaluated under IMPACT as well, the only teacher-relevant contributing factors to principal IMPACT scores are overall student achievement outcomes and a set of observed principal behaviors that are thought to contribute to teachers' support and development; teachers' IMPACT ratings are not considered in their principals' IMPACT ratings.

Second, we employ teachers' initial IMPACT scores, rather than the scores they may have received post-appeal. Doing so substantially mitigates against score manipulation and avoids violation of the exogeneity assumption. As shown in figure 2.3, there is a slight (fuzzy) discontinuity in the probability of assignment to treatment given a teacher's initial IMPACT score in AY 2012-13. When final, post-appeal IMPACT scores are used, there could be some manipulation occurring around the cut points, though potential effects of this manipulation are small, given that few teachers' IMPACT ratings are successfully appealed. In the 2012-13 through 2014-15 academic years, only 56 of the initial IMPACT ratings for Group 1 and Group 2 teachers across all of the ratings thresholds were changed following revisions or appeals, representing less than one percent of all ratings across the three years. Most of these appeals (82%) were granted in the first year of IMPACT 3.0, while the number of successful appeals granted in AY 2013-14 and AY 2014-15 declined respectively to 1 and 9. The use of initial, pre-appeal scores could diminish the external validity of findings; however, given that so few teachers succeed in their attempts at revising initial scores, any differences in findings would likely be negligible had there been no score revisions (or had the analysis been of treatment-on-treated, rather than intent-to-treat, effects). In addition, fuzziness effects are largely isolated to the 2012-13 academic year, following an error in the calculation of teachers' IVA scores.

Density tests of the distribution of observations through the ME threshold provide direct empirical evidence that manipulation of the assignment variable did not occur (McCrary, 2008). Specifically, we use the local-polynomial density estimators proposed by Cattaneo, Jansson, and Ma (2017, 2018) to test for discontinuity in the density of observations around the ME/D threshold. This test relies on the assumption that if there were no systematic manipulation of scores around the threshold we would observe continuous changes in the density of observations at the cut-off; conversely, evidence of discontinuous density at the threshold would suggest possible non-random sorting of teachers to ME or D ratings. We run this falsification test for each year of IMPACT 3.0 individually and for all three years in aggregate, finding no statistical difference in densities across the threshold within or across years. This evidence, presented graphically in appendix figure A.2.1, further supports our assumption that treatment is exogenous at the ME/D threshold.

Third, for an RD to be internally valid, an additional requirement is that the average outcome (in this case, either retention or next-year IMPACT scores) is a continuous function of teachers' current-year IMPACT scores, conditional on their IMPACT rating. Concerns about the violation of this assumption would be raised if the relationship between the two outcomes and teachers' IMPACT scores indicated discontinuities at points other than the consequential threshold. If there were no treatment effect, we would expect the relationship between initial IMPACT scores and retention or next-year performance to continue as is, without additional discontinuities beyond the consequential cut points. The graphs in figure 2.3 suggest that this assumption is not violated at the ME/D threshold, though because this relationship is noisy it is difficult to assess purely though visual evidence. To further test that this assumption is met, we run a series of RD models, using “placebo” cut points. Assuming that there is a discontinuity, or treatment effect, at the consequential threshold, there should be no other detectable effects at thresholds where we would not expect to see them. These placebo tests (shown in appendix table A.2.2) produce no significant results at any point other than the cut-off between ME and D ratings.

Another potential threat to the validity of our estimates is the possibility of differential attrition from the sample across the threshold of analysis (WWC, 2017). There are, however, two key reasons why attrition is not a concern in this context for teachers’ retention. First, we assess intent-to-treat effects based on initial IMPACT score assignment, thereby defining treatment as the threat of dismissal associated with having initially scored at the ME level; treatment cannot be defined separately from the running variable, and attrition from the sample is in this context the outcome of interest. Second, we use the full set of administrative data from DCPS during this period, such that no teacher is omitted from the analysis, regardless of treatment status, and we are therefore able to define retention status for all teachers in the sample, and on both sides of the consequential threshold.

There is risk of differential attrition, however, when examining effects on next-year IMPACT scores. For example, while our administrative data allow us to follow teachers’ retention decisions, there are cases in which a teacher might be technically retained in DCPS but not receive IMPACT scores the following year, such as when a teacher goes on maternity leave too early in the academic year to earn an IMPACT score. Our performance estimates would be biased, for example, if there were a differential probability of a teacher not receiving a next-year IMPACT rating across the ME/D threshold, conditional upon being retained in DCPS. We assess this by estimating our analytic model with the probability of receiving a next-year IMPACT score in the left-hand side of the equation. Our estimates indicate that predicted attrition rates are no different (.012, $p = .623$) for treated (.054) and untreated (.042) teachers; across the overall analytic sample, 4.42% of retained teachers receive IMPACT scores the following year.

RESULTS

Descriptive Evidence

Most teachers experience meaningful improvement in measured effectiveness over time under IMPACT 3.0. In figure 2.2, we sort teachers by their initial (pre-appeal) rating in a given year (t) and follow their performance over the next two years ($t + 2$). In t, most teachers score at least at the Effective level (27.01% HE and 43.36% E), with about one in five teachers (21.5%) scoring at the Developing level, and 6.2% achieving a score that places them at the Minimally Effective level. Fewer than 2% are rated Ineffective in a given year and these teachers are omitted as they are immediately dismissed. Teachers at each performance level, however, exhibit somewhat different trajectories over the next two years.

Among HE teachers, for example, most (59%) are still rated HE two years later, and 17% are rated E. Few HE teachers (1.8%) receive IMPACT ratings below the E level in year $t + 2$ and virtually none were involuntarily dismissed (0.04%). By year $t + 2$, approximately 20% of teachers who were HE in t have departed from DCPS of their own accord; among the voluntary exiters in this group, half exited in the year following their HE rating, and half exited in year $t + 2$ having received mostly ratings of E or above. Annual attrition rates of 10 percent are relatively low compared to most urban districts (Papay, Bacher-Hicks, Page, & Marinell, 2017).

At the E level, a majority of teachers are still earning HE (24%) or E (36%) ratings two years later, with 9% scoring at Developing level, and 2% either ME or I. Here again, few teachers (0.25%) are dismissed within two years of having received an E rating, and 27 percent of these teachers are choosing to voluntary exit teaching in DCPS over this two-year period. Among those Effective teachers who exit, about half (14%) exited immediately after receiving their E rating, and a third of the overall group (9%) chose to leave within two years, even with consecutive ratings of at least Effective.

Developing teachers encompass the new performance category under IMPACT 3.0 that includes a score band under which teachers would have previously been considered Effective. While we do not conduct our RD analysis of IMPACT effects at the D/E threshold in this paper, this score band is nonetheless one of interest given that such scores would have placed teachers in the Effective range where they were at no risk for separation under IMPACT 2.0. If this category were true to its name, we would expect "developing" teachers to improve their performance the following year, and indeed this is on average the case for the D teachers who remain. Among the teachers rated D in a given year t, more than a third (38.5%) have improved to E or HE two years later, another third approximately (34.7%) has elected to leave DCPS of their own accord, and

18.5% are performing at or below Developing. A small but not inconsequential share (6.3%) are involuntarily dismissed within the following two years.

ME teachers, who make up a small share of the overall performance of DCPS educators (6.2%), not surprisingly are performing at higher rating levels, on average, when they are still teaching in DCPS in year $t + 2$, yet many (33%) are dismissed by year $t + 2$ and a similar proportion (35.5%) opt to leave by this point. Notably, most of the voluntary exiters who have received an ME rating do so in the year immediately following ($t + 1$) rather than in year $t + 2$; 24.9% of the ME teachers leave in the year following their ME rating.

Figure 2.2 provides descriptive evidence that teachers' ratings often improve in IMPACT when they are retained, and shows that teachers at lower performance levels leave at meaningfully higher rates than those with higher IMPACT ratings, but it does not illuminate the extent to which IMPACT *causes* teachers to improve or to voluntarily leave DCPS. The RD analysis, on the other hand, explicitly attempts to answer this question.

Regression Discontinuity Analysis

First-Stage Effects. Figure 2.3 shows that the assignment to treatment is not strictly continuous across all IMPACT 3.0 years, due to teachers successfully appealing their IMPACT scores to attain higher ratings. These appeals are concentrated in AY2012-13, which saw a slightly higher share of successful appeals following an error in the value-added calculation for some teachers, with six percent of ME teachers successfully appealing their scores to upgrade to a D rating. For the remaining IMPACT 3.0 years, initial and final rating assignments are nearly strictly discontinuous, with no more than two ME teachers in the sample successfully appealing to a higher rating (D) in a given year.

Regardless, we employ an intent-to-treat analysis with the assumption—supported by Dee and Wyckoff's (2015) findings—that the threat of dismissal associated with an initial rating of ME would be sufficiently compelling for a teacher to either leave the DCPS teaching force or to stay and improve.

Retention. Figure 2.4 provides graphical evidence of large unconditional retention effects (top panel), with far lower average retention among teachers who have scored just below the ME/D threshold in IMPACT 3.0 than those who scored at the D level. When estimated parametrically (table 2.4), we find that these results are large and robust to the inclusion of teacher covariates and school fixed effects, with teachers just below the threshold approximately 11 percentage points less likely to return the following year, an increase in attrition of approximately 40 percent. These estimates are similar in magnitude to those in IMPACT 2.0, where estimates demonstrate roughly a 9 percentage-point decrease in retention. These results suggest that IMPACT 3.0 was at least equally effective at inducing low-performing teachers to voluntarily exit.

We ran additional analyses to explore the sensitivity of results to varying bandwidths and higher-order polynomials—both tests for the functional form of the relationship between IMPACT scores and retention. The inclusion of a quadratic produces a slightly higher point estimate (14%), though the Aikake information criterion (AIC) suggests that the linear model with teacher controls and school fixed effects is a slightly better model fit. In addition, we explore the use of triangular-kernel-weighted observations, in lieu of the uniform weights presented in table 2.4, where greater weight is placed on units closer to the threshold. We find that the use of triangular kernel weights produces estimates at least as large as those with uniform weights (appendix table A.2.3), yet our estimates are sensitive to our choice of bandwidth, highlighting the importance of our assumptions about the functional form between teachers' IMPACT scores and retention for estimating internally valid treatment effects. While larger bandwidths introduce greater precision, they can increase potential bias given that observations further from the cut point could bias effects seen at the threshold. At the bandwidths that balance squared bias and variance to minimize the asymptotic approximation to the mean-squared error (MSE) of the regression discontinuity point estimator (between 9 and 13 points from the ME/D threshold, depending on the method used; see Cattaneo, Idrobo, & Titiunik, 2018a), retention effects are even larger—ranging from 21 to 24 percentage points (see appendix table A.2.3). The estimates at these smaller bandwidths are nearly double that of the estimated retention effect at the full bandwidth (11 percentage points with a bandwidth of ± 50 points). A series of local linear regressions at increasingly small bandwidths, illustrated in appendix figure A.2.2, show that retention effects are larger at smaller bandwidths, and become smaller as the bandwidth increases to 50 points from the consequential threshold, yet the estimated treatment effects remain substantively large across bandwidth choices, and are significantly different from zero at nearly every bandwidth above a size of ten.

Performance. The lower panel of figure 2.4 suggests that there may be performance effects from assignment to treatment for those teachers who choose not to resign from DCPS, with approximately ten points higher average performance among teachers just scoring below D, than those just above the threshold. Parametrically, we estimate an IMPACT 3.0 treatment effect of 12.89 IMPACT points in our unconditional model, which becomes an increase of 11.99 points significant at $\alpha = 0.05$ when we control for teacher covariates and the schools in which they teach. This represents an increase of 27 percent of a standard deviation of IMPACT scores. These performance gains are similar to those observed in the two years of IMPACT 2.0. The inclusion of a quadratic term reduces the size and precision of the estimated performance effect such that it is a no longer statistically distinguishable from zero, though the slightly higher AIC for this model suggests that the linear model with teacher controls and school fixed effects is a better fit.

These performance effects are robust to bandwidth choice, with similar estimated treatment effects on next-year IMPACT scores at MSE-optimal bandwidths (between 10 and 11 IMPACT points) to those at the full potential bandwidth (see appendix table A.2.3). While performance effects at the ME/D threshold are of similar magnitude across the full range of bandwidths, they are imprecisely estimated even at most larger bandwidths where the inclusion of additional observations might be expected to improve precision—at best, treatment effects on teachers' next-year performance are significant at $\alpha = 0.10$. Results from these local linear regressions are presented in the second panel of appendix figure A.2.2. When estimated using triangular kernel weights, effects are also of a similar magnitude (between seven and 11 IMPACT points), though are statistically insignificant across each model specification.

Other Considerations. It is possible that the overall IMPACT 3.0 intent-to-treat effects we observe on both retention and performance mask heterogeneity in treatment effects by year. We therefore estimated effects on retention and performance by year (not shown). Within year, particularly for retention, results are similar in magnitude, though imprecisely estimated. In IMPACT 3.0, our samples decrease substantially due to a combination of compositional changes and the restructuring of rating categories which shrank the size of our treatment and control score bands. Our by-year estimates of treatment effects on teachers' next-year IMPACT scores are fairly stable from year to year, but are in some years more sensitive to decisions about the model specification. Regardless, these by-year estimates, while underpowered, provide suggestive evidence that there may be meaningful effects in each year of IMPACT 3.0, and that the overall effects we see are not driven entirely or even primarily by the first year of program revisions.

DISCUSSION AND CONCLUSION

Ten years ago, several reformers touted teacher evaluation as a mechanism to improve teacher effectiveness and student achievement. Despite often heated debate, virtually every state and school district redesigned its teacher evaluation system in response. Much of the recent public discourse has characterized these reforms as a costly failure that should be abandoned. However, the existing evidence suggests a more nuanced portrait in which these reforms were well implemented and effective in some settings and poorly implemented and ineffective in others. Recent research (Marsh et al., 2017; Donaldson & Woulfin, 2018; Cohen, Loeb, Miller, & Wyckoff, 2019) has informed our understanding of this variation in the implementation of teacher evaluation systems (e.g., suggesting the key role of principal take-up). Without a more thorough and rigorous understanding of whether teacher evaluation can improve outcomes for teachers and students across a variety of contexts and how its design and implementation should be altered to be most productive, it seems rash to label it as yet another failed policy.

There is much yet to be learned about the design and implementation of teacher evaluation across a broad set of contexts to realize and sustain its potential. In this paper, we document how the design of IMPACT has changed since its controversial introduction a decade ago and examine whether the initial effectiveness of IMPACT is sustained in the face of major changes in design and context. There are good reasons to believe these effects may have attenuated in subsequent years. First, the large effects of IMPACT on the improvement in teaching found in AY2010-11 (Dee and Wyckoff, 2015) may have been a singular response to the firings and financial rewards that teachers received in the first year of IMPACT. Second, the context surrounding IMPACT substantially changed over the subsequent eight years. Two new Chancellors and other leadership changes, meaningful design modifications, implementation fatigue and competing priorities, and pressure from stakeholders all could reduce the effects of IMPACT. The large effects we identify here suggest that rigorous teacher evaluation can be sustained over at least an eight-year period. We observe these effects across years implying IMPACT has led to a cumulative improvement in teaching quality and student achievement. These gains benefit students who primarily come from nonwhite, low-income households.

That IMPACT has caused some teachers to improve their skills as measured by TLF is important. The paper shows that IMPACT's differential incentives lead to improved outcomes. Are such incentives sufficient to induce improved outcomes? Null outcomes from experiments where the treatment is solely teacher pay-for-performance cast doubt on this hypothesis. It is more compelling that incentives embedded in a system with the strong supports for teacher improvement produce gains in teacher skills. This hypothesis is consistent with our IMPACT findings. Teachers receive multiple classroom observations per year and formal feedback and coaching following each of these evaluations. This feedback may be key to giving teachers the information necessary to make improvements. In fact, analysis of changes in DCPS teaching practice at consequential thresholds under IMPACT 2.0 (Adnot, 2016) suggests that teachers strategically improve their practices, as measured by the TLF, when incentivized by IMPACT.

The sustained improvements in teacher effectiveness resulting from IMPACT raise important questions about the national discussion of teacher evaluation. First, an important aspect of improvement in DCPS results from the voluntary exit of teachers who face a dismissal threat. Many districts may find dismissal as employed in DCPS an unrealistic sanction for weak performance. Political or labor market constraints may limit performance-based exits. Evidence from districts confronting different contexts would be very useful.

Second, disillusionment with teacher evaluation reform is largely premised on the observation that there has been little change in the percentage of teachers rated less than effective. We know very little about teachers' behavioral responses to being rated as effective in a system where there is a highly effective category. To what extent do teachers rated as effective actively engage to improve their performance? Faithfully implementing teacher evaluation is expensive in time and financial resources. Done well, teacher evaluation requires evaluators to be normed and to visit classrooms at least three times during the year. It also requires thoughtful feedback. While evidence on the extent to which states and districts made these investments is limited, it appears doing so may be the exception.

Finally, virtually everyone agrees that differences in teaching effectiveness make a substantial difference for students across a variety of proximal and distal outcomes. Evidence presented in this paper suggests that the rigorous diagnosis of teaching strengths and weaknesses coupled with feedback intended to improve weaknesses is a powerful form of professional development. We may disagree about the design of teacher evaluation systems—it is easy to disagree in the face of limited evidence—but it seems difficult to make a persuasive case that teachers should not understand their teaching strengths and weaknesses and be provided with expert feedback on how to improve.

TABLE 2.1
IMPACT Score Components 2009-10 through 2015-16

	IMPACT 1.0-2.0		IMPACT 3.0		
IMPACT	2009-10 to 2011-12		2012-13 to 2013-14		2014-15 to 2015-16
Components	Group 1	Group 2	Group 1	Group 2	Groups 1 & 2
Individual Value Added (IVA)	50%	0%	35%	0%	0%
Teaching and Learning Framework (TLF)	35%	75%	40%	75%	75%
Teacher-Assessed Student Achievement Data (TAS)	0%	10%	15%	15%	15%
Commitment to the School Community (CSC)	10%	10%	10%	10%	10%
School Value-Added (SVA)	5%	5%	0%	0%	0%

Notes. Group 1 consists only of those reading and mathematics teachers in grades for which it is possible to define value added with the available assessment data. IMPACT scores can also be adjusted downwards for "Core Professionalism" (CP) violations reported by principals. Group 1 teachers did not have IVA calculated during the first two years of the PARCC exam (AY2015 & AY2016); in those years, Group 1 teachers had the same score components and weights as Group 2 teachers.

TABLE 2.2

IMPACT Ratings, Separation and Extra Compensation Criteria, 2009-10 to 2014-15

Category			2009-10 to 2011-12	2012-13 to 2014-15
Scoring Bands for Performance Ratings			100-174: Ineffective (I) 175-249: Minimally Effective (ME) 250-349: Effective (E) 350-400: Highly Effective (HE)	100-199: I 200-249: ME 250-299: Developing (D) 300-349: E 350-400: HE
Separation Criteria			Separation after 1 I rating, or 2 consecutive ME ratings	Separation after 1 I rating, 2 consecutive ME ratings, 1 D followed by 1 ME rating, or 3 consecutive ratings below E
Compensation	Bonus Pay	Eligibility	Teachers in all schools scoring HE	Teachers in all schools scoring HE
		FRPL >= 60%	$10,000, plus $10,000 for teachers in Group 1, plus $5,000 for teachers in high-need subject	$10,000, plus $5,000 for teachers in Group 1, plus $10,000 for teachers in 40 lowest-performing schools
		FRPL < 60%	$5,000, plus $5,000 for teachers in Group 1, plus $2,500 for teachers in high-need subject	$2,000, plus $1,000 for teachers with value-added
	Base Pay Increase	Eligibility	Teachers in all schools	Only teachers in schools with >=60% FRPL[1]
		FRPL >= 60%[1]	2 consecutive years of HE ratings = Masters' band + 5-year service credit	Advanced teacher: 2-year service credit Distinguished teacher: Master's band + 5-year service credit Expert teacher: PhD band + 5-year service credit
		FRPL < 60%	2 consecutive years of HE ratings = Masters' band + 3-year service credit	None

[1] Teachers must be "teaching in a high-poverty school during the year in which you qualify for a service credit, and during the following school year" in order to be eligible for the base salary increase (LIFT guidebook, 2012-13, page 18).

TABLE 2.3
Sample Characteristics

Variable	Mean (SD)
Retention Next Year	0.75
Next-Year IMPACT Score	297
Initial IMPACT Score	269
Group 1	0.25
Female	0.72
Gender Missing	0.01
Black	0.56
White	0.20
Hispanic	0.05
Graduate Degree	0.62
0-3 Years of Experience	0.32
4-9 Years of Experience	0.30
10+ Years of Experience	0.35
AY 2012-13	0.35
AY 2013-14	0.34
AY 2014-15	0.31

Note. The sample consists of 1,809 general-education teachers in the 2012-13 through 2014-15 academic years who either received a Minimally Effective or Developing rating, but were not rated Minimally Effective in the prior year. See text for further details.

TABLE 2.4

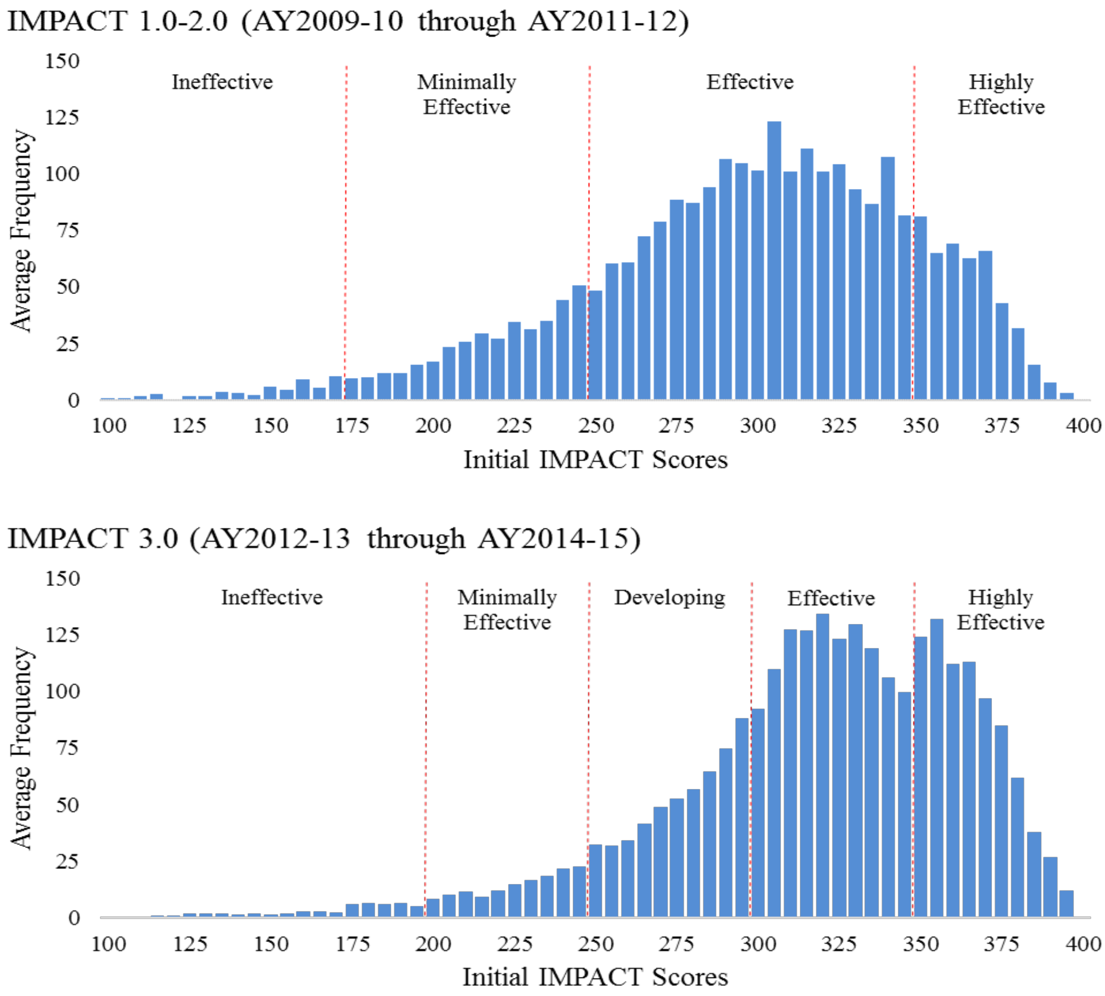
Reduced-Form Minimally Effective ITT RD Estimates on Teacher Retention and Performance, by IMPACT Phase

	Retention				Next-Year IMPACT Score			
Sample	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
IMPACT 2.0								
$I(S_{it} < 0)$	-0.093 *	-0.090 *	-0.092 *	-0.092	9.03 $^{+}$	8.01 $^{+}$	7.03	8.73 $^{+}$
	(0.046)	(0.044)	(0.042)	(0.062)	(4.93)	(4.80)	(4.43)	(6.41)
	1,874	*1,874*	*1,874*	*1,874*	*1,439*	*1,439*	*1,439*	*1,439*
Teacher controls	No	Yes	Yes	Yes	No	Yes	Yes	Yes
School fixed effects	No	No	Yes	Yes	No	No	Yes	Yes
Quadratic of running variable	No	No	No	Yes	No	No	No	Yes
AIC	2018	1909	1727	1729	14380	14339	14106	14110
IMPACT 3.0								
$I(S_{it} < 0)$	-0.117 *	-0.104 *	-0.114 *	-0.138 *	12.89 *	12.19 $^{+}$	11.99 *	8.31
	(0.052)	(0.050)	(0.047)	(0.069)	(6.52)	(6.45)	(6.04)	(9.09)
	1,809	*1,809*	*1,809*	*1,809*	*1,270*	*1,270*	*1,270*	*1,270*
Teacher controls		X	X	X		X	X	X
School fixed effects			X	X			X	X
Quadratic of running variable				X				X
AIC	2130	2034	1853	1856	12752	12736	12509	12511

Notes. Robust standard errors are in parentheses and sample sizes are in italics. Models include year fixed effects and employ uniform kernel weights. Treatment effects are estimated off of teachers who were not rated ME in the prior year. Teacher covariates include gender, race, education, experience, and an indicator for whether the teacher is in a tested grade and subject (Group 1). We exclude AY2009-10 (IMPACT 1.0) because of evidence that IMPACT was not truly implemented at that point.

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; and $^{+}$ $p < 0.10$

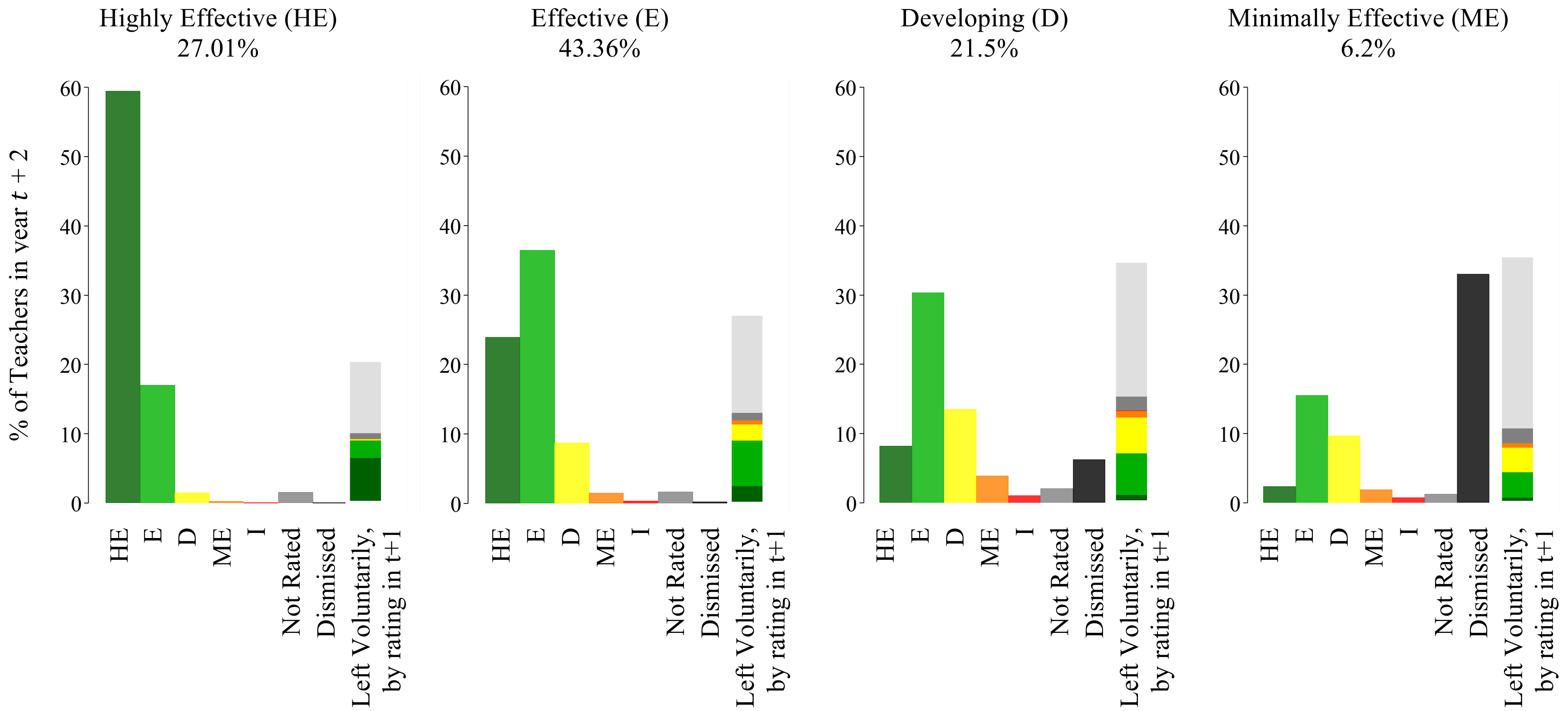
FIGURE 2.1. *Distribution of IMPACT Scores by Year and Rating*


IMPACT 1.0-2.0 (AY2009-10 through AY2011-12)

IMPACT 3.0 (AY2012-13 through AY2014-15)

Notes. IMPACT scores reported here are initial scores, assigned prior to the appeals process. Very few appeals result in revised scores. Sample consists of all general education teachers in DCPS between AY2009-10 and AY2014-15. The distribution of scores around the Effective/Highly Effective scores may indicate potential manipulation of scores; while it is possible such manipulation is occurring at this point in the distribution—given that teachers with consistently high performance are subject to fewer classroom observations and can therefore see their overall scores more-easily changed by a single classroom observation—this threshold is not one we focus on in this paper.

FIGURE 2.2. *Rating & Retention in Year t+2, by Initial Year t Rating*



Notes. Figures exclude teachers rated Ineffective (I), given that an I rating is grounds for immediate dismissal. Fewer than 2% of all teachers have received an I rating in IMPACT 3.0. Teachers who are not rated are those still employed by DCPS but not teaching in a given year (e.g., teachers on temporary leave). Reported ratings are based on teachers' initial IMPACT scores, assigned before the opportunity to appeal for a higher rating. As discussed in the following section, however, few teachers successfully appeal and receive different final scores from those initially assigned.

FIGURE 2.3. *First Stage: Effect of Initial IMPACT Score on Pr(Minimally Effective) at the Consequential Cut-Off*

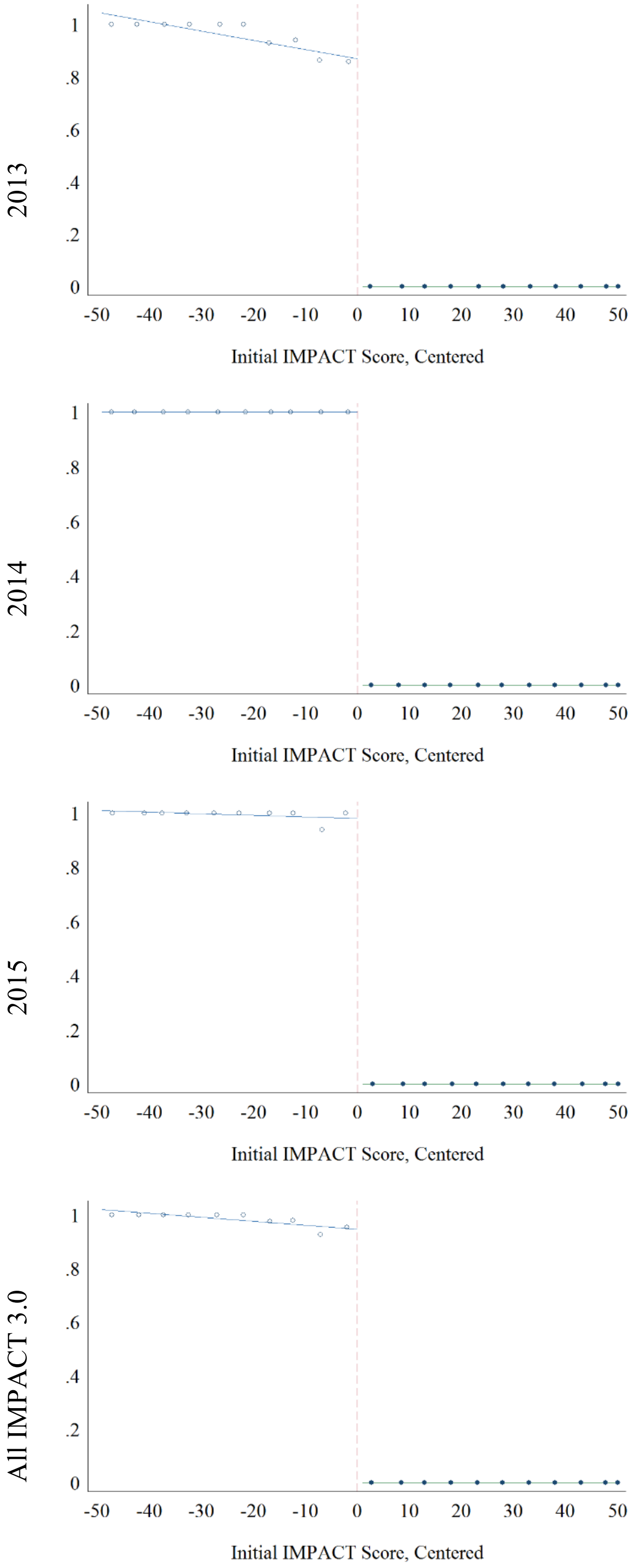
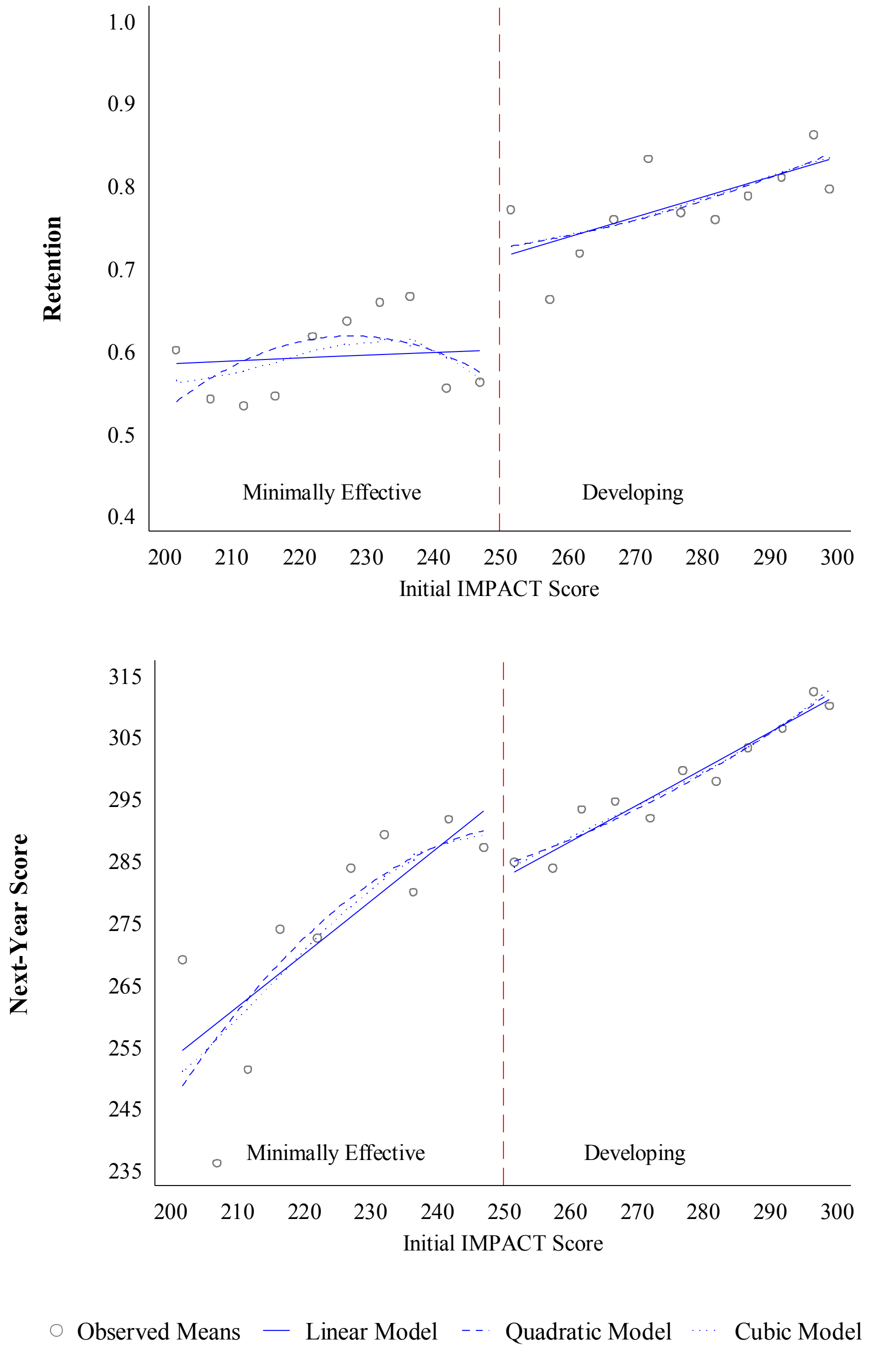


FIGURE 2.4. *Treatment Effects at the Minimally Effective Threshold*



Notes. Each plotted point represents the mean outcome for a given bin (width=5 IMPACT points) of initial (pre-appeal) IMPACT scores. Note that we test for discontinuous retention effects below the ME threshold, given that there is an apparent drop in the probability of retention for teachers with initial IMPACT scores between 240 and 244. We do this by running a regression with placebo treatment effects at points away from the true cut-off (shown in appendix table A.2.2), and by testing for differences in mean retention and mean teacher characteristics across bins (not shown); neither test indicates discontinuous effects at any point other than the true threshold.

CHAPTER 3

Exploring the Development of Teaching Skills in DCPS

(with Eric Taylor and James Wyckoff)

Abstract – Teachers are critical for their students' success across myriad outcomes, and effects are more pronounced for students who come from the most disadvantaged backgrounds. There is, however, considerable variation in the quality of teachers that students have access to, and much of this variation exists not simply across, but also within, teachers. It has now been well established empirically that teachers make significant improvements to their effects on student achievement as they gain experience teaching, particularly in their earliest years in the classroom. Less known, however, is whether teachers make similarly large early-career gains in terms of the practices and skills they exhibit in the classroom and whether these improvements might facilitate student learning. In this paper, we engage in one of the first attempts to quantify such returns to experience for novice teachers' practice, using administrative data from the District of Columbia Public Schools (DCPS), which has an unusually rich panel of data that includes scores on a rigorously-administered teacher observation rubric. We find that there are large returns to experience for overall practice measured in DCPS, but with variation in gains across practices, as well as across teachers. Importantly, we also establish an association between teachers' improved practices and their students' learning gains, suggesting that when teachers improve their skills, their students benefit, as well.

INTRODUCTION

Teachers are critical for their students' success, significantly impacting student learning (Aaronson, Barrow, & Sander, 2007; Chetty, Friedman, & Rockoff, 2014b; Kane & Staiger, 2008; Rivkin, Hanushek, & Kain, 2005; Rockoff, 2004), motivation (Ruzek, Domina, Conley, Duncan, & Karabenick, 2015), and social development (Gregory, Allen, Mikami, Hafer, & Pianta, 2016), these effects can persist well into their students' adulthood (Chetty, Friedman, & Rockoff, 2014b). The extent to which individual teachers drive these outcomes, however, can vary substantially. Whether measured by observations of teaching in the classroom (Kane & Staiger, 2012), their students' achievement scores (Rivkin et al., 2005; Rockoff, 2004) or non-test outcomes (Ruzek et al., 2014), there are large differences in teachers' effectiveness. Meanwhile, the least effective teachers may disproportionately teach in classrooms with low-income, low-achieving students (Goldhaber, Quince, & Theobald, 2016; Sass, Hannaway, Xu, Figlio, & Feng, 2012), though the size of these differences may not be large (Isenberg et al., 2016).

Key to this variation are not simply differences across, but also within, teachers. Teacher effectiveness is not fixed within teacher; rather, as teachers gain experience, they generally become significantly better at facilitating student learning, as measured by achievement on standardized tests, particularly early in their careers (e.g., Rockoff, 2004). Such "returns to experience" have been documented over time and across states and school districts (e.g., Harris & Sass, 2011). The literature on teacher improvement consistently demonstrates that the typical novice teacher meaningfully improves over the first five years of her career. These studies, however, have to date relied almost entirely on analyses of student test score data. With few notable exceptions (e.g., Kraft & Papay, 2014; Kraft, Papay, & Chi, 2018; Papay & Laski, 2018), these studies simply demonstrate teachers' improvements without providing guidance on which teaching skills may explain these improvements or how improvements in teaching skills directly relate to student learning gains.

Two distinct literatures explore the early career development of teachers. First, developmental and educational psychologists have long hypothesized how teachers build the many inter-related skills that improve their effectiveness. More recently, policy analysts with access to increasingly rich administrative data have examined how teachers grow in their ability to improve student achievement over the early years of their careers. These two literatures have largely persisted in isolation of each other, to the detriment of our understanding of teacher development.

Understanding how teachers improve—and how those improvements relate to student learning and behavioral development—is vital for developing effective teacher preparation programs and improving existing professional development curricula. While researchers have for years investigated the ways in which pre-service training and in-service professional development cultivate and facilitate quality teaching, we still lack robust empirical evidence as to the core teaching practices that are foundational for teachers' skill and effectiveness in the classroom. By improving our knowledge of teachers' patterns and trajectories of skill attainment, and which skills are most clearly associated with student gains, we might better understand how to produce more-highly skilled teachers from the outset, as well as how to help both pre-service and in-service teachers develop the skills necessary for their students to have the best opportunity to succeed. This is particularly of concern for novice educators, who on average perform well below their teaching potential yet are disproportionately assigned to teach the least advantaged children (Boyd, Lankford, Loeb, Rockoff, & Wyckoff, 2008; Clotfelter, Ladd, & Vigdor, 2005, 2006; Jackson, 2009; Kalogrides, Loeb, & Bèteille, 2012; Lankford, Loeb, & Wyckoff, 2002; Sass et al., 2012).

RESEARCH QUESTIONS

As modern teacher evaluation systems mature, we are newly able to collect nuanced and reliable data on teachers' practice over time for large samples of educators, and improved data collection systems allow us to link these teacher evaluation scores to student outcomes. The teacher evaluation system in the District of Columbia Public Schools (DCPS), known as IMPACT, has been in place since the 2009-10 academic year (AY) and provides a rich panel of data from which we can better understand teacher development over time, and how that development relates to student achievement.

Using these panel data, we first explore patterns in teacher development over time to understand how teachers improve during their early careers. Despite an abundance of evidence that teachers improve their ability to influence student learning over their early careers, we know little about the specific skills teachers are developing over that time, which skills may be most important for teachers' development and student learning, or the extent to which these returns to experience are acquired through experience teaching in a specific context. Specifically, in this paper we ask 1) what is the pattern of overall skill development during teachers' early careers; and 2) which practices best explain this gain (i.e., which practices contribute most to teachers' overall improvement)?

A key assumption for the importance of these question however, is that practice-based returns to experience are important for student learning. We test this assumption with a secondary research question, which asks how teachers' improvements in practice relate to student achievement gains. Specifically, we ask whether 1) teachers whose skills improve the most also experience the largest gains in student achievement; and 2) whether some teaching skills are more directly associated with improvements on student achievement than others.

Before we estimate whether teachers make meaningful gains to their practice over their early career years and whether teachers' improvement on the TLF is related to improvements in their students' learning gains, we must first establish that there is sufficient variation in TLF scores, given that many observation systems fail to adequately distinguish teacher quality (Kraft & Gilmour, 2017; Weisberg, Sexton, Mulhern, & Keeling, 2009). Figure 3.1 shows the distribution of TLF scores for all DCPS teachers between the 2009-10 and 2015-16 academic years, with scores from master educators (MEs) in the top panel and scores from all observers in the bottom panel. While the median teacher receives an overall score above 3 on a scale that ranges from 1 to 4, we observe scores across the full range of the TLF regardless of the type of rater assigning scores. The average teacher scores 3.02 on observations by external evaluators (3.13 when including administrator-assigned scores), with a standard deviation of 0.52 (0.48 for all observers).

BACKGROUND

Returns to Experience for Student Achievement

Given the large body of research documenting the importance of teachers for student achievement and other outcomes and the extent to which teachers' effectiveness varies across and within settings, researchers have in turn established that teachers' experience is far more informative than traditional measures of teacher qualifications (e.g., certification and licensure, advanced degree attainment, and preparation route), which are generally poor predictors of teachers' effectiveness in the classroom, teachers' experience is far more informative. Numerous studies making use of value-added scores—estimates of teachers' contributions to student achievement on standardized assessments—demonstrate that teachers make considerable gains to their performance based on their years of experience, and particularly so during their first three to five years of teaching (Boyd, Lankford, Loeb, Rockoff, & Wyckoff, 2008; Harris & Sass, 2011; Papay & Kraft, 2015; Rockoff, 2004; Wiswall, 2013).

There is, however, considerable heterogeneity in the gains that teachers make over their early careers. For instance, Atteberry, Loeb, and Wyckoff (2015) used panel data from the New York City Department of Education to examine teachers' early-career improvement, and found that the lowest-performing quintile of new teachers makes the steepest gains in value added over their first five years in the classroom, yet these teachers fail by that point to attain the average value-added of first-year teaches overall. Xu, Özek, and Hansen (2015) find similar performance patterns in Florida and North Carolina.

This finding raises the question of what might explain differences in teachers' learning trajectories, in addition to their variation in initial performance. Some of these gains can be explained by the contexts in which teachers teach. For example, some evidence suggests that teachers in low-poverty schools may benefit from larger returns to experience than teachers in high-poverty schools (Sass et al., 2012), though these differences may be attributable to sorting (Xu et al., 2015). Other school contexts may also be important for teachers' development on the job; teachers working in more supportive professional environments exhibit larger gains in value-added (Kraft & Papay, 2014) and teachers' improvements in general can also be influenced by the quality of their peers (Jackson & Bruegmann, 2009; Papay, Taylor, Tyler, & Laski, 2016; Sun, Penuel, Frank, Gallagher, & Youngs, 2013).

Consistency of teaching assignments is also associated with returns to experience for teachers' value added. A small handful of studies have attempted to distinguish generic returns to experience from task-specific returns to experience—performance gains associated with teaching a certain course or grade level. Ost (2014) finds that elementary teachers in grades three through five who remain teaching at the same grade level exhibit larger returns to experience than those who switch grades, with the largest grade-specific human capital effects for the least experienced teachers. Cook and Mansfield (2017) find that task-specific experience is similarly important for high school teachers; teachers with additional years teaching the same subject experience larger returns to experience than similarly-experienced teachers who have switched subjects.

Returns to Experience for Teachers' Practice

This literature establishes potential mechanisms for teachers' improvement over time, but provides little evidence in support of the knowledge and skills teachers develop over their early careers and how such skill development relates to teachers' improvement in value added. Two largely distinct bodies of literature have contributed to our understanding of teachers' development: the education policy literature examines teacher development from the perspective of changes in teachers' abilities to improve student outcomes but has only recently attempted to understand how and why teaching skills and capacity evolve. The teacher education and development literature has attempted to understand the factors that enable or hinder teacher development, but have typically done so for narrowly defined components of the development process, at small scale, or without attention to the rigor of methods.

Conceptualizing Teacher Development

The recent educational psychology literature depicts the development of teaching skills as multi-dimensional, situational, and dynamic, as opposed to defining teachers simply as effective or ineffective. Teachers need to develop many instructional skills, which may develop in different ways during their early years of teaching, in order to be their most effective (Malmberg, Hagger, Burn, Mutton, & Colls, 2010). Recent analyses of thousands of classroom observations conducted with multiple observational measures suggest at least two core domains of teaching skills—general instructional methods and classroom management (Ferguson & Danielson, 2014). Teachers need general instructional skills such as how to effectively convey content, get students thinking deeply about their own learning, and provide responsive feedback, but they also need classroom management skills so they can successfully engage groups of children and adolescents in meaningful learning opportunities. Of course, teachers also need a range of content-specific instructional methods to support students' learning in areas such as reading, math, and science. However, for the purpose of this paper, which seeks to understand the development of teaching skills across grade levels (PK-12) and content areas, we focus on domains of teaching that are relevant irrespective of subject matter.

The development of teaching skills is also conceptualized as situational and dynamic—that is, dependent on the settings in which teachers work and the complex ways a teacher's characteristics intersect with those settings over time. Although some of the knowledge and expertise needed to display these skills can be developed during pre-service teacher education programs (Cochran-Smith et al., 2016), teachers learn much about how to be effective as they first interact with students in the classroom and as they intersect with other professionals in the schools in which they work. Teachers' opportunities to develop are shaped by the new knowledge available to them in both formal (e.g., mentoring, workshops, courses, coaching) and informal (e.g., grade-level planning meetings, hallway conversations about a challenging student) learning opportunities (c.f., Desimone et al., 2014). Teachers apply this knowledge in their classrooms, refine their skills through trial and error and, over time, learn how to more effectively engage students and convey content.

Teachers' Instructional Improvements Over Their Early Careers

This research on teachers' development notably lacks detailed documentation of observed changes in teaching practice, and studies are often small scale or do not include tests of statistical significance. In a review of the literature on beginning teacher mentoring and induction, Ingersoll and Strong (2011) report that across five studies with observations of beginner teachers during their first years of teaching, the number of beginning teachers ranged from 6 to 287 and two of the five studies did not carry out tests of statistical significance. Some of these studies report interesting trends, but issues of generalizability and the rigor of measurement limit the usefulness of these studies. This makes clear that large observational studies documenting teachers' improvements can play an important role in expanding our understanding of returns to experience in teaching.

Until recently, data on teachers' observed performance over time has been largely limited to individual research studies spanning short periods. Estimation of returns to experience for teachers' practice has only recently been a feasible research venture due to considerable data limitations. Even a decade ago, most existing evaluation programs failed to differentiate levels of quality across teachers and rarely provided educators with actionable feedback for improving their instruction. Observations were often infrequent and uninformative; teachers who had already received tenure were particularly unlikely to receive meaningful feedback on the quality of their teaching (Weisberg, Sexton, Mulhern, & Keeling, 2009), leaving few opportunities to track teacher development over several years on a consistent evaluation measure. Following a series of reforms to teacher evaluation encouraged by research from the Measures of Effective Teaching project (Kane, McCaffrey, Miller, & Staiger, 2013) and federal initiatives such as Race to the Top, the Teacher Incentive Fund, and waivers under No Child Left Behind (NCLB) states and districts began to formalize and standardize their classroom observation processes, with the number of states requiring annual observations for each teacher nearly doubling between 2009 and 2015 (Doherty & Jacobs, 2015).

Early evidence from Tennessee (TN), a state that received Race to the Top funding and with a similarly long-lasting standards-based observation system to DCPS, suggests that teachers make large gains in their practice over their early careers that are similar in shape to the performance gains made to value-added for novice teachers. Papay and Laski (2018) show that teachers in TN make the steepest gains on the TEAM observation rubric in their first few years in the classroom, improving by .80 standard deviations on average by their fifth year of teaching, and by close to a full standard deviation after ten years in the classroom. Papay and Laski find that TN teachers exhibit larger returns to experience in the Instruction domain of the TEAM rubric than they do in the Environment domain, which comprises subskills including management of student behavior and fostering a respectful culture, and larger gains in Environment than Planning (i.e., instructional plans, student work, and assessment). It is unclear from their data, however, how teachers' development interacts across the three TEAM-assessed domains, or how teachers' improvement on the overall measure or its subcomponents relates to student learning gains.

The Development of Teaching Skills and Changes in Student Achievement

The goal of understanding the development of teaching skills is, ultimately, to understand the ways in which this development promotes positive outcomes for students. Rigorously conducted observations of teachers' skill are associated with gains in student learning. Although effect sizes are typically moderate (Taylor & Tyler, 2012), these associations have been documented across multiple contexts, grade levels, and using multiple observational rubrics. Well-designed and implemented PD, including instructional coaching, that changes teaching practice can lead to improvements in student learning (Desimone & Garet, 2015; Kraft, Blazar, & Hogan, 2016), although the effect sizes on student outcomes are often modest. At least one coaching intervention found preliminary evidence that improvements in student learning were, in part, mediated by improvements in observed measures of teaching (Allen, Pianta, Gregory, Mikami, & Lun, 2011). But beyond this PD and intervention research, we know little about whether systematic improvements in observed measures of teaching practice in the early years of teaching, among a large and diverse sample of teachers, are associated with improvements in student learning or engagement.

DATA AND SAMPLE

We have access to administrative data from DCPS that are particularly well-suited to examining the early-career development of teaching skills and distinctive in several respects. At the teacher level, our data include teachers' race/ethnicity, gender, age, and teaching experience. Most importantly, we are able to observe early-career improvement in specific teaching skills for several cohorts of individual teachers, beginning with the cohort of educators who entered teaching in the 2009-10 academic year (AY). These data include teachers' performance across several evaluation components, including on the district's observation rubric and its subcomponents.

Observations of Teaching

DCPS employs a rigorous evaluation system that is relatively more mature than in other school districts, making their data well suited to understanding how teachers' practice develops with experience. Importantly, DCPS has used a rigorously-implemented, standards-based observation protocol—the Teaching and Learning Framework (TLF)—to evaluate every teacher's instruction several times a year since they introduced their teacher evaluation program, IMPACT, in the fall of 2009. [32] The TLF consists of nine teaching practices on which teachers are evaluated up to five times per year (see table 3.1).[33] Teachers receive a score for each subcomponent, averaged across observations within a given year, as well as a total TLF score, which is equal to the unweighted average of each of the nine TLF subcomponents. Teachers are rated on a scale from 1 through 4, with each score value within an individual classroom observation and for a given TLF component reflecting rubric-defined performance criteria.

While teachers in DCPS are evaluated by a combination of internal (administrator) and external ("Master Educator") evaluators, we rely primarily on the scores assigned by external evaluators, though we explore the sensitivity of our analysis of TLF performance by the type of rater. We do this because, while school administrators typically have more reliable scores, external evaluators generally assign scores that are more strongly associated with objective measures of teacher quality, even when adjusting for reliability (Gill, Shoji, Cohen, & Place, 2016; Ho & Kane, 2013; Meyer, 2016; Whitehurst, Chingos, & Lindquist, 2014); the external evaluators' scores are likewise less subject to ceiling effects, as administrators in our sample tend to rate teachers' performance more highly than the master educators (MEs). A recent G study of reliability

[32] The TLF is a modified version of Danielson's Framework for Teaching, and was used in DCPS between AY 2009-10 and AY 2015-16. TLF scores are part of a high-stakes teacher evaluation system called IMPACT. For more discussion of IMPACT, see Dee and Wyckoff (2015).

[33] The number of observations a DCPS teacher receives is dependent on prior performance and experience. All teachers received five formal observations through AY 2011-12. Starting in AY 2012-13, the number of formal observations decreased to four for most teachers, and the district introduced a new career ladder that, in addition to compensation, tied the number of evaluations a teacher received to performance over time such that teachers who consecutively scored in the highest performance band could receive as few as two observations annually by the end of their fifth year in the classroom.

and validity for the TLF found ME reliabilities to range between 0.64 and 0.69 between academic years 2010 and 2014 (Meyer, 2016), levels consistent with those found in other studies of classroom observation measures (Kane & Staiger, 2012). TLF scores are also moderately correlated with teachers' value-added scores, with correlations exceeding 0.30 in each year of analysis.

To ease interpretation, we reduce the nine teaching domains to two distinct factors. As is common with classroom observations (Adnot, 2016), the domains of practice captured by the TLF are inter-correlated; by reducing the number of dimensions to two uncorrelated factors, we can better isolate trends in teachers' returns to experience. In other observation rubrics, the assessed domains typically load onto at least two factors: instruction and classroom management (Ferguson & Danielson, 2014; Hafen et al., 2015; Kane & Staiger, 2012). The TLF, however, fails to load onto more than one factor at commonly-accepted thresholds for factor loadings, perhaps due to the high-stakes nature of evaluation in DCPS (Adnot, 2016). Conducting a principal-components factor analysis, we therefore force the data to load onto a second factor (appendix table A.3.2). For this analysis, we use master-educator (ME) scores for all novice teachers in DPCS with complete evaluation scores between AY2009-10 and AY2011-12. This produces a dominant factor for this population that is highly correlated with Teach 3 and Teach 6 through 8, which are oriented around the *classroom environment and lesson accessibility*; the secondary factor captures TLF components that address *instructional clarity and student understanding* (predominantly Teach 1, Teach 2, and Teach 5). Though we rely primarily on the two-dimensional TLF factors for our analyses, we report on results from the full set of TLF sub-scores in the appendices to this paper.

Analytic Sample

To ensure appropriate comparisons across observer types, we first limit our sample to teachers who have received scores from master educators (MEs) and school administrators, as is built into the design of IMPACT. Typically, teachers are missing scores from one of these categories of evaluators near the beginning or end of the school year, suggesting that this may occur for teachers who move in and out of DCPS during the school year. In addition to ensuring that score comparisons are made across equivalent groups, this sampling restriction may also lessen the probability of attrition bias if, for example, teachers' scores tend to decline in the observation window leading up to their departure.

We additionally limit our primary analyses to teachers who have received scores from each category of observer and on each of the nine domains of the TLF. One domain in particular has relatively high missingness—Teach 6 (*Respond to student understanding*). In each year of our analysis, early childhood education teachers were able to waive these scores if student misunderstandings were not observed during their classroom evaluation; this occurs for approximately 8% of early-childhood teachers' evaluations, or 1-2% of overall teacher evaluations in each year. In AY2011-12, however, the language of the rubric was altered slightly, allowing this to occur for any teacher in DCPS, and close to 25% of all teachers did not receive a Teach 6 score that year. Across the years of analysis, there are 6,678 unique teachers, and 18,720 teacher-by-year observations, who have scores on all nine TLF standards from both Master Educators and school administrators. Given that occurrences of non-applicable Teach 6 scores are higher for teachers who tend to score more highly on the other Teach standards, this sampling restriction leads to lower overall TLF scores than might have been observed had these scores been included; for this reason, we include as robustness checks results across the sub-standards that include the set of teachers missing scores on Teach 6.

A final sampling restriction is applied to our first research question. Given that our primary interest is returns to experience for novice educators, for this research question we limit our sample to teachers for whom we have observation scores in their first five consecutive years of teaching; this yields three cohorts of DCPS teachers totaling 120. By limiting the sample to novice teachers with consecutive experience, we limit the risk of confounding improvement with teachers' selection out of teaching. Nevertheless, we test the robustness of estimates to sample-restriction decisions. Similarly, some of the analyses require data from later years of experience to avoid confounding experience effects with year effects, as described in the following section of this paper. With the exception of one estimation method, which identifies experience effects off of teachers with discontinuous careers, these analyses make use of all observations for teachers who are observed with consecutive years of teaching in DCPS.

Our second research question pertains to student achievement, and so for this question we limit the sample to those teachers who can be linked to students in tested subjects and grades and are in their first five years of teaching. DCPS provides us with linking and dosage rosters that allow us to connect approximately 40,000 individual students to 1,079 teachers in tested grades and subjects, and adjust teacher effects for the amount of time each student spends with his or her teacher when estimating teachers' value-added scores. Student-level data include such information as students' demographic characteristics (e.g., race/ethnicity, gender, age, grade level, English language learner status, and special education status), and test scores.[34] Within our larger analytic sample, there are 847 teacher-year observations that are in their first five years of teaching and can be linked to either of the two student outcomes we explore here (i.e., math and reading achievement), of which 275 unique teachers are observed at more than one level of experience.

METHODS

Establishing Returns to Experience on the TLF

To explore teachers' returns to experience on the TLF, we first standardize teachers' TLF scores across each year of experience to the mean and standard deviation of novice-year TLF scores. This allows us to interpret each teachers' subsequent TLF score gains relative to the typical first-year teacher's performance on the measure.

We first explore these trends graphically, by observing intercepts and learning trajectories for teachers who are observed in DCPS for their first five consecutive years of teaching. We do this for the overall TLF score earned, as well as for the core domains identified by our factor analysis, and for each of the nine subcomponents that comprise the TLF. The parameters for our visual analysis are estimated using the restricted sample of 120 teachers who are observed for at least five consecutive years from their initial year teaching. To do so, we fit:

$$TLF_{jt} = f\left(YrsExp_{jt}\right) + \tau_j + \theta_t + u_{jt}, \qquad (1)$$

where TLF_{jt} is teacher j's observation score (or sub-score) in year t, τ_j is a teacher fixed effect which permits us to estimate experience effects based on within-teacher variation

[34] Test data in DCPS come from the DC Comprehensive Assessment System (CAS) exam through AY 2013-14, and from the Partnership for Assessment of College and Career Readiness (PARCC) exam from AY 2015-16 onward.

in observation scores, θ_t is a year fixed effect to account for possible changes in how teachers in DCPS were scored on the TLF over time, and u_{jt} is an idiosyncratic error term.

Experience is included in the model as a set of indicators for teachers' experience, with each indicator representing e years of experience up to year E, as in (2):

$$f(YrsExp_{jt}) = \sum_{e=1}^{(E-1)} \beta_e \mathbf{1}\{YrsExp_{jt} = e\} + \beta_E \mathbf{1}\{YrsExp_{jt} \geq E\} \qquad (2)$$

The first year of teaching ($e = 0$) is the reference year in this model. The inclusion of $\mathbf{1}\{YrsExp_{jt} \geq E\}$ in the model prevents collinearity between experience and time; this indicator allows us to identify year effects from teachers at or above experience level E. The estimates of β_e can then be plotted to illustrate the slope of improvement for the average DCPS teacher, with confidence intervals that account for the use of multiple observations per teacher per year.[35]

This restricted-sample approach, however, limits us to teachers who are retained in DCPS; these teachers may experience different growth trajectories than those who have attrited, given that low-performing early-career teachers have been documented to leave the profession voluntarily at higher rates than their relatively higher-performing peers (Goldhaber, Gross & Player, 2011; Hanushek, Kain, O'Brien, & Rivkin, 2005). This is particularly a concern in DCPS, where the high-stakes nature of IMPACT has been shown to incentivize low-performing teachers to voluntarily leave the district (see

[35] The fixed effect estimator in this model will assign greater weight to the levels of experience at which teachers exhibit the most variation, making β_e a weighted average across teachers. We therefore also test the robustness of β_e estimates to a version of (1) in which each teacher is equally weighted.

Dee & Wyckoff, 2015, and chapter 2 of this dissertation).[36] While this approach allows us to estimate longitudinal within-teacher returns to experience with potentially high internal validity, it limits the external validity of our estimates; these teachers may well experience different performance trajectories than their peers who attrit or who can be observed only in non-consecutive early-career years. Importantly, the restrictions imposed by this sample also reduce the precision with which we can estimate returns to experience, given that it only comprises 120 teachers, and values of E cannot exceed 5 in our seven-year panel while avoiding collinearity between year and experience. [37] By estimating experience effects based on a larger population (i.e., those who enter and *exit* the DCPS teaching force), we can have better statistical power, while still estimating within-teacher returns to experience.

This approach for estimating returns to experience is sometimes referred to as a censored growth model (CGM), and makes the assumption that there are no additional returns to experience after year E. There are a number of additional methods for estimating returns to experience, described in detail by Papay and Kraft (2015), each of which makes different assumptions for the identification of internally valid estimates. While the other methods may introduce new sources of bias, as discussed in the

[36] Given the high-stakes context of DCPS's teacher evaluation program, we may observe higher returns to experience than would occur in other districts, as teachers at certain performance levels are directly incentivized to improve (see Dee & Wyckoff, 2015, and chapter 2 of this dissertation). For this reason, each of our analytic samples exclude any teachers whose pattern of performance over time will have led to involuntary separation. Regardless, the returns to experience observed in DC might, conditional on the types of teachers who select to teach in DCPS, exhibit steeper performance trajectories than in a typical district with lower stakes for teachers' performance.

[37] If E is defined at a level where teachers are still demonstrating returns to experience, estimates may bias year effects by capturing experience effects for teachers with experience at or above E (Rockoff, 2004; Papay & Kraft, 2015). We test for sensitivity to different values of E, giving preference to values at which estimates of β_e do not change and above which there is sufficient variation that year and experience are not collinear.

following section, we estimate each in order to test the robustness of our CGM results to alternative modeling decision.

The indicator variable model, for instance, used by Harris and Sass (2011), models experience as a function of dummy variables representing bins of years of teacher experience (e.g., 1-2, 3-4, 5-9, 10-14, 15-24, and 25 or more). This approach has the benefit of allowing for the modeling of effects throughout teachers' careers but is arguably least suited to our purpose in this paper because: 1) our research questions are focused specifically on teachers' early careers; and 2) estimates from this method can be biased if teachers' skills change meaningfully within bins of experience—something we expect to be true for the bins of interest.

Another approach that has previously been used to estimate returns to experience makes use of the full sample of teachers, including those with discontinuous levels of experience (e.g., from taking a year of leave), as proposed by Wiswall (2013). These teachers, who are typically omitted from analyses of returns to experience, in this instance serve to identify changes in experience. While solving the problem of perfect collinearity between experience and year, this discontinuous career model (DCM) requires that teachers with discontinuous teaching be representative of the rest of their district in terms of their experience trajectories. The cause of this discontinuity must likewise be unassociated with teachers' improvement trajectories, an assumption which would be violated if for instance teachers were likely to temporarily leave their district following an illness which also negatively impacted their performance.

Finally, Papay and Kraft (2015) propose a fourth approach to estimating returns to experience, using a two-stage model. The first stage estimates (1) without teacher effects, and in the second stage uses the first-stage-estimated year effects in lieu of year effects from equation (1). This approach however, assumes that initial teacher effectiveness does not change within the years of the panel, which may not be the case in DCPS. Figure 3.2 shows average TLF scores for teachers new to DCPS in each year of the panel, and suggests that the effectiveness of new cohorts of teachers has not been static over time. For this reason, we do not rely on the two-stage model for our primary analyses, as these differences in entry ability across cohorts could bias estimates of their returns experience.

Identifying key practices for teachers' overall returns to experience

In addition to understanding the extent to which teachers improve their practice during their early careers, we also explore the heterogeneity of these gains across teachers' entry-level skills and teaching practices to better understand the relative importance of these practices and skills for teachers' improvement in their early careers. To do this, we first estimate returns to experience within sub-skills, identifying entry-level performance (i.e., intercepts) and improvements made with experience (i.e., slopes). We conduct this analysis both graphically and with non-parametric regressions, as with the preceding analyses.

Next, we extend these analyses to better understand variation in entry performance and improvements across teachers. For this analysis, we first divide our restricted sample of 120 teachers with consecutive experience from entry into quartiles based on their initial performance. We determine these performance quartiles using the first two years of all novice teachers' TLF scores. In addition, we break first-year teachers into quartiles based on their location in the overall distribution of first-year teachers' TLF performance (i.e., not just relative to those who persist for five consecutive years) and graph performance trajectories across quartiles; this method provides insight into whether there is variation in teachers' returns to experience. To avoid capturing regression to the mean in lieu of true underlying trends, these quartiles are defined by average performance in teachers' first two years of performance, rather than a single year. Because of non-random attrition among lower-performing teachers, the number of teachers within each quartile in our analytic sample (i.e., novice teachers who are observed in their first five consecutive years of experience) will vary by quartile.

We estimate returns to experience for each quartile, as in equation (1), to better understand the extent of variation in intercepts and slopes of TLF performance across teachers. We then calculate the differences in entry performance for the overall TLF and the nine Teach domains between the top and bottom quartiles in their first and fifth years teaching to identify the size of performance gaps for these teachers as they gain experience in the classroom. Next, we subtract the fifth-year difference from the first-year difference to calculate the gap closure over time. Finally, given that the overall TLF score is equal to an unweighted average of the nine Teach domains, we decompose overall TLF gains into the portions attributable to improvements on each sub-score. This will provide insight into whether certain teaching skills account for a greater share of teachers' overall development, as measured by the TLF, than others.

The Relationship Between Changes in Teaching Practice and Changes in Student Outcomes

Understanding whether and how teacher development influences student outcomes is crucial to employing our analysis to develop hypotheses regarding teacher preparation and professional development. A rigorous assessment of this question depends on long panels of teacher and student data. These data allow us to use only within-teacher variation over time, and thus estimate whether a teacher's changes in skills are associated with changes in her students' achievement.

To estimate this relationship, we use teachers with up to five years of experience from the full analytic sample, fitting regression specifications of the form:

$$\bar{A}_{jt}^{*} = \beta_{jt} + \delta TLF_{jt} + f\left(YrsExp_{jt}\right) + \tau_j + u_{jt} \qquad (6)$$

where $\bar{A}_{jt}^{*}$ is the average outcome for students taught by teacher j in school year t, residualized for a vector of student characteristics (including prior achievement, lagged absences, gender, race/ethnicity, grade level, free- or reduced-price lunch status, special education status, English proficiency, and whether the student changed schools), TLF_{jt} is again teacher j's classroom observation score or sub-score(s) measured in year t, τ_j represents teacher fixed effects, and u_{jt} is a random error term. Our focus is on estimates of δ. $\bar{A}_{jt}^{*}$ is weighted by students' exposure to teacher j, given that some students are assigned to multiple teachers in a given year. While we have a variety of student outcome data, we focus our analyses on those where we expect it would be difficult for teachers to manipulate scores or data, and outcomes for which we are confident we aren't capturing reverse causality. The student outcomes we analyze here are student achievement in math and reading, as measured by standardized test scores.

We also control explicitly for teacher experience, $f\left(YrsExp_{jt}\right)$, which takes the same form as in (2), given that student achievement and teacher skills both co-vary meaningfully with teacher experience. In addition, we are identifying effects based on teachers who may move in and out of tested grades and subjects and therefore some of the teachers in the sample exhibit jumps in experience.[38] We further reduce bias in δ by controlling for many unobserved factors; our within-teacher estimator controls for any factors, including characteristics of teacher j or the students she is assigned, which are fixed over the period we study.

While the above regression is run using only teachers in tested grades and subjects, to ease interpretation of TLF effects, we standardize observation scores to the mean and standard deviation of all teacher-year observations with TLF scores.[39] This allows us to more meaningfully translate effect sizes to that of the larger teaching population in DCPS.

Ideally, we would estimate these effects using our restricted sample of teachers with continuous experience for their first five years in the classroom so that the relationship between student achievement and teachers' practice would not be confounded by teachers' selection out of teaching in DCPS. Because of switching out of tested grades and subjects, however, this sample decreases further when limited to teachers who can be linked to their students' outcomes, leaving us with only two dozen unique teachers who would be able to contribute to this analysis. Instead, we rely on our larger sample of teachers with any level of consecutive experience, removing observations above the fifth year in the classroom. By limiting our analyses to teachers with five or fewer year of experience, we can omit potentially different relationships between these two measures of teacher quality from more-experienced teachers.[40]

[38] We also explore models that omit teachers' experience.

[39] We also explore standardizing these scores to have the same mean and standard deviation teachers in tested grades and subjects for whom we have TLF scores and who can be linked to individual student outcomes. In DCPS, both samples have similar distributions of TLF scores, and so each standardization method yields similar results.

[40] We may, however, still over- or under-estimate the association between improvements on practice and student achievement if the students of teachers who leave before their first five years exhibit different

Considerations across models and research questions

We test the sensitivity of these models for several potential sources of bias beyond those discussed above. The first source of bias is the potential sorting of teachers to certain types of students according to their level of experience (Boyd et al., 2008; Jackson, 2009; Kalogrides et al., 2012). We test for this potential bias in two ways: with the addition of controls for average student characteristics and with school fixed effects.[41]

Estimates of overall experience effects from each of the specifications above are additionally subject to possible bias from endogenous shocks that affect teachers' yearly performance gains in addition to their selection out the district (or out of the profession, which we are unable to distinguish in our data from departure from DCPS). For example, a teacher might experience an illness that causes her to perform lower than she otherwise would have, and following the illness and drop in performance, the teacher chooses to leave DCPS; if teachers who experience larger performance shocks are more likely to attrit, we may observe downwardly biased estimates of overall experience effects. Following Papay and Kraft (2015), we test for this by regressing teachers' departure from

returns to practice than their retained colleagues. This could very well be true if performance on value-added scores compel teachers to leave at different rates than does performance on the TLF; value-added scores give teachers little information about how to improve their performance, while the TLF is much more prescriptive in terms of not simply defining ideal practices, but also giving in some cases prescriptive guidance on how to perform well on the measure (see, for example Adnot, 2016). To test the sensitivity of our results to such a potential source of bias, we define two subsamples to compare results: the first requires at least three years of consecutive experience in a tested grade and subject from the novice year, so as to isolate these changes to the same teachers across annual changes. The second estimates these associations for any teacher up to their third year in the classroom. We use a smaller band of novice experience so as to maximize the sample size while also retaining variation in the second sample in terms of year of attrition. These two sets of results (not shown) produce qualitatively similar results to each other, suggesting that the potential bias from sample attrition is not large.

[41] Students can only be linked to teachers in tested grades and subjects, so we average these student characteristics at the school level.

DCPS on her TLF scores, conditional on prior TLF performance.[42] The coefficient on the TLF scores will be negative if teachers tend to depart following worse-than-expected performance. If we find that the coefficient on lagged experience is significant, it would imply that the timing of departure from DCPS is correlated with a teachers' change in performance. Indeed, as shown in appendix table A.3.1 (panel A), teachers who attrit perform approximately 10 percent of a standard deviation lower on the TLF in their departure year, conditional on prior-year scores, than teachers who remain in DCPS.

This result is inconclusive about the direction of causality, given that teachers may exhibit smaller gains after having decided to leave DCPS, or they may leave DCPS in response to their limited improvements relative to the preceding year. We therefore also estimate attrition effects relative to lagged TLF scores, conditional on twice-lagged scores (panel B). These estimates by definition cannot be conducted for the least-experienced teachers in our sample, but might better capture the relationship between teachers' performance trajectories and their attrition, rather than year-specific shocks to their performance, which could be contributing to the estimates in panel B. Here we find no association between teachers' prior-year improvements and the likelihood of attrition in a given year. Together, these two sets of results suggest that, while teachers who leave may experience performance dips in the year of their departure, they may not have different underlying returns to experience than their peers who stay.

Finally, in order to identify returns to experience for teachers' practice, we must also consider our definition of teachers' initial performance. Teachers receive multiple

[42] We omit from all of our analyses teachers whose IMPACT scores would have led to involuntary dismissal, so the attrition we investigate here represents only voluntary attrition, rather than officially enforced departure due to inadequate performance.

TLF scores throughout a year, so when TLF_{jt} is defined by an average TLF score, it will include some performance gains that are made during the first year. Across the literature on returns to experience in terms of value-added scores, the steepest gains are typically made in the earliest phases of teachers' careers; if this trend can be generalized to within-first-year performance, then any definition of TLF_{jt} that averages scores at the year level will attenuate returns to experience. To test whether this is the case, we explore the sensitivity of estimates to varying definitions of $e = 0$. In primary specifications, starting performance is the average TLF score assigned by MEs in teachers' first year in DCPS, which occurs in the first half of the school year, while the end-of-year TLF score will be equal to the second ME score, which is determined in the second half of the school year.

To estimate a "truer" initial performance level, we define teachers' experience by the observation cycle in which an evaluation occurs instead of by year. This approach may reduce bias in estimates of returns to experience, but will also reduce the precision of estimates, given that reliability for any single observation score will typically be lower than an average of multiple observation scores. Conversely, this approach could increase bias if certain types of teachers are more likely to be evaluated in only one window of the year; indeed, as teachers advance through DCPS's career ladder—which is determined by patterns of performance—they are subject to fewer evaluations each year.[43] We test for such observation-cycle attrition bias using the same technique we use to estimate bias from district-level attrition, regressing teachers' occurrence of not receiving a TLF score

[43] More information about this career ladder, the Leadership Initiative for Teachers (LIFT), is available on the DCPS website at https://dcps.dc.gov/page/leadership-initiative-teachers-lift.

in a next observation window on her current TLF score, conditional on prior TLF performance.

Here, we find similar associations between attrition and current-year changes in TLF as with year-level attrition (appendix table A.3.1, panel C); teachers who do not receive evaluation scores in a subsequent observation cycle demonstrate lower scores in the current cycle than their colleagues, conditional on their experience and prior-cycle TLF scores. When estimated using prior-cycle TLF scores as the outcome variable and controlling for twice-lagged cycle evaluation scores, these effects reverse direction; teachers who do not receive scores in a given window have seen larger performance gains in the preceding observation windows. This is consistent with the decreased observation requirements laid out by DCPS's career ladder; we would expect teachers to be subject to fewer evaluations as their performance improves. This evidence suggests, however, that observation-window-level estimates of returns to experience may understate teachers' returns to experience, as teachers with greater improvements are not observed continuously across cycles in DCPS.

RESULTS

Establishing Patterns of Skill Development in Teachers' Early Careers

Overall TLF Improvement

Figure 3.3 illustrates returns to experience on the TLF for our restricted sample—those teachers who we observe with complete TLF scores for five consecutive years. Using scores assigned by external evaluators, who we believe may be less subject to evaluation bias and ceiling effects, we find that within this sample of teachers, the average educator improves to 0.91 standard deviations above the first-year average by her fifth year in the classroom, equivalent to close to a half of point on the TLF. These gains are large, but due to the small sample size are not statistically different from zero beyond the first two years. Using a censored growth model to estimate returns to experience with the larger sample of teachers (i.e., teachers who are observed for any length of consecutive experience in DCPS), we find similar overall improvement trajectories. A model that controls only for teacher and year fixed effects yields an estimated improvement of 0.885 standard deviations (see table 3.2), an effect that is robust to the inclusion of school-averaged student characteristics (0.864 standard deviations, also plotted in figure 3.4) and school fixed effects (0.875 standard deviations).

Estimates are additionally robust to other modeling approaches, shown in appendix table A.3.3. Point estimates are at least as large when the GGM model is adjusted for equal teacher weights as described in footnote 4 (top left panel of appendix table A.3.3) and when estimated using all observers' evaluation scores (appendix table A.3.4); estimates are similarly large across different censoring levels (appendix figure A.3.1), and are likewise substantial when estimated using a discontinuous career approach and a two-stage approach (top right and bottom left panels, respectively). While still producing large estimates, the indicator variable model yields meaningfully smaller effects (roughly a half a standard deviation, bottom right panel of appendix table A.3.3); we would expect this approach to return smaller point estimates, given that the model cannot account for changes in experience effects within early-career bins. However, even these likely downwardly-biased results support the conclusion that teachers make meaningful improvements on their practice and skills over their early careers.

The steep returns to experience for teachers practice that we demonstrate here indicate that teachers' practice improves over their early careers in a pattern that is similar to what we observe for teachers' effects on student achievement—with large overall gains in their early careers, which are most heavily concentrated in teachers' first couple of years in the classroom. These sizeable returns to experience suggest that teachers' practice is on average highly malleable at the start of their careers and implies that targeted interventions and supports might be able to shift teachers' practice upward earlier in their careers, if not before their formal entry into the profession.

Meanwhile, all of the approaches we rely on for our analyses may attenuate returns to experience somewhat, given that we define teachers' initial performance as their average scores across their first year in the classroom. Given that we have established that there are meaningfully large experience effects across the first few years of teachers' practice, it stands to reason that there may also be large returns to experience *within* year, and particularly so for teachers' earliest years in the classroom. We therefore estimate our censored growth model using teachers' initial observation window, in the fall of their first year, as the starting point, and replace the years in our experience function, $f(YrsExp_{jtw})$, with year-by-observation-window indicators. Indeed, this approach, estimated on the unrestricted sample, demonstrates slightly larger returns to experience (0.933 standard deviations) than the same method using year-averaged TLF scores (appendix figure A.3.2).

Together, these overall TLF trajectories indicate large returns to experience for early-career teachers, providing a fuller picture of the ways in which teachers improve beyond the evidence generated from within-teacher effects on student achievement. Even the more-conservative estimates that we derive from our primary sampling and modeling specifications demonstrate that early-career teachers are improving substantially upon the expectations for teaching and learning that are laid out by the TLF. Next, we explore the role that individual practices play in teachers' overall early-career improvements to better understand how different skills contribute to these within-teacher performance trajectories.

Early-career improvements across sub-skills

Figure 3.5 displays teachers' early-career improvement on the two TLF domains identified in our factor analysis for the restricted sample of 120 novice teachers with complete score data and at least five continuous years of experience. These teachers start with higher initial performance on their *classroom environment and lesson accessibility* skills than on *instructional clarity and student understanding*. While imprecisely estimated, teachers in this sample improve their *classroom environment and lesson accessibility* by a half of a TLF point (0.501), scaled to the distribution of novice educators' overall, unstandardized scores, or a full (1.014) standard deviation of novice teachers' TLF scores, while making smaller and similarly non-significant gains to *instructional clarity and student understanding* (0.15 points when scaled to the distribution of novice teachers' overall scores, or 0.308 standard deviations); point estimates are provided in appendix table A.3.5.

When estimated on the larger sample of DCPS educators, the difference in effects for the two core TLF domains persist, though to a lesser degree, with an estimated five-year return-to-experience of 0.734 standard deviations of the overall TLF for *classroom environment and lesson accessibility* and just over a half of a standard deviation (0.557) for *instructional clarity and student understanding*. For both samples, tests of the equality of the two domains' returns to experience indicate that teachers make different improvements across the two subskills and at each level of experience ($p < 0.001$). Importantly, teachers not only make smaller gains to *instructional clarity and student understanding* than *classroom environment and lesson accessibility*, but start lower on this practice as well. This suggests that *instructional clarity and student understanding* is more difficult but may also be less malleable than *classroom environment and lesson accessibility*.

Across the nine Teach domains defined by the TLF, we likewise find that teachers' entry skills and their performance trajectories vary across subskills. As shown in appendix figure A.3.3 and appendix table A.3.5, the novice teachers in our sample make far steeper improvements on Teach 2 and Teach 6 (*explain content clearly* and *respond to student misunderstandings*, respectively) than they do on Teach 1 (*lead well-organized, objective-driven lessons*) and Teach 7 (*develop higher-level understanding through effective questioning*). Meanwhile, they enter with relatively high performance on Teach 8 and 9 (*maximize instructional time* and *build a supportive, learning-focused environment*, respectively) and lower performance on Teach 2 and Teach 7.[44] These results demonstrate that teachers in DCPS are not necessarily making the largest gains on the skills where they enter with the lowest scores, nor are these teachers on average unable to continue to improve meaningfully on the skills where they enter with already

[44] Given that a sizeable number of DCPS teachers are missing Teach 6 scores, as described on pages 13 and 14, we also plot these trends—less Teach 6—for teachers who have otherwise complete scores (see appendix figure A.3.4); though these teachers exhibit slightly larger overall gains, on average, than those in our primary, restricted sample, the patterns in terms of relative slopes and intercepts for each of the sub-scores across the two samples are largely the same.

high achievement. This suggests that even high-performing teachers in some areas may to continue to improve on certain subskills, while low-performing teachers in others may have limited room for growth on other subskills; these performance patterns indicate that each of these subskills is not equally malleable and may not yield similar returns to professional development investments.

Identifying key practices for teachers' overall returns to experience

The overall returns to experience that we document in figure 3.1 may mask heterogeneity not simply across individual skills, as we explored above, but also heterogeneity across teachers. We explore this question by first sorting first-year teachers into quartiles based on their location in the overall distribution of first-year teachers' TLF performance (i.e., not just relative to those who persist for five consecutive years) and plotting performance trajectories across quartiles; this method provides insight into the extent to which there is variation in teachers' overall returns to experience. To avoid capturing regression to the mean in lieu of true underlying trends, we define our quartiles by averaging performance in teachers' first two years in the classroom, rather than a single year. Figure 3.6 shows the returns to experience for each quartile of teachers in our restricted sample. The lowest quartile of teachers enters DCPS with TLF scores more than a full point—roughly 2.3 standard deviations of novice-year performance ($p < 0.001$)—below their highest-scoring colleagues. These bottom-quartile teachers make substantial improvements over their first five years; however, their peers in the top quartile on average are performing only somewhat better by their fifth year of teaching than they were in their first year (0.19 TLF points, or 0.49 standard deviations, $p < 0.05$). Table 3.3 documents the overall differences between top and bottom quartiles, for teachers in their first and fifth year teaching. The first row documents these differences for overall TLF scores, for which the lowest-quartile of teachers have closed 83% of their performance gap relative to their peers with higher entering performance.

We also estimate performance gaps for each sub-domain of the TLF, which we then use to calculate the contribution of bottom-quartile teachers' relative improvements on each sub-skill to overall TLF gap closures. Within the two core domains identified by our factor analysis, top and bottom-quartile teachers score more than a standard deviation apart in their initial year teaching, yet by their fifth year, there is no statistical difference on the first factor (*classroom environment and lesson accessibility*) across these quartiles by the time teachers enter their fifth year. Bottom-quartile teachers have effectively improved enough relative to their higher-performing peers on this skill to have made up more than 90 percent of the difference in performance; this gap closure is attributable entirely to bottom-quartile teachers catching up to their higher-performing colleagues, rather than both groups making improvements, albeit at different rates. Meanwhile, when it comes to the second factor (*instructional clarity and student understanding*), while smaller than in their first year of teaching, a substantial performance gap remains across top- and bottom-performing teachers five years into their careers (0.76 standard deviations, $p<0.05$); however, in this case both groups of teachers have managed to improve. Given that performance gaps remain for this domain of teaching, and it is a secondary factor in our analysis (see appendix table A.3.1), improvements on this factor account for a smaller share of overall TLF performance gap reduction (40%) than the first factor (60%).

We conduct the same analysis across the nine teaching domains explicitly defined in the TLF, each of which exhibit large differences in entry-year performance between top- and bottom-quartile teachers. Across most of these practices, the performance gaps have closed by at least 76% by teachers' fifth year in the classroom. Teach 6 (*respond to student misunderstandings*), however, stands out as having a still-substantial performance gap in year 5. While the entry gap for this subskill was not among the largest (at 1.99 standard deviations, $p < 0.001$), the year-five gap demonstrates a reduction of only 58%. The relative lack of performance gap closure is attributable to substantive improvements made by both quartiles of teachers, with a relatively higher share of overall gains coming from high-performing teachers than we observe with other Teach domains. Bottom-quartile teachers' average scores increase by 1.79 standard deviations ($p < 0.001$) and top-quartile teachers' score improve on average by 0.63 standard deviations ($p < 0.05$). This suggests that the higher slopes that we observe for Teach 6 in appendix figure A.3.3 and table A.3.5 are due to gains made across novice teachers, rather than being driven only or predominantly by initially-low-performing educators.

These performance trajectories provide potential insights into ways to differentially target feedback and supports to teachers across performance levels. Given that there are some practices (e.g., Teach 6) where both high- and low-skill novice teachers make meaningful improvements over time, supports on such practices could be widely implemented across schools and districts in order to effect meaningful changes in early-career teachers' practices, as could training for pre-service teachers. Meanwhile, other areas may demonstrate diminishing returns as teachers improve on those practices (e.g., the practices associated with *classroom environment and lesson accessibility*); for these practices, supports would be most-efficiently delivered to only relatively low-skilled teachers. Importantly, however, across all of the domains of the TLF, teachers on average make statistically and substantively meaningful gains over their first five years in the classroom, indicating that early-career teachers' practice is malleable across many domains.

Another important implication of the varying improvement trajectories that we observe across practices and subgroups of teachers is that there is sufficient variation in teacher skill development to assess how differences in growth of teaching skills relate to changes in student outcomes. While the existing literature has extensively documented that teachers become considerably better at improving student achievement as they advance through their careers, and the preceding analyses demonstrate similar trends for teachers' practice, it is not a given that changes in practice are necessarily accompanied by changes in student outcomes. We turn to this question next.

The Association Between Teachers' Skill Development and Their Students' Achievement Gains

Table 3.4 shows the results from our regression model estimating the association between teachers' practice and their students' outcomes. Given that the model estimates this relationship within teacher, each coefficient can be interpreted as the effect of a standard deviation improvement on the TLF on changes in student outcomes. Overall, the directionality of the coefficients is largely what we would expect, though imprecisely estimated. Within teacher, higher TLF scores are associated with greater student learning and fewer student absences in the first five years of novice educators' experience.

When estimated without experience or school effects, an improvement on the TLF of one standard deviation is associated with a 3.7% standard deviation increase in math achievement, though point estimates become smaller and less precisely estimated as additional controls are added. Among the two main factors of the TLF, only the second factor (*instructional clarity and student understanding*) demonstrates a significant association with changes in math achievement; a within-teacher standard deviation increase in TLF scores on this factor is associated with an increase of 3.3 to 4 percent of a standard deviation of average student achievement in math.

In reading, overall TLF effects are small and imprecise, but across the two core teaching domains identified by our factor analysis, we observe similar effects to those in math. While effects are null for the model with just teacher effects, when estimated with experience controls and school fixed effects, a standard deviation gain in a teacher's *instructional clarity and student understanding* is associated with an average student gain in reading scores of approximately 4% ($p < 0.05$).

Across both subjects, these effects appear to be driven largely by Teach 1 (*lead well-organized, objective-driven lessons*), the only sub-score for which coefficients are consistently statistically different from zero (see appendix table A.3.6), at roughly 4 percent of a standard deviation for both subjects. In math, Teach 7 (*develop higher-level understanding through effective questioning*) is an additional a skill where improvements are associated with higher student achievement—approximately 3 percent of a standard deviation, across specifications.

In spite of our low statistical power and potentially attenuated results, the effects we find are still meaningfully large. If we were to assume that the average effect of a standard deviation improvement on the TLF across the first five years of teaching on test scores is—as we estimate here—3.7 percent of a standard deviation gain in student math achievement (i.e., if we were to extrapolate these results to the sample we use to estimate overall returns to experience), and that the average TLF trajectories of the teachers from which we identify our math-outcomes regression estimates mirror those of their peers in non-tested grades and subjects, the overall five-year increase on TLF scores of 0.864 standard deviations would be associated with student learning gains of approximately 3.2 percent of a standard deviation; at these grade levels, this is roughly equivalent to three additional weeks of learning in math (Hill, Bloom, Black, & Lipsey, 2008).

We know, however, that there is heterogeneity in terms of teachers' returns to experience on the TLF. The lowest quartile of entering teachers, for example, improve by more than two (2.3) standard deviations of novice DCPS teachers' TLF scores, relative to 0.49 standard deviations of improvement made by the top quartile. This suggests that the gains made in the first five years by the lowest-performing entering teachers may yield as large as an 8.5 percent of a standard deviation learning gain in math—approximately equal to two additional months of schooling (Hill et al., 2008).

These data also are striking in juxtaposition with the skill-specific returns to experience that we identify, as the practices for which we are sufficiently powered to find associations between teachers' improvements and their students' learning outcomes are the practices where teachers enter with relatively low performance, on average, and demonstrate the lowest relative gains. These skills may be more difficult for teachers to develop, but our results suggest that they are more strongly associated with student achievement gains, and may therefore be areas to focus professional development for in-service teachers or teacher preparation curricula for pre-service teachers.

DISCUSSION AND CONCLUSION

This paper begins to build our knowledge of practice-based returns to experience, but introduces new questions about the nature of these improvements that we cannot currently answer with the limited size of our panel. For example, we have established that there are large returns to experience for teachers on average, and that there is heterogeneity in these returns, but we do not have sufficient statistical power with our DCPS data to improve our understanding of this heterogeneity. The literature suggests that school contexts are important for teachers' development (Kraft & Papay, 2014; Johnson, Kraft, & Papay, 2012); an analysis that decomposes the importance of school-specific relative to general experience for the development of teaching skills, for example, would help identify the variation in teachers' improvement that is attributable to the contexts in which they are teaching.

There may, additionally, be important district-level differences that affect teachers' performance as they transition to their school districts, which we cannot observe in an analysis of a single district such as DCPS. The specific policies unique to a given district may also affect teachers' entry performance and development over time, as these policies determine, among other things, the peers with which a teacher works and the professional development a teacher receives. Indeed, the many analyses of returns to experience for student achievement suggest wide variation, depending on the population being studied (Atteberry et al., 2015; Papay & Laski, 2018). District-specific experience is likely to be particularly important in DCPS, given that IMPACT is a uniquely high-stakes system; teachers entering DCPS with prior experience are unlikely to have come from similar districts. DCPS teachers, for example, face large incentives to perform well and improve from year to year, and they receive multiple formal observations followed by written and in-person feedback sessions in which they and their evaluators discuss next steps for their professional growth.

Additionally, DCPS's teacher evaluation system is uniquely high stakes relative to other districts nationally, leading to differential retention of low-performing teachers, as well as performance gains for those who remain (Dee & Wyckoff, 2015; Chapter 2 of this dissertation). The incentive to perform well on IMPACT, along with the feedback that is a formal part of the program, may be strong enough to induce additional returns to experience beyond those a teacher would have exhibited in another district. Indeed, Adnot (2016) finds that incentivized teachers—including those performing at IMPACT thresholds associated with risk of dismissal as well as those near the threshold for substantial financial rewards—make larger gains to their practice than teachers who do not face such incentives to improve. This suggests that teachers in DCPS may demonstrate larger returns to experience on their practice than teachers in other districts; on the other hand, the types of teachers who select into a high-stakes context such as IMPACT may have different potential performance trajectories than those who choose to teach in other settings.

Meanwhile, low-performing teachers in DCPS attrit at higher rates than might be observed in otherwise comparable districts, given that IMPACT explicitly removes teachers who fail to perform above a given level or make sufficient gains over time. In addition, teachers who perform below key thresholds voluntarily attrit at higher rates than their higher-performing colleagues (Dee & Wyckoff, 2015; Chapter 2 of this dissertation). This high attrition of low-performers under IMPACT might produce different returns to experience than would be observed under more-typical teacher evaluation policies. While the lowest-performing (i.e., Ineffective) teachers might make substantial gains to their performance were they to remain teaching, these account for a small share of the overall attrition in DCPS. That IMPACT includes as a condition for dismissal failure to attain an Effective rating over time, however, means that teachers with lower returns to experience are inherently omitted from these analyses, as they are no longer teaching in the district. While the similarity of results from the restricted sample of non-attriting novice teachers to results from the full sample of DCPS teachers suggests that this attrition bias may not be large, these returns to experience may still be steeper than they would be in the absence of IMPACT. Future research building upon the findings we show here with additional districts' evaluation data can demonstrate whether the large returns to experience for practice that we observe in DCPS are representative of teachers' early-career improvements more broadly, or if they represent the potential of rigorous evaluation policies for facilitating teachers' development.

Regardless, our findings indicate that teachers make meaningful gains to their overall practice in their early years. These estimates are consistent with early evidence from Tennessee (Papay & Laski, 2018), with a correspondingly steep slope to the performance trajectories documented in the value-added literature. Even our more conservative estimates suggest that teachers improve their practice by at least 80 percent of a standard deviation in their first five years. These large average gains to experience likewise provide evidence of the malleability of teaching skills over teachers' early careers. It shows that trajectories are large in terms of teachers' overall practice, but that gains are not equally large for all of the sub-skills measured by the TLF, nor are they equal across teachers.

We find, for example, that the lowest-performing teachers at entry make larger overall gains to their practice than their higher-performing peers. While there may be some ceiling effects that prohibit larger gains for top-quartile teachers, it is likely a combination of factors that causes their limited growth, given that few teachers in DCPS ever receive perfect scores from external evaluators, as documented in figure 3.1. Rather, this could in part reflect a disproportionate urgency to improve for lower-performing teachers, or nonlinearity in the feasibility of making improvements on these skills past a certain performance threshold. Unlike analyses of the heterogeneity of student-achievement improvements across teachers by initial performance (e.g., Atteberry et al., 2015; Xu et al., 2015), by their fifth year in the classroom the bottom-quartile teachers are performing nearly as well as their highest-skilled peers, and better on average than the typical first-year teacher in DCPS.

By decomposing the variance in gains across these quartiles of initial performance, we reveal some nuance to these skill developments. Importantly, none of the gaps that remain are due to a lack of development on the part of the lowest-quartile teachers; rather, the areas where gaps persist are generally skills and practices where teachers across the skill distribution have made some degree of improvement. The skills where performance gaps have disappeared after a few years are those where the top-quartile teachers have made few gains, but bottom quartile teachers have improved substantially. Across these skills, we see evidence of malleability, albeit to different degrees depending on the teachers' skill level.

Importantly, in spite of power limitations, we also identify a consequential association between gains in practice and gains in student achievement. This suggests that interventions to improve practice may have meaningful effects for student outcomes, and this relationship provides support for the hypothesis that improvements in measured TLF observation scores reflect substantive gains in teaching skills that are associated with improvements in student outcomes—particularly in math. These student achievement gains are likewise consistent with the magnitude of the student-achievement-level returns to experience that have been observed in other studies, where by the end of a teacher's first five years in the classroom her students are scoring roughly five to 10 percent of a standard deviation higher than in her novice year, depending on the subject area, population, and model specification (see figure 1 in Atteberry et al. for a compilation of these studies). In our analyses, we find that a standard deviation of improvement on the TLF is associated with approximately 4 percent of a standard deviation gain in student achievement; given that teachers improve at least 80 percent of a standard deviation on the TLF in their first five years in the classroom, our estimates would extrapolate to a magnitude of expected returns to student achievement over the same period associated with teachers' improved practice of roughly three percent of a standard deviation—a significant rate of improvement relative to prior studies of teachers' returns to experience.

Together, these findings illustrate that measures of teachers' practice can be instrumental for understanding teachers' development, and may provide insights into levers for improving student achievement. The finding that early-career teachers make large gains on classroom observations—which are by design both formative and summative measures—can facilitate informed policy and programmatic decisions in a way that value-added scores, as purely summative measures, cannot. Identifying malleable skills using the TLF and similar measures can provide policymakers, school districts, and other stakeholders involved in teachers' preparation and development with essential information as to where to target training so that teachers not only more-rapidly improve once they are in the classroom, but also enter the profession with stronger skills. Because novice teachers are more likely to be teaching the least-advantaged students, the implications of teachers' reaching their performance potential earlier in their careers are significant.

TABLE 3.1

The DCPS Teaching and Learning Framework (TLF) Components

Teach 1: Lead well-organized, objective–driven lessons
Teach 2: Explain content clearly
Teach 3: Engage students at all learning levels in accessible and challenging work
Teach 4: Provide students multiple ways to move toward mastery
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 3.2

Estimates of Returns to Experience from Censored Growth Models

		(1)	(2)	(3)
Experience	1	0.403 ***	0.400 ***	0.395 ***
		(0.039)	(0.039)	(0.039)
	2	0.599 ***	0.586 ***	0.586 ***
		(0.054)	(0.054)	(0.055)
	3	0.713 ***	0.699 ***	0.696 ***
		(0.071)	(0.071)	(0.072)
	4	0.885 ***	0.864 ***	0.875 ***
		(0.088)	(0.088)	(0.088)
Teacher FE		X	X	X
Year FE		X	X	X
School-level student characteristics			X	X
School FE				X
Sample Size		10,399	10,399	10,399

Notes. Data are teacher-by-year, using master-educator-assigned evaluation scores only. Units are standardized such that each coefficient estimate represents SD gains on the TLF relative to experience=0. Models are censored at *E*=15. Robust standard errors are in parentheses.

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; and $^{+}$ $p < 0.10$

TABLE 3.3

Gaps in Teaching Quality, by Quartile of Entry Performance, Across the First Five Years of Experience

	Difference between top & bottom quartiles				Year 1 Difference minus Year 5 Difference		Percent of Performance Gap Reduced	Sub-score Contribution to Gap Reduction
	Year 1		Year 5					
	TLF Score	SD Units	TLF Score	SD Units	TLF Score	SD Units		
Total TLF	1.13	2.41 ***	0.19	0.41 *	0.94	2.00 ***	83%	--
Environment	1.05	2.23 ***	0.06	0.13	0.98	2.10 ***	94%	60%
Clarity	1.25	2.66 ***	0.37	0.80 *	0.87	1.86 ***	70%	40%
Teach 1	1.23	2.62 ***	0.29	0.62 *	0.94	2.00 ***	76%	0.11
Teach 2	0.95	2.03 ***	0.22	0.47	0.73	1.57 ***	77%	0.09
Teach 3	1.25	2.67 ***	-0.03	-0.06	1.28	2.72 ***	102%	0.15
Teach 4	1.51	3.21 ***	0.28	0.60 *	1.22	2.61 ***	81%	0.14
Teach 5	1.16	2.48 ***	0.27	0.57 $^{+}$	0.90	1.91 ***	77%	0.11
Teach 6	0.99	2.10 ***	0.41	0.88 *	0.57	1.22 *	58%	0.07
Teach 7	1.08	2.31 ***	0.04	0.09	1.04	2.22 ***	96%	0.12
Teach 8	1.20	2.56 ***	0.05	0.11	1.15	2.45 ***	96%	0.14
Teach 9	0.80	1.70 ***	0.19	0.40	0.61	1.30 **	76%	0.07

Notes. Quintiles are defined across the average TLF scores for the first two years of teaching for all teachers entering the profession in AY2009-10, AY2010-11, and AY2011-12. Data are teacher-by-year, using master-educator-assigned evaluation scores only, from a restricted sample of 120 teachers who entered DCPS with no prior experience, continued teaching for at least five consecutive years between 2009-10 and 2015-16, have scores in each year from master educators and school administrators, and were not involuntarily separated due to their IMPACT scores. Standard deviations are estimated off of all first-year teachers' ME-assigned, overall TLF scores; the standard deviation of novice teachers' ME scores in this period is 0.494.

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; and $^{+}p < 0.10$

TABLE 3.4

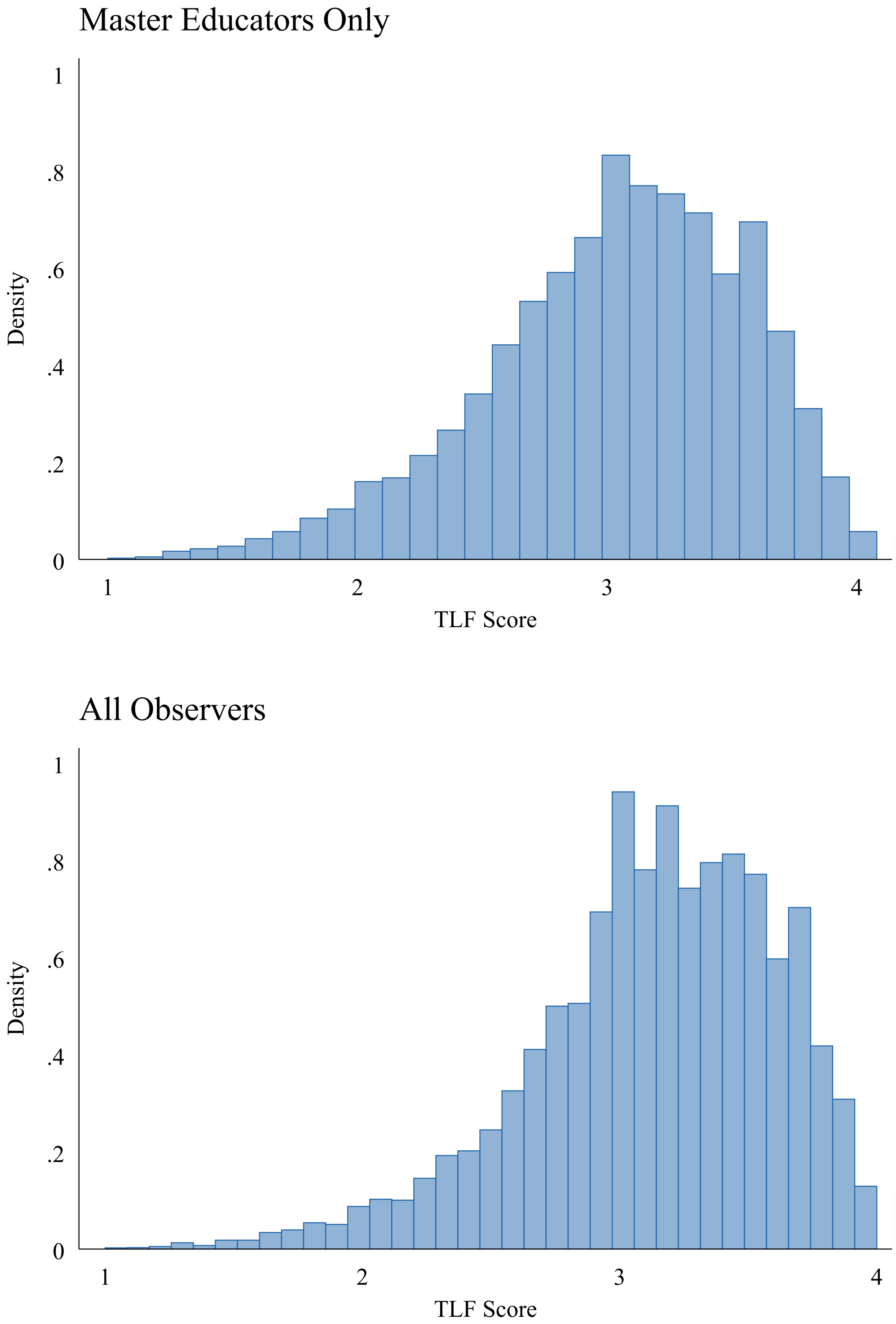
The Relationship Between Student Achievement and Changes in Teachers' Practice

	Math			Reading		
	(1)	(2)	(3)	(1)	(2)	(3)
TLF Overall	0.037 *	0.031	0.026	0.002	0.013	0.012
	(0.018)	(0.020)	(0.021)	(0.019)	(0.021)	(0.022)
	{0.040}	{0.134}	{0.225}	{0.925}	{0.529}	{0.596}
Factor 1: *Classroom Environment & Lesson Accessibility*	0.013	0.008	0.005	-0.015	-0.011	-0.013
	(0.018)	(0.019)	(0.020)	(0.019)	(0.020)	(0.021)
	{0.484}	{0.668}	{0.788}	{0.449}	{0.588}	{0.526}
Factor 2: *Instructional Clarity & Student Understanding*	0.040 *	0.036 *	0.033 $^{+}$	0.020	0.037 **	0.038 *
	(0.016)	(0.018)	(0.019)	(0.014)	(0.014)	(0.015)
	{0.012}	{0.047}	{0.082}	{0.135}	{0.010}	{0.010}
Teacher FE	X	X	X	X	X	X
Experience		X	X		X	X
School FE			X			X
n	664	664	664	662	662	662

Notes. Outcomes are averaged at the teacher level, after residualizing using a vector of student characteristics including race/ethnicity, gender, lagged absences and achievement, poverty status, special education status, grade level, and indicators for limited English proficiency and whether the student is in a new school. TLF scores are standardized relative to the distribution of all DCPS teachers' overall master-educator (ME)-assigned TLF scores. Point estimates for each subdomain are from separate regressions. Robust standard errors are in parentheses; *p*-values are in brackets.

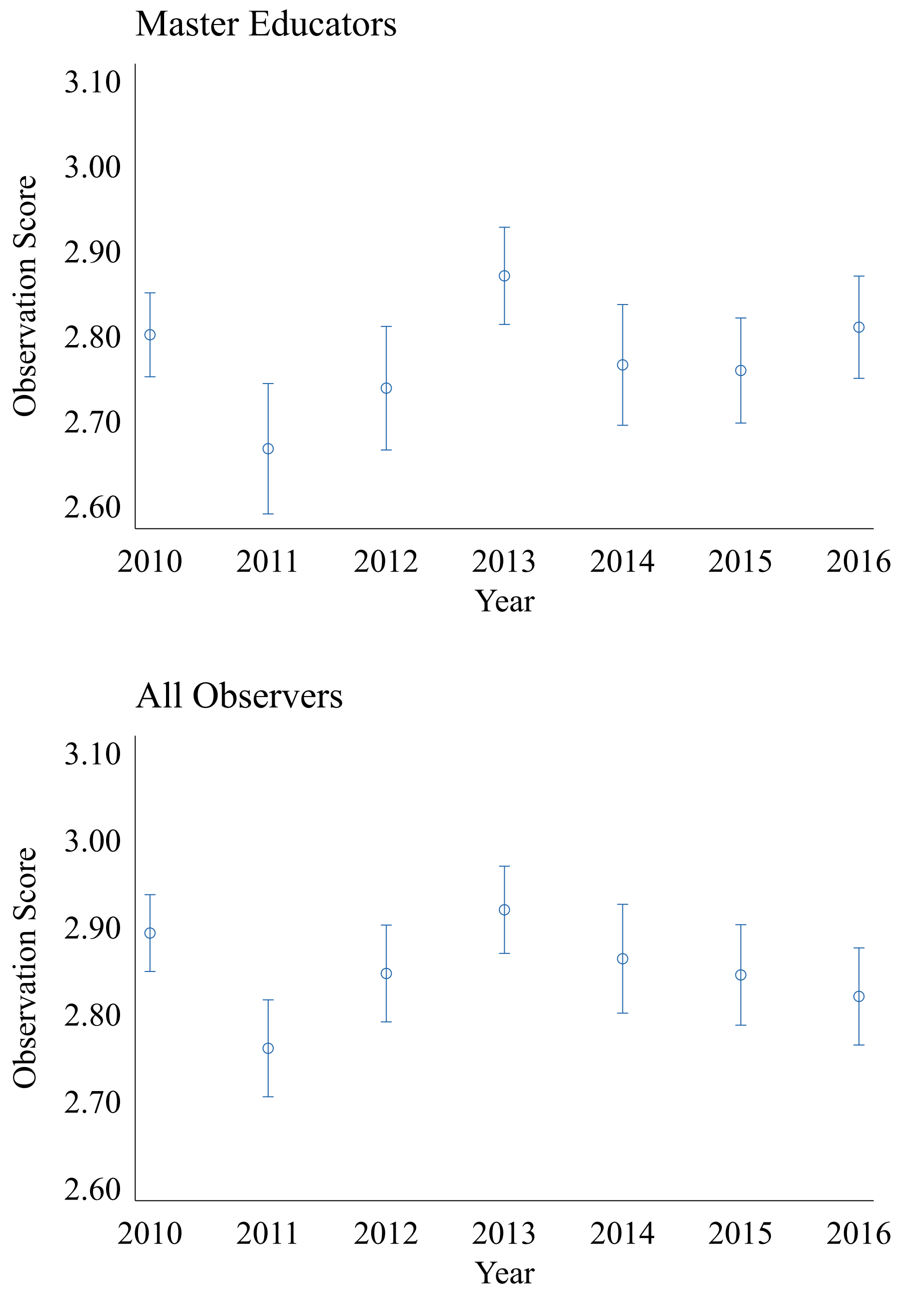
$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; and $^{+}p < 0.10$

FIGURE 3.1. *Distribution of Teacher Observation Scores in DCPS*



Note. Histograms of teacher-by-year TLF scores for all DCPS teachers from AY 2009-10 through AY 2015-16 with scores on each of the nine Teach domains of the TLF.

FIGURE 3.2. *Observation Ratings for Novice Teachers in DCPS*

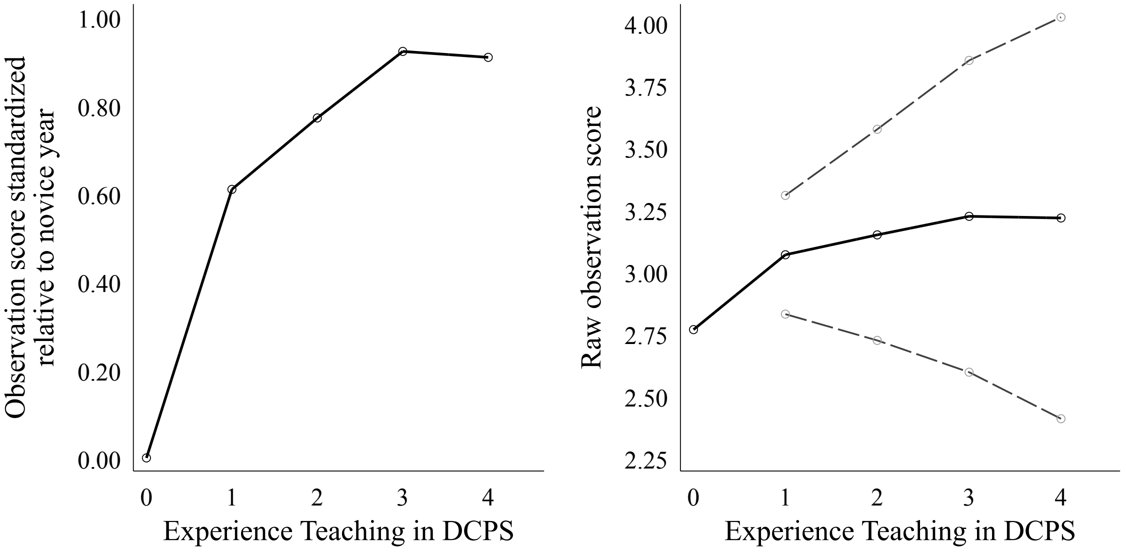


Master Educators

All Observers

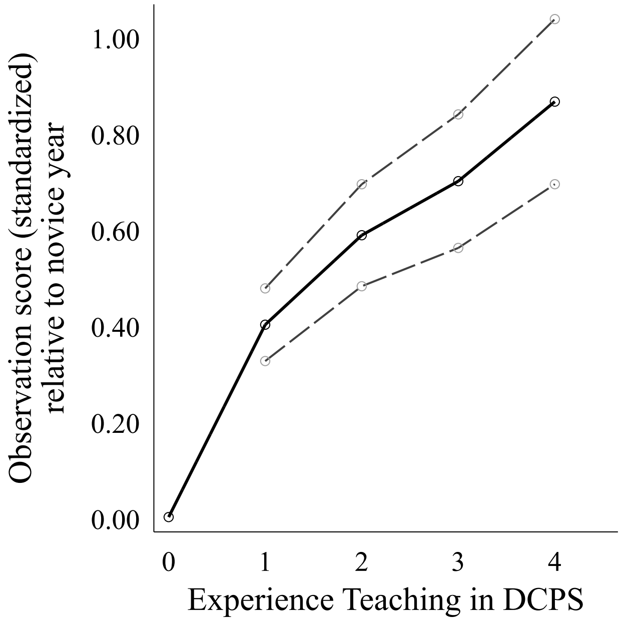
Note. Average TLF scores for teachers in DCPS between AY 2009-10 and AY 2015-16 with scores on each of the nine Teach domains of the TLF who have no prior teaching experience.

FIGURE 3.3. *Returns to Experience for Novice Teachers' Practice, Primary Analytic Sample*



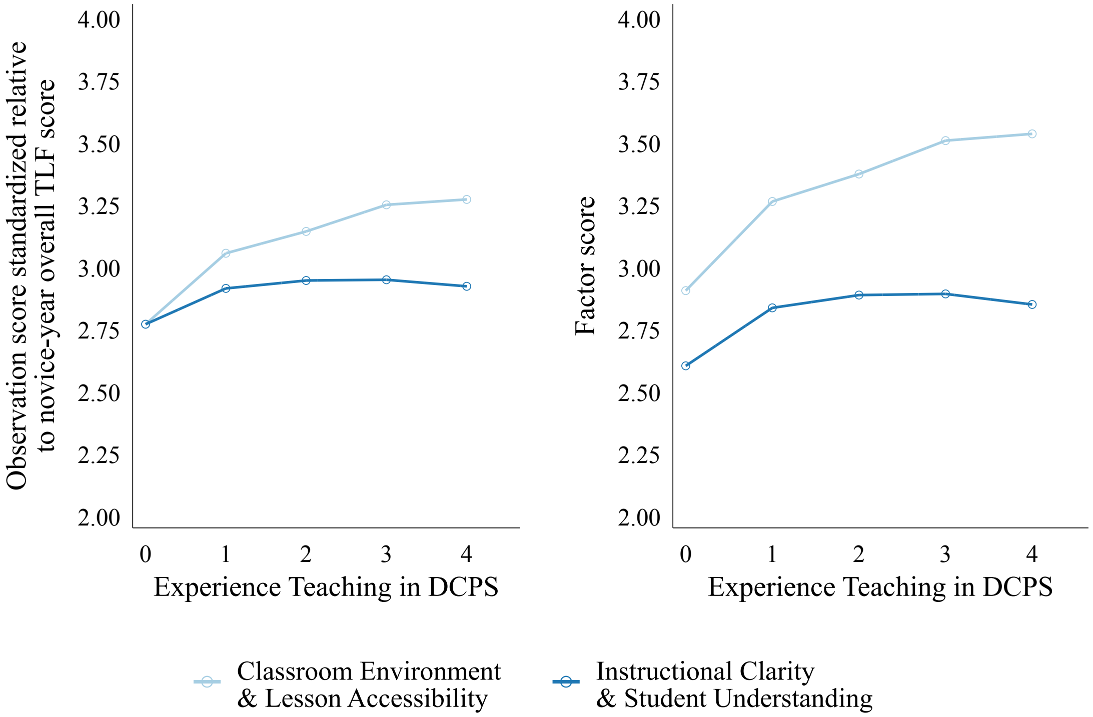
Notes. Point estimates first five years obtained from fitting Equation 1 with $E = 5$ using a restricted sample of 120 teachers who entered DCPS with no prior experience, continued teaching for at least five consecutive years between 2009-10 and 2015-16, have scores in each year from master educators and school administrators, and were not involuntarily separated due to their IMPACT scores. Regressions include teacher and year fixed effects, as well as school-averaged student characteristics. The dependent variable is the average teacher observation score; in the graph on the left, scores are scaled to the distribution of novice teachers' overall ME-assigned TLF scores; the right shows these scores in their raw (i.e., unstandardized) form.

FIGURE 3.4. *Returns to Experience for Novice Teachers' Practice, Full Analytic Sample*



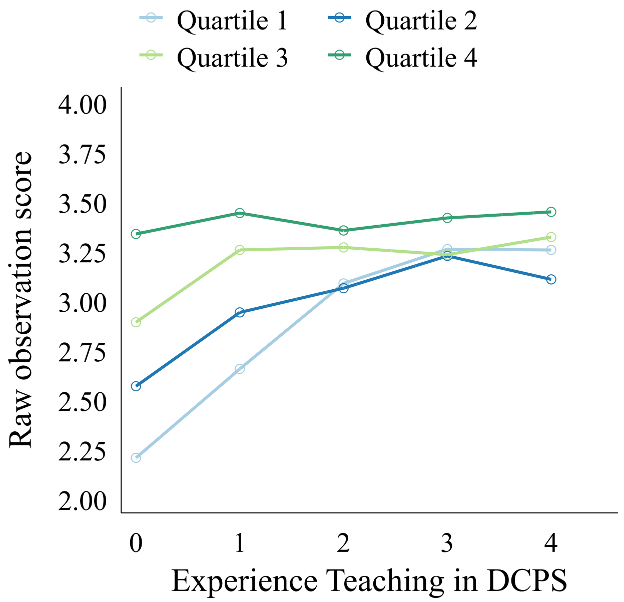
Notes. Point estimates and 95% confidence intervals for first five years obtained from fitting Equation 1 with $E = 15$ using the sample of 3,407 teachers with at least two years of continuous experience in DCPS between 2009-10 and 2015-16 who have scores in each year from master educators and school administrators and were not involuntarily separated due to their IMPACT scores. Regressions include teacher and year fixed effects, as well as school-averaged student characteristics. The dependent variable is the average teacher observation score assigned by master educators (MEs); scores are scaled to the distribution of novice teachers' overall ME-assigned TLF scores.

FIGURE 3.5. *Returns to Experience for Teachers' Performance on* Classroom Environment *and* Instructional Clarity *domains*



Notes. Point estimates first five years obtained from fitting Equation 1 with $E = 5$ using a restricted sample of 120 teachers who entered DCPS with no prior experience, continued teaching for at least five consecutive years between 2009-10 and 2015-16, have scores in each year from master educators and school administrators, and were not involuntarily separated due to their IMPACT scores. Regressions include teacher and year fixed effects, as well as school-averaged student characteristics. The dependent variable is the average teacher observation score; in the graph on the left, scores are scaled to the distribution of novice teachers' overall ME-assigned TLF scores; the right shows these scores in their raw (i.e., unstandardized) form. The *classroom environment* and *instructional clarity* domains are derived from a principal-components factor analysis (see appendix table A.3.1.)

FIGURE 3.6. *Returns to Experience by Quartile of First-Year Performance on the TLF*



Notes. Point estimates first five years obtained from fitting Equation 1 with $E = 5$ using a restricted sample of 120 teachers who entered DCPS with no prior experience, continued teaching for at least five consecutive years between 2009-10 and 2015-16, have scores in each year from master educators and school administrators, and were not involuntarily separated due to their IMPACT scores. Regressions include teacher and year fixed effects, as well as school-averaged student characteristics. The dependent variable is the average teacher observation score assigned by master educators (MEs).

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.

Adnot, M. K. (2016). *Effects of incentives and feedback on instructional practice: Evidence from the District of Columbia Public Schools' IMPACT teacher evaluation system.* (Doctoral Dissertation, University of Virginia).

Adnot, M., Dee, T., Katz, V., & Wyckoff, J. (2017). Teacher turnover, teacher quality, and student achievement in DCPS. *Educational Evaluation and Policy Analysis, 39*(1), 54-76.

Ajayi, L. (2016). High school teachers' perspectives on the English language arts Common Core State Standards: An exploratory study. *Educational Research for Policy and Practice, 15*(1) 1-25.

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

Apperson, J., Bueno, C., Sass, T. (2016). Do the cheated ever prosper? The long-run effects of test-score manipulation by teachers on student outcomes. CALDER Working Paper. https://caldercenter.org/sites/default/files/Do%20The%20Cheated%20Ever%20Prosper%20Final.pdf

Atteberry, A. Loeb, S., & Wyckoff, J. (2015). Do first impressions matter? Predicting early career teacher effectiveness. *AERA Open, 1*(4), 1-23.

Backes, B., Cowan, J., Goldhaber, D., Koedel, C., Miller, L., & Xu, Z. (2018). The common core conundrum: To what extent should I worry that changes to assessments and standards will affect test-based measures of teacher performance? *Economics of Education Review, 62*(1), 48-65.

Baker, E. L. Barton, P. E., Darling-Hammond, L., Haertel, E. Ladd, H. F., Linn, R. L., et al. (2010). *Problems with the use of student test scores to evaluate teachers* (EPI Briefing Paper #278). Washington, DC: Economic Policy Institute.

Balch, R. & Springer, M. G. (2015). Performance pay, test scores, and student learning objectives. *Economics of Education Review, 44*(1) 114-125.

Balingit, M. & Tran, A. B. (2018, January 6). Before a graduation scandal made headlines, teachers at D.C.'s Ballou High raised an alarm. *Washington Post*. Retrieved from: https://www.washingtonpost.com/local/education/before-a-graduation-scandal-made-headlines-teachers-at-dcs-ballou-high-raised-an-alarm/2018/01/06/ad49f198-df6a-11e7-89e8-edec16379010_story.html?utm_term=.7a95adfa3e20

Bill and Melinda Gates Foundation. (2010). *Fewer, clearer, higher: Moving forward with consistent, rigorous standards for all students*. Seattle, WA: Bill and Melinda Gates Foundation.

Bill & Melinda Gates Foundation 2018 Annual Letter. (2018, Feburary 13). *The toughest questions we get.* Seattle, WA: Bill & Melinda Gates Foundation. Retrieved from https://www.gatesnotes.com/2018-Annual-Letter

Boyd, D. J., Lankford, H., Loeb, S., Rockoff, J. E., & Wyckoff, J. (2008). The narrowing gap in New York City teacher qualifications and its implications for student achievement in high-poverty schools. *Journal of Policy Analysis and Management*, *27*(4), 793-818.

Brown, E. (2016, February). Report: Kids who took Common Core test online scored lower than those who used paper. *The Washington Post*.

Brown, E., Strauss, V., & Stein, P. (2018, March 10). It was hailed as the national model for school reform. Then the scandals hit. *The Washington Post.* Retrieved from: https://www.washingtonpost.com/local/education/dc-school-scandals-tell-me-that-its-not-great-and-that-youre-dealing-with-it/2018/03/10/b73d9cf0-1d9e-11e8-b2d9-08e748f892c0_story.html?utm_term=.52f11dc57cbf

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys, 22*(1), 31-72.

Campbell, S. L. & Ronfeldt, M. (2018). Observational evaluation of teachers: Measuring more than we bargained for? *American Educational Research Journal,* 1-35.

Carmichael, S. B., Martino, G., Porter-Magee, K., & Wilson, W. S. (2010). *The state of state standards—and the Common Core—in 2010*. Washington, DC: Thomas B. Fordham Foundation.

Cattaneo, M. D., Idrobo, N., & Titiunik, R. (2018a). *A practical introduction to regression discontinuity designs: Part I*, In preparation for Cambridge Elements: Quantitative and Computational Methods for Social Science, Cambridge University Press.

Cattaneo, M. D., Idrobo, N., & Titiunik, R. (2018b). *A practical introduction to regression discontinuity designs: Part II*, In preparation for Cambridge Elements: Quantitative and Computational Methods for Social Science, Cambridge University Press.

Cattaneo, M. D., Jansson, M., & Ma, X. (2017). *Simple local polynomial density estimators*. Working paper, University of Michigan.

Cattaneo, M. D., Jansson, M., & Ma, X. (2018). Manipulation testing based on density discontinuity. *The Stata Journal, 18*(1), 234-261.

Chang, K. (2013, September 2). With Common Core, fewer topics but covered more rigorously. *New York Times*. Retrieved from http://www.nytimes.com/2013/09/03/science/fewer-topics-covered-more-rigorously.html

Charalambous, C. Y., Hill, H. C., & Mitchell, R. N. (2012). Two negatives don't always make a positive: Exploring how limitations in teacher knowledge and the curriculum contribute to instructional quality. *Journal of Curriculum Studies, 44*(4), 289-513.

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

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

Chiang, H., Speroni, C., Herrmann, M. Hallgren, K., Burkaender, P., Wellington, A., & Warner, E. (2017). *Evaluation of the Teacher Incentive Fund: Final report on implementation and impacts of pay-for-performance across four years* (NCEE 2018-4005). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Chuong, C. (2014, September 8). The inconsistent implementation of teacher evaluation reforms. *EducationNext*. Retrieved from: http://educationnext.org/inconsistent-implementation-teacher-evaluation-reforms/

Clotfelter, C. T., Glennie, E., Ladd, H. F., & Vigdor, J. L. (2008). Would higher salaries keep teachers in high-poverty schools? Evidence from a policy intervention in North Carolina. *Journal of Public Economics*, *92*(5-6), 1352-1370.

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

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

Cochran-Smith, M., Villegas, A.M., Abrams, L., Chavez Moreno, L., Mills, T., & Stern, R. (2016). Research on teacher preparation: Charting the landscape of a sprawling field. In Gitomer, D. & Bell, C. (Eds.). *Handbook of Research on Teaching* (5th ed., pp. 439-547). Washington, DC: American Educational Research Association.

Cohen, J. (2015). Challenges in identifying high-leverage practices. *Teachers College Record, 117*(7), 1-41.

Cohen, J. Loeb, S. Miller, L. C., & Wyckoff, J. H. (2019). *Policy implementation, principal agency, and strategic action: Improving teaching effectiveness in New York City middle schools.* EdPolicyWorks Working Paper, University of Virginia.

Common Core or something else? A map of state academic standards. (2015). *Education Week, 34*(36).

Common Core State Standards Initiative. (2010a). Common core state standards for English language arts & literacy in history/social studies, science, and technical subjects.

Common Core State Standards Initiative. (2010b). Common core state standards for mathematics.

Conley, D. T. (2014). *The Common Core State Standards: Insight into their development and purpose.* Washington, DC: Council of Chief State School Officers.

Cook, J. B. & Mansfield, R. K. (2017). Task-specific experience and task-specific talent: Decomposing the productivity of high school teachers. *Journal of Public Economics, 140*, 51-71.

Corcoran, S. P., Jennings, J. L., & Beveridge, A. A. (2013, January). *Teacher effectiveness on high- and low-stakes tests*. Paper presented at the Annual Meeting of the American Economic Association, San Diego, CA.

Cowan, J. (April 24, 2017). *tfxreg* Stata package, version 1.4. Retrieved from https://github.com/jecowan/tfxreg

Cunningham, E. (2014). Opportunity costs of the Common Core in high school ELA. *English Journal, 104*(2), 34-40.

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

Darling-Hammond, L., & Adamson, F. (2014). *Beyond the bubble test: How performance assessments support 21st century learning*. John Wiley & Sons.

District of Columbia Public Schools [DCPS]. (2016). *FY2016 Performance Accountability Report*. Retrieved from https://oca.dc.gov/page/performance-plans-and-reports-agency

Dee, T.S., Dobbie, W., Jacob, B. & Rockoff, J. "The Causes and Consequences of Test Score Manipulation: Evidence from the New York Regents Examinations," *American Economic Journal: Applied Economics*, forthcoming.

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

Dehejia, R., & Wahba, S. (2002). Propensity score matching methods for non-experimental causal studies. *The Review of Economics Statistics, 84*, 151–161.

Desimone, L. M., & Garet, M. S. (2015). Best practices in teachers' professional development in the United States. *Psychology, Society and Education*, *7*(3), 252-263.

Desimone, L., Hochberg, E., Polikoff, M., Porter, A., Schwartz, R., & Johnson, L. (2014). Formal and informal mentoring: Compensatory, complementary, or consistent? *Journal of Teacher Education, 65*(2), 88-110.

Dinerstein, M. & Opper, I. M. (2017). *Does incentivizing value added make it more or less meaningful?* Paper presented at the 2017 annual meeting of the Association for Education Finance and Policy. Washington, DC.

Doherty, K. M., & Jacobs, S. (2015). *State of the states 2015: Evaluating teaching, leading and learning.* Washington, DC: National Council on Teacher Quality.

Donaldson, M. L., & Papay, J. P. (2015). Teacher evaluation for accountability and development. In H. Ladd F., & M. E. Goertz (Eds.), *Handbook of research in education finance and policy* (2nd ed., pp. 174-193). New York, NY: Routledge.

Donaldson, M. L., & Woulfin, S. (2018). From tinkering to going "rogue": How principals use agency when enacting new teacher evaluation systems. *Educational Evaluation and Policy Analysis*.

Doorey, N. & Polikoff, M. (2016). *Evaluating the content and quality of next generation assessments*. Washington, DC: Thomas B. Fordham Institute.

Dorans, N. J. & Holland, P. W. (2000). Population invariance and the equatability of tests: Basic theory and the linear case. *Journal of Educational Measurement, 37*(4), 281-306.

Duncan, A. (2014, August 21). A back-to-school conversation with teachers and school leaders [Blog post]. Retrieved from http://www.smartbrief.com/original/2014/08/back-school-conversation-teachers-and-school-leaders

Dynarski, M. (2016). *Teacher observations have been a waste of time and money*. Washington, DC: Brookings Institution. Retrieved from https://www.brookings.edu/research/teacher-observations-have-been-a-waste-of-time-and-money/

Ferguson, R. F. & Danielson, C. (2015) How framework for teaching and tripod 7Cs evidence distinguish key components of effective teaching. In T. J. Kane, K. A. Kerr and R. C. Pianta (Eds.), *Designing teacher evaluation systems*. San Francisco, CA: Wiley & Sons, Inc. doi: 10.1002/9781119210856.

Feuer, M. J., Holland, P. W., Green, B. F., Bertenthal, M. W., Hemphill, F. C. (Eds.). (1999). *Uncommon measures: equivalence and linkage among educational tests.* Washington, DC: The National Academies Press.

Fixsen, D. L., Naoom, S. F., Blase, K. A., Friedman, R. M., & Wallace, F. (2005). *Implementation research: A synthesis of the literature*. Tampa, FL: The University of South Florida, Louis de la Parte Florida Mental Health Institute, The National Implementation Research Network (FMHI Publication #231).

Fryer, R. (2013). Teacher incentives and student achievement: Evidence from New York City public schools. *Journal of Labor Economics, 31*(1), 373-427.

Fulbeck, E. S. (2014). Teacher mobility and financial incentives: A descriptive analysis of Denver's ProComp. *Educational Evaluation and Policy Analysis, 36*(1), 67-82.

Gill, B., Shoji, M., Coen, T., & Place, K. (2016). *The content, predictive power, and potential bias in five widely used teacher observation instruments (REL 2017-191).* Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Mid-Atlantic.

Glazerman, S., & Seifullah, A. (2012). *An evaluation of the Chicago teacher advancement program (Chicago TAP) after four years: Final report*. Washington, DC: Mathematica Policy Research, Inc.

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., Gross, B., & Player, D. (2011). Teacher career paths, teacher quality, and persistence in the classroom: Are public schools keeping their best? *Journal of Policy Analysis and Management, 30*(1), 57-87.

Goldhaber, D. & Hansen, M. (2013). Is it just a bad class? Assessing the long-term stability of estimated teacher performance. *Economica, 80*, 589-612.

Goldhaber, D., Quince, V., & Theobald, R. (2017). Has it always been this way? Tracing the evolution of teacher quality gaps in U.S. public schools. *American Educational Research Journal, 55*(1), 171-201.

Gordon, R., Kane, T. J., & Staiger, D. O. (2006). *Identifying effective teachers using performance on the job* (The Hamilton Project Discussion Paper No. 2006-01). Washington, DC: The Brookings Institution.

Gregory, A., Allen, J. P., Mikami, A. Y., Hafen, C. A., & Pianta, R. C. (2014). Effects of a professional development program on behavioral engagement of students in middle and high school. *Psychology in the Schools*, *51*(2), 143-163.

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

Haberman, S. J. (2015). Pseudo-equivalent groups and linking. *Journal of Educational and Behavioral Statistics, 40*(3), 254-273.

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.

Hanushek, E. A. &, Kain, J. F., O'Brien, D. M., & Rivkin, S. G. (2005). *The market for teacher quality* (NBER Working Paper No. 11154). Cambridge, MA: National Bureau of Economic Research.

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

Herold, B. (2016). PARCC scores lower for students who took exams on computers: Discrepancy raises questions about fairness. *Education Week, 35*(20), 1-11.

Hill, H. (2001). Policy is not enough: Language and the interpretation of state standards. *American Educational Research Journal, 38*(2), 289-318.

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

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2007). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives, 2*(3), 172-177.

Ho, A. D., & Kane, T. J. (2013). *The reliability of classroom observations by school personnel.* Seattle, WA: Measures of Effective Teaching Project, Bill and Melinda Gates Foundation.

Holland, P. W. & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 187-220). Westport, CT: American Council on Education.

Huggins, A. C., & Penfield, R. D. (2012). An NCME instructional module on population invariance in linking and equating. *Educational Measurement: Issues and Practice*, *31*(1), 27-40.

Iasevoli, B. (2018, February 15). Teacher-evaluation efforts haven't shown results, say Bill and Melinda Gates. Retrieved from http://blogs.edweek.org/edweek/teacherbeat/2018/02/teacher_evaluation_efforts_haven%27t_shown_results_bill_melinda_gates.html?cmp=soc-edit-tw

Ingersoll, R. M., & Strong, M. (2011). The impact of induction and mentoring programs for beginning teachers: A critical review of the research. *Review of Educational Research, 81*(2), 201-233.

Isenberg, E., Max, J. Gleason, P., Johnson, M., Deutsch, J., & Hansen, M. (2016). *Do low-income students have equal access to effective teachers? Evidence from 26 districts*. NCEE 2017-4007. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Isenberg, E. & Walsh, E. (2014). *Final report: Measuring teacher value added in DC, 2013-2014 school year.* Washington, DC: Mathematica Policy Research.

Jackson, C. K. (2009). Student demographics, teacher sorting, and teacher quality: evidence from the end of school desegregation. *Journal of Labor Economics, 27*(2), 213-256.

Jackson, C. K. & Bruegmann, E. (2009). Teaching students and teaching each other: The importance of peer learning for teachers. *American Economic Journal: Applied Economics, 1*(4), 85-108.

Jacob, B. A. (2005). Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics, 89*, 761-796.

Jacob, B. A., & Levitt, S. D. (2003). Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *Quarterly Journal of Economics, 118*, 843-877.

Jacob, B. A., Rockoff, J. A., Taylor, E. S., Lindy, B., & Rosen, R. (2018). Teacher applicant hiring and teacher performance: Evidence from DC public schools. *Journal of Public Economics, 166*(1), 81-97.

Jennings, J. L. & Lauen, D. L. (2016). Accountability, inequality, and achievement: The effects of the No Child Left Behind act on multiple measures of student learning. *Russell Sage Foundation Journal, 2*, 220–241.

Jochim, A. & McGuinn, P. (2016). The politics of the Common Core assessments: Why states are quitting the PARCC and Smarter Balanced consortia. *Education Next, 14*(4), 44-52.

Johnson, S. M., Kraft, M. A., & Papay, J. P. (2012). How context matters in high-need schools: The effects of teachers' working conditions on their professional satisfaction and their students' achievement. *Teachers College Record, 114*(10), 1-39.

Kalogrides, D. Loeb, S., & Bèteille, T. (2012). Systematic sorting: Teacher characteristics and class assignments. *Sociology of Education, 86*(2), 103-123.

Kane, T. J., McCaffrey, D. F., Miller, T. & Staiger, D. O. (2013). *Have we identified effective teachers? Validating measures of effective teaching using random assignment*. Seattle, WA: The Bill and Melinda Gates Foundation, Measures of Effective Teaching Project.

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

Kane, T. J., Owens, A. M., Marinell, W. H., Thal, D. R., & Staiger, D. O. (2016). *Teaching higher: Educators' perspectives on Common Core implementation*. Cambridge, MA: Center for Education Policy Research.

Kane, T. J., & Staiger, D. O. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation.* (NBER Working Paper No. 14607). Cambridge, MA: National Bureau of Economic Research.

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

Katz, V. & Wiseman, E. (2018). *Using financial incentives to attract and retain high-performing teachers in low-performing and low-income schools: Evidence from DCPS using a 7-year panel*. Paper presented at the 2018 annual meeting of the Association for Education Finance and Policy. Portland, OR.

Kim, S. & Lu R. (2018). *The pseudo-equivalent groups approach as an alternative to common-item equating. ETS RR-18-02*. Princeton, NJ: ETS Research Report.

Kim, Y. K. & DeCarlo, L. T. (2016). *Evaluating equity at the local level using bootstrap tests*. (Research Report No. 2016-4). New York, NY: The College Board.

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

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

Kolen, M. J. & Brennan, R. L. (2014). *Test equating, scaling, and linking: methods and practices* (3rd ed.). New York, NY: Springer.

Kolluri, S. (2018). Student perspectives on the Common Core: The challenge of college readiness at urban high schools. *Urban Education*, 1-28.

Kraft, M. A., Blazar, D., & Hogan, D. (2018). The effect of teacher coaching on instruction and achievement: A meta-analysis of the causal evidence. *Review of Educational Research, 88*(4), 547-588.

Kraft, M. A. & Gilmour, A. (2016). Can principals promote teacher development as evaluators? A case study of principals' views and experiences. *Educational Administration Quarterly, 52*(5), 711-753.

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*, *36*(4), 476-500.

Kraft, M.A., Papay, J.P., & Chi, O.L. (2018). *Teacher skill development: Evidence from performance ratings by principals*. Brown University Working Paper.

Lankford, H., Loeb, S., & Wyckoff, J. (2002). Teacher sorting and the plight of urban schools: A descriptive analysis. *Educational Evaluation and Policy Analysis, 24(*1), 37-62.

Lee, E., Lee, W. C., & Brennan, R. L. (2010). *Assessing equating results based on first-order and second-order equity*. (CASMA Research Report No. 31). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa.

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

Lockwood, J. R., McCaffrey, D. F., Hamilton, L., Stecher, B., Le, V. N., & Martinez, J. F. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Education Measurement, 44*(1), 47-67.

Longford, N. T. (2015). Equating without an anchor for nonequivalent groups of examinees. *Journal of Educational and Behavioral Statistics, 40*(3), 227-253.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Malmberg, L.E., Hagger, H., Burn, K., Mutton, T., & Colls, H. (2010). Observed classroom quality during teacher education and two years of professional practice. *Journal of Educational Psychology*, *102*(4), 916-932.

Marsh, J. A., Springer, M. G., McCaffrey, D. F., Yuan, K., Epstein, S., Koppich, J., Kalra, N., DiMartino, C., Peng, X. (2011). *A big apple for educators: New York City's experiment with schoolwide performance bonuses: Final evaluation report*. Santa Monica, CA: RAND Corporation.

McCaffrey, D. F., Sass, T. R., Lockwood, J. R., & Mihaly, K. (2009). The intertemporal variability of teacher effect estimates. *Education Finance and Policy, 4*(4), 672-606.

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

McDonnell, L. M. & Weatherford, M. S. (2013). Evidence use and the Common Core State Standards movement: From problem adoption to policy adoption. *American Journal of Education, 120*(1), 1-25.

McDuffie, A. R., Drake, C., Choppin, J., Davis, J. D., Magaña, M. V., & Carson, C. (2017). Middle school mathematics teachers' perceptions of the Common Core State Standards for Mathematics and related assessment and teacher evaluation systems. *Educational Policy, 31*(2), 139-179.

McGee, K. (2018, January 25). Most DCPS teachers feel pressure to pass students, teacher union survey says. *WAMU.* Retrieved from: https://wamu.org/story/18/01/25/dcps-teachers-feel-pressure-pass-students-teacher-union-survey-says/

McGuinn, D. (2012). Stimulating reform: Race to the Top, competitive grants, and the Obama education agenda. *Educational Policy, 26*(1), 136-159.

McNeil, M. (2014, March 19). Race to Top reports detail winners' progress, challenges. *Education Week*. Retrieved from: https://www.edweek.org/ew/articles/2014/03/19/26rtt.h33.html

Meyer, J. P. (2016). *Reliability of and validity evidence for Teaching Learning Framework scores for the District of Columbia public school system*. Unpublished manuscript, Curry School of Education, University of Virginia, Charlottesville, VA.

Mihaly, K., McCaffrey, D., Lockwood, J. R., & Sass, T. (2010). Centering and reference groups for estimates of fixed effects: Modifications to felsdvreg. *Stata Journal, 10*, 82-103.

Morris, C. N. (1982). On the foundations of test equating. In P. W. Holland & D. B. Rubin (Eds.), *Test equating*. New York, NY: Academic Press.

National Council for Teacher Quality [NCTQ] (2017). *Running in place: How new teacher evaluations fail to live up to promises.* Report. Retrieved from https://www.nctq.org/publications/Running-in-Place:-How-New-Teacher-Evaluations-Fail-to-Live-Up-to-Promises

National Governors Association [NGA. (2008). *Benchmarking for success: Ensuring U.S. students receive a world-class education*. Washington, DC: National Governors Association, Council of Chief State School Officers, and Achieve, Inc.

Office of the State Superintendent for Education [OSSE]. (2011a). ELA Common Core Crosswalk [Microsoft Excel spreadsheet]. Washington, DC: Retrieved from https://osse.dc.gov/service/common-core-state-standards.

Office of the State Superintendent for Education [OSSE]. (2011b). Math Common Core Crosswalk [Microsoft Excel spreadsheet]. Washington, DC : Retrieved from https://osse.dc.gov/service/common-core-state-standards.

Ost, B. (20149). How do teachers improve? The relative importance of specific and general human capital. *American Economic Journal: Applied Economics*, 127-151.

Ost, B., Gangopadhyaya, A., & Schiman, J. C. (2017). Comparing standard deviation effects across contexts. *Education Economics, 25*(3), 251-265.

Otis, A. S. (1922). The method for finding the correspondence between scores in two tests. *Journal of Educational Psychology, 13*(1), 529-545.

Papay, J. P. (2011). Different tests, different answers: The stability of teacher value-added estimates across outcome measures. *American Educational Research Journal, 48*(1), 163-193.

Papay, J. P., Bacher-Hicks, A., Page, L. C., & Marinell, W. H. (2017). The challenge of teacher retention in urban schools: evidence of variation from a cross-site analysis. *Educational Researcher, 46*(8), 434-448.

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, 1*(30), 105-119.

Papay, J. P. & Laski, M. (2018). *Exploring teacher improvement in Tennessee: A brief on reimagining state support for professional learning*. Nashville, TN: TN Education Research Alliance.

Papay, J. P., Taylor, E. S., Tyler, J. H., & Laski, M. (2016). *Learning job skills from colleagues at work: Evidence from a field experiment using teacher performance data* (NBER Working Paper No. 21986). Cambridge, MA: National Bureau of Economic Research.

Penney, J. (2017). A self-reference problem in test score normalization. *Economics of Education Review, 61*(4), 79-84.

Peterson, P. E., Barrows, S., & Gift, T. (2016). After Common Core, states set rigorous standards. *Education Next, 16*(3), 9-15.

Polikoff, M. S. (2012). Instructional alignment under No Child Left Behind. *American Journal of Education, 118*, 341–368.

Polikoff, M. S. (2015). How well aligned are textbooks to the Common Core Standards in mathematics? *American Educational Research Journal, 52*(6), 1185-1211.

Polikoff, M. S. (2017). Is Common Core "working"? And where does Common Core research go from here? *AERA Open*, *3*(1), 1-6.

Pope, N. (2019). The effect of teacher ratings on teacher performance. *Journal of Public Economics*, 172(1), 84-110.

Reardon, S. F. (2013). The widening income achievement gap. *Educational Leadership, 30*(8).

Reardon, S.F., Kalogrides, D., & Ho, A. (2019). *Validation methods for aggregate-level test scale linking: A case study mapping school district test score distributions to a common scale* (CEPA Working Paper No.16-09). Retrieved from Stanford Center for Education Policy Analysis: http://cepa.stanford.edu/wp16-09

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

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

Rothstein, J. (2015). Teacher quality policy when supply matters. *American Economic Review, 105*(1), 100-130.

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

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.

Schmidt, W. H., McKnight, C. C., & Raizen, S. (Eds.). (2002). *A splintered vision:An investigation of U.S. science and mathematics education* (1st ed.). https://doi.org/10.1007/0-306-47209-0

Schultz, S. R., Michaels, H. R., Dvorak, R. N., & Wiley, C. R. H. (2016). *Evaluating the content and quality of next generation high school assessments*. Alexandria, VA: Human Resources Research Corporation.

Springer, M. G., 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, Vanderbilt University.

Springer, M. G., 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*(4), 367-390.

Springer, M. G., Swain, W. A., & Rodriguez, L. A. (2016). Effective teacher retention bonuses: Evidence from Tennessee. *Educational Evaluation and Policy Analysis, 38*(2), 199-221.

Stacy, B., Guarino, C., & Wooldridge, J. (2018). Does the precision and stability of value-added estimates of teacher performance depend on the types of students they serve? *Economics of Education Review, 64*, 50-74.

Stecher, B. M., Holtzman, D. J., Garet, M. S., Hamilton, L. S., Engberg, J. Steiner, E. D., . . . Chamber, J. (2018). *Improving teaching effectiveness: Final report: The Intensive Partnerships for effective teaching through 2015-16*. Santa Monica, CA: RAND Corporation.

Steele, J. L., Murnane, R. J., & Willett, J. B. (2010). Do financial incentives help low-performing schools attract and keep academically talented teachers? Evidence from California. *Journal of Policy Analysis and Management, 29*(3), 451-478.

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

Steinberg, M. P., & Sartain, L. (2015). Does teacher evaluation improve school performance? Experimental evidence from Chicago's Excellence in Teaching project. *Education Finance and Policy, 10*(4), 535-572.

Steiner, P. M., Cook, T. D., Shadish, W. R., & Clark, M. H. (2010). The importance of covariate selection in controlling for selection bias in observational studies. *Psychological methods*, *15*(3), 250.

Stosich, E. L. (2016). Joint inquiry: Teachers' collective learning about the Common Core in high-poverty urban schools. *American Educational Research Journal, 53*(6), 1698-1731.

Strauss, V. (2015, January 1). Teacher evaluation: Going from bad to worse. *The Washington Post*. Retrieved from: https://www.washingtonpost.com/news/answer-sheet/wp/2015/01/01/teacher-evaluation-going-from-bad-to-worse/?utm_term=.798b25c88256

Student Achievement Partners (2013). Introduction to the ELA/literacy shifts of the Common Core State Standards [PowerPoint slides]. Retrieved from the Achieve the Core website: http://achievethecore.org/page/394/introduction-to-the-ela-literacy-shifts

Student Achievement Partners (2014). Introduction to the math shifts of the Common Core State Standards [PowerPoint slides]. Retrieved from the Achieve the Core website: http://achievethecore.org/page/399/introduction-to-the-math-shifts

Student Achievement Partners. (2017a). *Grade 3 math alignment module* [PowerPoint slides]. Retrieved from https://achievethecore.org/page/2886/mathematics-assessment-item-alignment-modules.

Student Achievement Partners. (2017b). *Grade 5 ELA/literacy alignment module* [PowerPoint slides]. Retrieved from https://achievethecore.org/page/2885/ela-literacy-assessment-item-alignment-modules.

Sun, M., Mutcheson, R. B., & Kim, J. (2015). Teachers' use of evaluation for instructional improvement and school supports for this use. In J. Grissom, & P. Youngs (Eds.), *Making the most of multiple measures: The impacts and challenges of implementing rigorous teacher evaluation systems* (First ed., pp. 102-115). New York, NY: Teachers College Press.

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

Thorndike, E. L. (1922). On finding equivalent scores in tests of intelligence. *Journal of Applied Psychology, 6*(1), 29-33.

Toch, T. (2018). *A policymaker's playbook: Transforming public school teaching in the nation's capital.* Washington, DC: Future Ed, Georgetown University https://www.future-ed.org/wp-content/uploads/2018/06/APOLICYMAKERSPLAYBOOK.pdf

Tong, Y. & Kolen, M. J. (2005). Assessing equating results on different equating criteria. *Applied Psychological Measurement, 29*(6), 418-432.

U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, Common Core of Data (CCD), "Local Education Agency (School District) Universe Survey", 2013-14 v.1a; "Public Elementary/Secondary School Universe Survey", 2008-09 v.1b.

U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2007 Reading and Math Assessments. Reports generated using the NAEP Data Explorer. http://nces.ed.gov/nationsreportcard/naepdata

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness*. New York, NY: The New Teacher Project.

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

What Works Clearinghouse [WWC]. (2017). *What Works Clearinghouse procedures and standards handbooks* (Version 4.0). Washington, DC: Institute for Education Sciences. Retrieved from https://ies.ed.gov/ncee/wwc/handbooks#procedures

Wiswall, M. (2013). The dynamics of teacher quality. *Journal of Public Economics, 100*, 61-78.

Xu, Z., Özek, U., & Hansen, M. (2015). Teacher performance trajectories in high- and lower-poverty schools. *Educational Evaluation and Policy Analysis, 37*(4), 458-477.

Yuan, K. & Le, V. (2012). *Estimating the percentage of students who were tested on cognitively demanding items through the state achievement tests*. Santa Monica, CA: RAND Corporation.

APPENDIX A

Additional Tables and Figures

TABLE A.1.1

The 9 Teach standards of the Teaching and Learning Framework

STANDARD	DESCRIPTION OF HIGHLY EFFECTIVE TEACHING
Teach 1 *Lead well-organized, objective-driven lessons*	*Lesson Organization* 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 all students toward mastery of the objective. *Lesson Objective* The objective of the lesson is clear to students and conveys what students are learning and what they will be able to do as a result of the lesson. Students also can authentically explain what they are learning and doing beyond simply repeating the stated or posted objective. *Objective Importance* Students understand the importance of the objective. Students also can authentically explain why what they are learning and doing is important, beyond simply repeating the teachers' explanation.
Teach 2 *Explain content clearly*	*Clear, Coherent Delivery* Explanations of content are clear and coherent, and they build student understanding of content. The teacher might provide explanations through direct verbal or written delivery, modeling or demonstrations, think-alouds, visuals, or questioning. Explanations of content also are delivered in as direct and efficient a manner as possible. *Academic Language* The teacher gives clear, precise definitions and uses a broad vocabulary that includes specific academic language and words that may be unfamiliar to students when it is appropriate to do so. Students also demonstrate through their verbal or written responses that they are internalizing academic vocabulary. *Emphasize Key Points* The teacher emphasizes key points when necessary, such that students understand the main ideas of the content. Students also can authentically explain the main ideas of the content beyond simply repeating back the teacher's explanations.

	Student Understanding Students show that they understand the explanations. When appropriate, concepts also are explained in a way that actively and effectively involves students in the learning process. For example, students have opportunities to explain concepts to each other. *Connections* The teacher makes connections with students' prior knowledge, students' experiences and interests, other content areas, or current events to effectively build student understanding of content.
Teach 3 *Engage students at all learning levels in accessible and challenging work*	*Accessibility* The teacher makes the lesson accessible to all students. There is evidence that the teacher knows each student's level and ensures that the lesson meets all students where they are. *Challenge* The teacher makes the lesson challenging to all students. There is evidence that the teacher knows each student's level and ensures that the lesson pushes all students forward from where they are. *Balance* There is an appropriate balance between teacher-directed and 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 move toward mastery*	*Multiple Ways Toward Mastery* The teacher provides students multiple ways to engage with content, and all ways move students toward mastery of lesson content. During the lesson, students are also developing deep understanding of the content. *Appropriateness for Students* The ways the teacher provides include learning styles or modalities that are appropriate to students' needs; all students respond positively and are actively involved in the work.

Teach 5 *Check for student understanding*	*Key Moments* The teacher checks for understanding of content at all key moments. *Accurate Pulse* The teacher always gets an accurate “pulse” at key moments by using one or more checks that gather information about the depth of understanding for a range of students, when appropriate.
Teach 6 *Respond to student understanding*	*Scaffolding* When students demonstrate misunderstandings or partial understandings, the teacher always uses effective scaffolding techniques that enable students to construct their own understandings, when appropriate. *Re-Teaching* The teacher always re-teaches effectively when appropriate, such as in cases in which most of the class demonstrates a misunderstanding or an individual student demonstrates a significant misunderstanding. The teacher also anticipates common misunderstandings (e.g., by offering a misunderstanding as a correct answer to see how students respond) or recognizes a student response as a common misunderstanding and shares it with the class to lead all students to a more complete understanding. *Probing* The teacher always probes students’ correct responses, when appropriate, to ensure student understanding.
Teach 7 *Develop higher-level understanding through effective questioning*	*Questions and Tasks* The teacher asks questions that push all students’ thinking; when appropriate, the teacher also poses tasks that are increasingly complex that develop all students’ higher-level understanding. *Support* After posing a question or task, the teacher always uses appropriate strategies to ensure that students move toward higher-level understanding. *Meaningful Response* Almost all students answer questions of complete complex tasks with meaningful responses that demonstrate movement toward higher-level understanding, showing that they are accustomed to being asked these kinds of questions.

Teach 8 *Maximize instructional time*	*Routines, Procedures, and Transitions* Routines, procedures, and transitions are orderly, efficient, and systematic with minimal prompting from the teacher. Students know their responsibilities and some students share responsibility for leading the operations and routines in the classroom. *Student Idleness* Students always have something meaningful to do. Lesson pacing is also student-directed or individualized, when appropriate. *Lesson Pacing* The teacher spends an appropriate amount of time on each part of the lesson. *Student Behavior* Inappropriate or off-task student behavior never interrupts or delays the lesson, either because no such behavior occurs or because when such behavior occurs the teacher efficiently addresses it.
Teach 9 *Build a supportive, learning-focused classroom community*	*Investment* Students are invested in their work and value academic success. Students are also invested in the success of their peers. For example, students can be seen helping each other or showing interest in other students' work without prompting from the teacher. *Risk-Taking* The classroom environment is safe for students, such that students are willing to take on challenges and risk failure. For example, students are eager to ask questions, feel comfortable asking the teacher for help, feel comfortable engaging in constructive feedback with their classmates, and do not respond negatively when a peer answers a question incorrectly. *Respect* Students are always respectful of the teacher and their peers. For example, students listen and do not interrupt when their peers ask or answer questions.

	Reinforcement The teacher meaningfully reinforces positive behavior and good academic work, when appropriate. Students also give unsolicited praise or encouragement to their peers, when appropriate. *Rapport* 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. There is also evidence that the teacher has strong, individualized relationships with some students in the class.

TABLE A.1.2

Pairwise Correlations Between Teaching and Learning Framework (TLF) Scores Across Testing Regimes

	Teach		Teach	Teach	Teach	Teach	Teach	Teach	Teach	Teach	Teach
	Domain	Overall	1	2	3	4	5	6	7	8	9
Both exams	Overall	1.00									
	Teach 1	0.78	1.00								
	Teach 2	0.79	0.59	1.00							
	Teach 3	0.81	0.58	0.60	1.00						
	Teach 4	0.82	0.63	0.61	0.65	1.00					
	Teach 5	0.79	0.55	0.60	0.59	0.59	1.00				
	Teach 6	0.75	0.49	0.57	0.56	0.52	0.59	1.00			
	Teach 7	0.78	0.55	0.58	0.62	0.59	0.57	0.54	1.00		
	Teach 8	0.76	0.53	0.50	0.57	0.58	0.54	0.46	0.51	1.00	
	Teach 9	0.73	0.50	0.49	0.54	0.53	0.51	0.46	0.48	0.67	1.00
CAS	Overall	1.00									
	Teach 1	0.75	1.00								
	Teach 2	0.80	0.57	1.00							
	Teach 3	0.81	0.54	0.61	1.00						
	Teach 4	0.81	0.59	0.61	0.64	1.00					
	Teach 5	0.80	0.54	0.62	0.60	0.61	1.00				
	Teach 6	0.75	0.46	0.58	0.58	0.52	0.58	1.00			
	Teach 7	0.78	0.53	0.59	0.60	0.58	0.59	0.54	1.00		
	Teach 8	0.76	0.51	0.52	0.58	0.58	0.55	0.47	0.52	1.00	
	Teach 9	0.74	0.48	0.50	0.56	0.54	0.53	0.48	0.49	0.66	1.00
PARCC	Overall	1.00									
	Teach 1	0.84	1.00								
	Teach 2	0.76	0.63	1.00							
	Teach 3	0.82	0.67	0.57	1.00						
	Teach 4	0.82	0.72	0.60	0.67	1.00					
	Teach 5	0.76	0.56	0.52	0.55	0.55	1.00				
	Teach 6	0.74	0.54	0.57	0.52	0.54	0.60	1.00			
	Teach 7	0.79	0.61	0.55	0.66	0.62	0.52	0.53	1.00		
	Teach 8	0.75	0.58	0.46	0.54	0.56	0.51	0.46	0.48	1.00	
	Teach 9	0.71	0.53	0.45	0.49	0.52	0.47	0.42	0.46	0.68	1.00

Note. Values are Pearson correlation coefficients; all are significant at $p<0.001$.

TABLE A.1.3

Factor Loadings from Exploratory Factor Analysis of the TLF

TLF Domain	Factor 1 *Instruction*	Factor 2 *Classroom Environment*	Uniqueness
Teach 1	**0.67**	0.28	0.48
Teach 2	**0.79**	0.22	0.33
Teach 3	**0.45**	**0.63**	0.41
Teach 4	**0.63**	**0.44**	0.41
Teach 5	**0.74**	0.27	0.38
Teach 6	**0.69**	0.32	0.42
Teach 7	**0.60**	0.42	0.46
Teach 8	0.20	**0.87**	0.20
Teach 9	0.25	**0.80**	0.30

Notes. N=35,055 observations from external ("Master Educator") evaluators in DCPS between AY2009-10 and AY2015-16. Varimax rotated factor loadings of 0.40 or higher are in bold. Results are from a principal-component factor analysis in which the TLF is forced to load onto more than one factor. The first factor has an eigenvalue of 4.80 and explains 53% of the variance in TLF scores; the second factor has an eigenvalue of 0.82 and explains an additional 9% of the variance.

TABLE A.1.4
Stability of Value-Added Percentile Ranks Across the Transition to a New Exam

	Math				ELA			
PARCC	-1.96	-1.22	0.13	0.18	-1.70	-1.31	-0.68	-0.19
	(1.74)	(1.81)	(1.80)	(2.39)	(1.64)	(1.70)	(1.78)	(2.56)
Classroom and teacher controls		X	X	X		X	X	X
School FE			X	X			X	X
Teacher FE				X				X
Constant	24.44 ***	14.96 ***	14.02	12.80	24.27 ***	-1.37	18.94 *	16.40
	(0.64)	(5.64)	(11.30)	(16.53)	(0.63)	(6.00)	(9.55)	(16.91)
n	1,111	1,111	1,111	1,111	1,168	1,168	1,168	1,168
R-squared	0.00	0.04	0.19	0.64	0.00	0.08	0.18	0.63

Notes. Classroom controls include the proportion of students who are male, black, Hispanic, another non-white race, eligible for free or reduced-price lunch, with limited English proficiency, or in special education, as well as the average lagged match test score and average lagged ELA test score. Teacher controls include experience level, prior IMPACT rating, and quintile of lagged value-added scores in the subject. PARCC is an indicator for years in which the PARCC exam was first administered (i.e., AY2014-15).
*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; and $^{+}$ $p < 0.10$

TABLE A.1.5a

Stability of Value-Added Percentile Ranks Across the Transition to a New Exam, by Class Characteristics

	Math			ELA		
Student characteristics						
% Male * PARCC	16.56	5.18	-5.72	16.90	3.58	-8.03
	(20.95)	(20.87)	(29.99)	(19.77)	(20.53)	(32.75)
% Black * PARCC	19.21	39.99 *	57.30 *	-2.51	22.19	37.62
	(15.30)	(17.51)	(27.09)	(17.88)	(17.84)	(23.61)
% Hispanic * PARCC	8.51	16.24	30.72	-10.90	15.31	34.53
	(16.81)	(18.48)	(25.69)	(18.85)	(18.32)	(24.45)
% Other race * PARCC	-2.92	2.11	53.10	-61.53	-45.65	-46.44
	(34.09)	(36.46)	(44.92)	(41.45)	(38.08)	(45.52)
% FRPL * PARCC	-15.47	-28.95 **	-38.13 *	-5.90	-17.17	-15.93
	(11.06)	(12.12)	(16.59)	(10.28)	(11.08)	(15.47)
% Limited English proficiency * PARCC	27.35	47.97	49.85	7.46	5.81	-19.03
	(33.97)	(30.39)	(40.19)	(30.62)	(31.85)	(42.19)
% Special education * PARCC	28.25	27.21	1.78	3.24	-13.16	10.25
	(30.17)	(27.44)	(31.86)	(23.05)	(22.53)	(37.91)
Mean lagged math score * PARCC	-13.06	-4.58	-3.73	4.62	3.03	8.40
	(10.15)	(8.92)	(13.24)	(7.35)	(8.02)	(12.15)
Mean lagged ELA score * PARCC	9.06	-0.69	-6.48	0.66	3.83	4.55
	(10.31)	(10.00)	(15.05)	(8.46)	(9.30)	(15.16)
Classroom and teacher controls	X	X	X	X	X	X
School FE		X	X		X	X
Teacher FE			X			X
n	1,111	1,111	1,111	1,168	1,168	1,168
R-squared	0.06	0.20	0.66	0.09	0.19	0.64

Notes. Classroom controls include the proportion of students who are male, black, Hispanic, another non-white race, eligible for free or reduced-price lunch, with limited English proficiency, or in special education, as well as the average lagged match test score and average lagged ELA test score. Teacher controls include experience level, prior IMPACT rating, and quintile of lagged value-added scores in the subject. PARCC is an indicator for years in which the PARCC exam was first administered (i.e., AY2014-15).

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; and $^{+}$ $p < 0.10$ before correcting for multiple hypothesis testing. Bonferroni corrections for multiple hypothesis testing show that none of the interacted teacher and classroom characteristics are significant at conventional levels.

TABLE A.1.5b

Stability of Value-Added Percentile Ranks Across the Transition to a New Exam, by Teacher Characteristics

	Math			ELA		
Teacher characteristics						
Experience < 3 years * PARCC	-8.43 *	-7.83 $^{+}$	-7.82	-4.10	-6.42	4.26
	(3.94)	(4.35)	(6.35)	(4.57)	(5.11)	(7.84)
Experience >= 10 years * PARCC	-7.25 $^{+}$	-6.93 $^{+}$	-4.57	-3.52	-2.97	-2.14
	(3.88)	(3.81)	(5.11)	(4.03)	(4.17)	(6.30)
Highly Effective in year *t* - 1 * PARCC	8.06 $^{+}$	7.14	6.93	-1.03	-4.83	-0.69
	(4.17)	(4.71)	(6.96)	(4.72)	(4.87)	(7.51)
Developing in year *t* - 1 * PARCC	4.72	-2.50	2.57	-6.60	-7.50	-6.55
	(4.92)	(4.96)	(7.25)	(4.26)	(4.58)	(7.50)
Minimally Effective in year *t* - 1 * PARCC	2.81	-2.35	-7.42	2.43	3.07	7.24
	(8.46)	(7.33)	(10.64)	(6.94)	(6.79)	(11.91)
Top quintile of IVA in year *t* - 1 PARCC	-0.49	-0.41	0.07	-3.90	-3.61	1.91
	(4.18)	(4.40)	(6.12)	(4.97)	(5.16)	(7.40)
Bottom quintile of IVA in year *t* - 1 * PARCC	-2.23	-1.93	-7.67	-7.63	-6.81	-16.97
	(6.81)	(5.75)	(9.85)	(6.49)	(6.45)	(10.78)
Classroom and teacher controls	X	X	X	X	X	X
School FE		X	X		X	X
Teacher FE			X			X
n	1,111	1,111	1,111	1,168	1,168	1,168
R-squared	0.06	0.20	0.66	0.09	0.19	0.64

Notes. Classroom controls include the proportion of students who are male, black, Hispanic, another non-white race, eligible for free or reduced-price lunch, with limited English proficiency, or in special education, as well as the average lagged match test score and average lagged ELA test score. Teacher controls include experience level, prior IMPACT rating, and quintile of lagged value-added scores in the subject. PARCC is an indicator for years in which the PARCC exam was first administered (i.e., AY2014-15).

$^{***} p < 0.001$; $^{**} p < 0.01$; $^{*} p < 0.05$; and $^{+} p < 0.10$ before correcting for multiple hypothesis testing. Bonferroni corrections for multiple hypothesis testing show that none of the interacted teacher and classroom characteristics are significant at conventional levels.

TABLE A.1.6

Robustness of PARCC-CAS Slopes to Sample Restrictions

		Math		
		Minimum years of VA available on each test		
		1 year	2 years	3 years
Minimum Prior Teaching Experience	3 years	0.326 ***	0.512 ***	0.458 ***
		-0.083	-0.134	-0.171
		176	*64*	*31*
		{66.10} ***	{13.29} ***	{10.08} **
	5 years	0.339 ***	0.525 ***	0.407 +
		-0.093	-0.184	-0.227
		133	*52*	*27*
		{50.24} ***	{6.64} *	{6.80} *
	10 years	0.691 ***	0.63 ***	0.447 +
		-0.111	-0.21	-0.258
		82	*35*	*20*
		{7.71} **	{3.09} +	{4.58} *

		ELA		
		Minimum years of VA available on each test		
		1 year	2 years	3 years
Minimum Prior Teaching Experience	3 years	0.474 ***	0.559 ***	0.69 ***
		-0.073	-0.124	-0.174
		165	*66*	*29*
		{51.84} ***	{12.69} ***	{3.16}
	5 years	0.435 ***	0.529 ***	0.709 ***
		-0.082	-0.133	-0.171
		134	*59*	*28*
		{47.40} ***	{12.60} ***	{2.91}
	10 years	0.406 ***	0.471 ***	0.583 ***
		-0.094	-0.131	-0.204
		81	*39*	*19*
		{40.17} ***	{16.45} ***	{4.17} +

Notes. Robust standard errors are in parentheses, sample sizes are in italics, and *F*-statistics for tests of the equality of slope coefficients to 1 are in curly brackets.
*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; and + $p < 0.10$

TABLE A.1.7

The Relative Association Between Teachers' Practice and Student Achievement Across Exams, Including Data from AY2014-15

	Math			ELA		
TLF Overall	-0.024 ***	-0.037 +	0.005	-0.002	0.026	0.019 ***
	(0.008)	(0.022)	(0.008)	(0.007)	(0.019)	(0.006)
Factor 1: *Instruction*	-0.025 **	-0.027	-0.011	-0.009	-0.002	0.007
	(0.008)	(0.018)	(0.008)	(0.008)	(0.016)	(0.007)
Factor 2: *Classroom Environment*	-0.005	-0.022	0.020 *	0.008	0.045 *	0.017 *
	(0.009)	(0.022)	(0.009)	(0.006)	(0.019)	(0.006)
Student and teacher controls	X	X	X	X	X	X
Teacher FE		X			X	
Student FE			X			X
n	21,739	21,739	21,739	27,109	27,109	27,109

Notes. Student controls include gender, race, eligibility for free or reduced-price lunch, limited English proficiency status, and special education status; teacher controls include race, gender, education, and experience. This table shows results from two regressions within each estimation model: the first rows show the interacted effects of overall TLF scores, standardized within year, and the PARCC exam on student achievement; the following rows show interacted effects between the PARCC exam and the *instruction* and *classroom environment* domains. PARCC exam scores used for this analysis are linked to the CAS scale and distribution using propensity-score matching followed by an equipercentile transformation (Approach 1 in appendix B); scores from each test are then standardized within subject and grade relative to the distribution of CAS scores in the final year of the CAS exam. TLF scores are standardized within year and use scores assigned only by external (i.e., master educators) evaluators. Data from AY2014-15 are included in this analysis.

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; and $^{+}p < 0.10$ after Bonferroni adjustment for multiple hypothesis testing.

TABLE A.1.8

The Relative Association Between Teachers' Practice and Student Achievement Across Exams: Results Across the Nine TLF Teach Domains

	Math			ELA		
Teach 1	0.016	0.202 $^{+}$	0.037	-0.011	-0.019	0.015
	(0.022)	(0.074)	(0.026)	(0.019)	(0.063)	(0.023)
Teach 2	-0.034	-0.335 **	-0.061	-0.038 $^{+}$	-0.069	-0.002
	(0.025)	(0.092)	(0.028)	(0.014)	(0.047)	(0.016)
Teach 3	0.006	-0.012	0.036	0.042	0.103	0.039
	(0.026)	(0.105)	(0.029)	(0.017)	(0.059)	(0.019)
Teach 4	-0.046	-0.157	-0.026	0.045 *	0.120	0.004
	(0.024)	(0.079)	(0.028)	(0.014)	(0.068)	(0.017)
Teach 5	0.010	0.135 $^{+}$	0.023	0.044 **	0.118	0.020
	(0.017)	(0.051)	(0.019)	(0.013)	(0.052)	(0.015)
Teach 6	-0.022	0.199 $^{+}$	-0.012	-0.023	-0.217 **	-0.010
	(0.018)	(0.073)	(0.020)	(0.014)	(0.058)	(0.015)
Teach 7	-0.005	0.029	-0.040	-0.022	0.050	-0.039
	(0.022)	(0.063)	(0.026)	(0.018)	(0.062)	(0.020)
Teach 8	0.020	-0.055	-0.008	-0.017	-0.038	-0.031
	(0.022)	(0.076)	(0.026)	(0.014)	(0.041)	(0.015)
Teach 9	0.059 *	0.047	0.089 ***	0.011	0.018	0.007
	(0.018)	(0.049)	(0.022)	(0.013)	(0.034)	(0.013)
Student and teacher controls	X	X	X	X	X	X
Teacher FE		X			X	
Student FE			X			X
n	16,616	16,616	16,616	21,153	21,153	21,153

Notes. Student controls include gender, race, eligibility for free or reduced-price lunch, limited English proficiency status, and special education status; teacher controls include race, gender, education, and experience. This table shows interacted effects between the PARCC exam and the nine Teach domains of the TLF. PARCC exam scores used for this analysis are linked to the CAS scale and distribution using propensity-score matching followed by an equipercentile transformation (Approach 1 in appendix B); scores from each test are then standardized within subject and grade relative to the distribution of CAS scores in the final year of the CAS exam. TLF scores are standardized within year and use scores assigned only by external (i.e., master educators) evaluators. Data from AY2014-15 are omitted from this analysis.

**** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; and $^{+}$ $p < 0.10$ after Bonferroni adjustment for multiple hypothesis testing.

TABLE A.1.9a

The Relative Association Between Teachers' Practice and Student <u>*Math*</u> *Achievement Across Exams: Estimates from Each Approach to Linking PARCC and CAS Scores*

	PSM	NAEP	Predicted	Not Linked
	Student & Teacher Controls			
TLF Overall	0.003	-0.005	-0.041***	-0.008
	(0.013)	(0.013)	(0.008)	(0.012)
Factor 1	-0.029*	-0.041**	-0.042***	-0.036**
Instruction	(0.013)	(0.012)	(0.008)	(0.012)
Factor 2	0.039**	0.041**	-0.010	0.032*
Classroom Environment	(0.013)	(0.013)	(0.008)	(0.013)
	Student & Teacher Controls + Teacher FE			
TLF Overall	0.034	0.005	-0.004	0.006
	(0.060)	(0.064)	(0.048)	(0.062)
Factor 1	0.025	0.000	0.014	0.006
Instruction	(0.045)	(0.047)	(0.035)	(0.046)
Factor 2	0.023	0.009	-0.017	0.004
Classroom Environment	(0.050)	(0.056)	(0.041)	(0.055)
	Student & Teacher Controls + Student FE			
TLF Overall	0.039***	0.035**	-0.011	0.034**
	(0.014)	(0.015)	(0.011)	(0.015)
Factor 1	-0.008	-0.018	-0.023^{+}	-0.012
Instruction	(0.014)	(0.015)	(0.011)	(0.014)
Factor 2	0.064***	0.071***	0.011	0.062***
Classroom Environment	(0.016)	(0.016)	(0.012)	(0.016)
n	15,765	15,765	15,765	15,765

Notes. Student controls include gender, race, eligibility for free or reduced-price lunch, limited English proficiency status, and special education status; teacher controls include race, gender, education, and experience. This table shows results from two regressions within each estimation model: the first rows show the interacted effects of overall TLF scores, standardized within year, and the PARCC exam on student achievement; the following rows show interacted effects between the PARCC exam and the *instruction* and *classroom environment* domains. Scores from each test are linked to the CAS scale using the approaches described in appendix B and are then standardized within subject and grade relative to the distribution of CAS scores in the final year of the CAS exam. TLF scores are standardized within year and use scores assigned only by external (i.e., master educators) evaluators. Data from AY2014-15 are omitted from this analysis.

**** *p* < 0.001; ** *p* < 0.01; * *p* < 0.05; and $^{+}$ *p* < 0.10 after Bonferroni adjustment for multiple hypothesis testing.

TABLE A.1.9b

The Relative Association Between Teachers' Practice and Student ELA Achievement Across Exams: Estimates from Each Approach to Linking PARCC and CAS Scores

	PSM	NAEP	Predicted	Not Linked
	Student & Teacher Controls			
TLF Overall	0.017 $^{+}$	0.011	0.004	0.009
	(0.009)	(0.009)	(0.007)	(0.009)
Factor 1 *Instruction*	0.012 (0.010)	0.011 (0.010)	0.004 (0.007)	0.009 (0.010)
Factor 2 *Classroom Environment*	0.013 (0.008)	0.009 (0.008)	0.005 (0.006)	0.006 (0.008)
	Student & Teacher Controls + Teacher FE			
TLF Overall	0.109 ***	0.106 ***	0.051	0.094 **
	(0.041)	(0.039)	(0.035)	(0.038)
Factor 1 *Instruction*	-0.004 (0.048)	0.029 (0.039)	-0.022 (0.040)	0.021 (0.039)
Factor 2 *Classroom Environment*	0.145 *** (0.034)	0.112 *** (0.032)	0.089 ** (0.029)	0.103 ** (0.031)
	Student & Teacher Controls + Student FE			
TLF Overall	0.000	0.000	-0.006	-0.002
	(0.010)	(0.010)	(0.009)	(0.010)
Factor 1 *Instruction*	0.009 (0.011)	0.013 (0.011)	0.007 (0.009)	0.011 (0.011)
Factor 2 *Classroom Environment*	-0.005 (0.009)	-0.009 (0.009)	-0.010 (0.008)	-0.008 (0.009)
n	21,739	21,739	21,739	21,739

Notes. Student controls include gender, race, eligibility for free or reduced-price lunch, limited English proficiency status, and special education status; teacher controls include race, gender, education, and experience. This table shows results from two regressions within each estimation model: the first rows show the interacted effects of overall TLF scores, standardized within year, and the PARCC exam on student achievement; the following rows show interacted effects between the PARCC exam and the *instruction* and *classroom environment* domains. Scores from each test are linked to the CAS scale using the approaches described in appendix B and are then standardized within subject and grade relative to the distribution of CAS scores in the final year of the CAS exam. TLF scores are standardized within year and use scores assigned only by external (i.e., master educators) evaluators. Data from AY2014-15 are omitted from this analysis.

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; and $^{+}p < 0.10$ after Bonferroni adjustment for multiple hypothesis testing.

TABLE A.1.10

The Relative Association Between Teachers' Practice and Student Achievement Across Exams: Estimates Including TLF Scores Assigned by School Administrators

	Math			ELA		
TLF Overall	-0.028^{+}	0.040	-0.017	0.031**	0.098	-0.026^{+}
	(0.015)	(0.075)	(0.017)	(0.013)	(0.060)	(0.013)
Factor 1 *Instruction*	-0.052*	0.068	-0.060**	0.032*	0.041	0.003
	(0.019)	(0.051)	(0.021)	(0.014)	(0.058)	(0.015)
Factor 2 *Classroom Environment*	0.014	-0.013	0.039	0.009	0.138^{+}	-0.038**
	(0.018)	(0.080)	(0.021)	(0.013)	(0.068)	(0.013)
Student and teacher controls	X	X	X	X	X	X
Teacher FE		X			X	
Student FE			X			X
n	16,616	16,616	16,616	21,153	21,153	21,153

Notes. Student controls include gender, race, eligibility for free or reduced-price lunch, limited English proficiency status, and special education status; teacher controls include race, gender, education, and experience. This table shows results from two regressions within each estimation model: the first rows show the interacted effects of overall TLF scores, standardized within year, and the PARCC exam on student achievement; the following rows show interacted effects between the PARCC exam and the *instruction* and *classroom environment* domains. PARCC exam scores used for this analysis are linked to the CAS scale and distribution using propensity-score matching followed by an equipercentile transformation (approach 1 in appendix B); scores from each test are standardized within subject and grade relative to the distribution of CAS scores in the final year of the CAS exam. TLF scores are standardized within year and use scores assigned by both external (i.e., master educators) and internal (i.e., school administrators) evaluators. Data from AY2014-15 are omitted from this analysis. $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; and $^{+}p < 0.10$ after Bonferroni adjustment for multiple hypothesis testing.

TABLE A.1.11

Difference-in-Differences Estimates of PARCC Effects on Teachers' Practice: Robustness to Alternative Specifications and Sampling Decisions

Overall TLF	-0.147 ***	-0.020	0.011	0.022	-0.054	-0.023	-0.010
	(0.042)	(0.039)	(0.038)	(0.038)	(0.039)	(0.040)	(0.042)
Factor 1	-0.164 ***	-0.065 +	-0.054	-0.049	-0.133 ***	-0.128 ***	-0.161 ***
Instruction	(0.038)	(0.035)	(0.040)	(0.040)	(0.042)	(0.043)	(0.046)
Factor 2	-0.028	0.047	0.084 *	0.094 **	0.078 +	0.120 ***	0.180 ***
Classroom Environment	(0.040)	(0.039)	(0.040)	(0.040)	(0.041)	(0.041)	(0.044)
Control for Level of Departmentalization		X	X	X	X	X	X
Teacher Controls		X	X	X	X	X	X
School FE				X			
Teacher FE			X	X	X	X	X
Rater FE		X	X	X			
Exclude AY2009-10						X	
Exclude AY2009-10 - AY2011-12							X
Master Educators Only	X	X	X	X		X	X
Master Educators and Administrators					X		
n	22,785	22,785	22,785	22,785	22,785	18,891	15,374

Notes. The outcome variable is the TLF score assigned by master educators (MEs), standardized relative to the overall mean and standard deviation of ME-assigned TLF scores across the years of analysis (AY2010-AY2016). Teacher controls include education level, race, gender, and experience. Robust standard errors, clustered at the teacher level, are in parentheses.

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; and $^{+}$ $p < 0.10$

TABLE A.1.12

CITS Estimates of PARCC Effects on Teachers' Practice

Overall TLF	Group 1 * PARCC	0.085	-0.111
	(2015 Effect)	(0.058)	(0.095)
	Group 1 * PARCC * Year	-0.081	-0.286 ***
	(2016 Effect)	(0.062)	(0.105)
Factor 1 *Instruction*	Group 1 * PARCC	-0.132 *	-0.208 *
	(2015 Effect)	(0.063)	(0.100)
	Group 1 * PARCC * Year	-0.103	-0.205 $^{+}$
	(2016 Effect)	(0.066)	(0.112)
Factor 2 *Classroom Environment*	Group 1 * PARCC	0.292 ***	0.080
	(2015 Effect)	(0.060)	(0.099)
	Group 1 * PARCC * Year	-0.004	-0.201 $^{+}$
	(2016 Effect)	(0.061)	(0.106)
Control for Level of Departmentalization		X	X
Teacher Controls		X	X
School FE			
Teacher FE		X	X
Rater FE			
Quadratic of Year			X
n		22,785	22,785

Notes. The outcome variable is the TLF score assigned by master educators, standardized relative to ME-assigned TLF scores across the years of analysis (AY2010-AY2016). The year variable is centered at 2015, when PARCC was first administered in DCPS. Teacher controls include education level, race, gender, and experience. Robust standard errors, clustered at the teacher level, are in parentheses.

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; and $^{+}$ $p < 0.10$

TABLE A.1.13

DiD Estimates by TLF Sub-Score

Teach 1	-0.125 ***	-0.121 ***	-0.130 ***	-0.072
	(0.039)	(0.038)	(0.037)	(0.044)
Teach 2	-0.075 $^{+}$	-0.075 $^{+}$	-0.093 **	-0.039
	(0.040)	(0.039)	(0.037)	(0.045)
Teach 3	-0.071 $^{+}$	-0.074 $^{+}$	-0.073 $^{+}$	-0.021
	(0.039)	(0.038)	(0.037)	(0.047)
Teach 4	0.013	0.016	0.027	-0.039
	(0.038)	(0.038)	(0.036)	(0.041)
Teach 5	-0.049	-0.048	-0.058	-0.037
	(0.040)	(0.039)	(0.038)	(0.047)
Teach 6	-0.150 ***	-0.147 ***	-0.166 ***	-0.128 ***
	(0.041)	(0.040)	(0.039)	(0.046)
Teach 7	-0.109 ***	-0.111 ***	-0.112 ***	-0.072
	(0.040)	(0.039)	(0.038)	(0.047)
Teach 8	0.018	0.012	0.023	0.084 $^{+}$
	(0.042)	(0.042)	(0.040)	(0.044)
Teach 9	-0.063	-0.070	-0.077 $^{+}$	0.036
	(0.044)	(0.043)	(0.041)	(0.044)
Control for Level of Departmentalization	X	X	X	X
Teacher Controls		X	X	X
School FE			X	
Teacher FE				X
Rater FE				
n	22,785	22,785	22,785	22,785

Notes. The outcome variable is for each TLF sub-score is standardized relative to ME-assigned TLF scores across the years of analysis (AY2010-AY2016). Teacher controls include education level, race, gender, and experience. Robust standard errors, clustered at the teacher level, are in parentheses.

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; and $^{+}$ $p < 0.10$

TABLE A.2.1

Covariate Balance at the ME/D Threshold

Female	-0.042
	(0.048)
Black	0.005
	(0.050)
White	-0.024
	(0.039)
Hispanic	-0.024
	(0.022)
Graduate Degree	-0.030
	(0.052)
Experience: 0-1 years	0.073
	(0.045)
Experience: 2-4 years	-0.043
	(0.041)
Experience: 5-9 years	-0.070 $^{+}$
	(0.042)
Experience: 10-15 years	0.034
	(0.034)
Experience: 15-19 years	-0.015
	(0.027)
Experience: Missing	-0.025
	(0.016)
Group 1	0.009
	(0.046)

Notes. Coefficients are estimated using the full ME/D bandwidth (h=50), regressing teacher characteristics on intent-to-treat, the centered IMPACT score, and their interaction, with year and school fixed effects. Robust standard errors in parentheses. n=1,809.
*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; and $^{+}$ $p < 0.10$

TABLE A.2.2

Estimated Treatment Effects at Placebo Cut Points

Retention

Placebo Cut Score	RD Estimate	SE
-30	-0.114	(0.128)
-20	0.067	(0.098)
-10	0.012	(0.052)
0	0.049	(0.047) *
10	-0.069	(0.104)
20	-0.157	(0.066)
30	-0.030	(0.044)

Next-Year IMPACT Score

Placebo Cut Score	RD Estimate	SE
-30	-14.67	18.01
-20	-1.49	12.76
-10	-0.30	6.42
0	4.35	5.24
10	-3.58	12.14
20	11.09	8.64
30	6.72	5.08

Notes. Estimates from a single RD regression estimating treatment effects for each placebo cut score (where 0 is the true treatment threshold) across the full bandwidth, with uniform kernel weights. Regressions include year fixed effects.

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; and $^{+}$ $p < 0.10$

TABLE A.2.3

Selection of Optimal Bandwidth

	Retention			Next-Year IMPACT Score		
	Optimal Bandwidth		Corresponding RD Estimate	Optimal Bandwidth		Corresponding RD Estimate
	left	right		left	right	
Mean Squared Error (MSE) Minimization						
Same bandwidth on each side						
Uniform kernel weight	9.25	9.25	-0.211 *	10.22	10.22	10.18
			(0.107)			(13.73)
			256			*180*
Triangular kernel weight	12.21	12.21	-0.240 *	15.83	15.83	11.13
			(0.105)			(12.10)
			264			*323*
Allowing for different bandwidths on either side						
Uniform kernel weight	9.37	10.50	-0.210 *	20.22	9.64	10.83
			(0.103)			(12.59)
			353			*274*
Triangular kernel weight	11.70	12.35	-0.212 *	24.31	15.95	10.32
			(0.104)			(11.64)
			335			*435*

Notes. The MSE method selects the bandwidth (*h*) that balances squared bias and variance to minimize the asymptotic approximation to the mean-squared error of regression discontinuity point estimator. Optimal bandwidths are estimated using Cattaneo et al.'s *rdplot* Stata program (for explanations of the MSE-optimization methods, see Cattaneo, Idrobo, & Titiunik, 2018a) with year fixed effects. Robust standard errors are in parentheses, and sample sizes in italics. $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; and $^{+}p < 0.10$

TABLE A.3.1

Test for Attrition Bias, by Year and Observation Window

	By Year			By Observation Window		
	(A)			(C)		
	Association between attrition and TLF scores in time *t*					
Attrition	-0.094 **	-0.095 **	-0.101 **	-0.091 **	-0.113 ***	-0.105 **
	(0.034)	(0.034)	(0.034)	(0.032)	(0.033)	(0.034)
Lagged TLF	-0.168 ***	-0.169 ***	-0.189 ***	-0.090 ***	-0.091 ***	-0.101 ***
	(0.015)	(0.015)	(0.015)	(0.011)	(0.011)	(0.011)
Teacher FE	X	X		X	X	
Teacher-by-school FE			X			X
Year FE		X	X		X	X
Sample Size	8,398	8,398	8,398	14,505	14,505	14,505
	(B)			(D)		
	Association between attrition and TLF scores in time *t - 1*					
Attrition	0.014	0.022	0.009	0.119 *	0.096 *	0.074
	(0.046)	(0.047)	(0.047)	(0.048)	(0.049)	(0.050)
Twice-lagged TLF	-0.226 ***	-0.228 ***	-0.240 ***	-0.147 ***	-0.151 ***	-0.163 ***
	(0.019)	(0.019)	(0.019)	(0.015)	(0.015)	(0.016)
Teacher FE	X	X		X	X	
Teacher-by-school FE			X			X
Year FE		X	X		X	X
Sample Size	4,626	4,626	4,626	7,028	7,028	7,028

Notes. All models include experience fixed effects. Data in the first three columns are reported at the teacher-by-year level and the remaining columns are at the teacher-by-observation-by-year level. Coefficients are standardized relative to the overall mean and standard deviation of master-educator-assigned TLF scores. Robust standard errors are in parentheses.

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; and $^{+}p < 0.10$

TABLE A.3.2

Factor Loadings from Exploratory Factor Analysis of the TLF

TLF Domain	Factor 1 *Classroom Environment & Lesson Accessibility*	Factor 3 *Instructional Clarity & Student Understanding*	Uniqueness
Teach 1	0.00	**0.84**	0.30
Teach 2	0.49	**0.60**	0.40
Teach 3	**0.83**	0.12	0.29
Teach 4	**0.51**	**0.53**	0.46
Teach 5	**0.50**	**0.61**	0.38
Teach 6	**0.67**	0.33	0.45
Teach 7	**0.58**	**0.44**	0.47
Teach 8	**0.76**	0.12	0.41
Teach 9	**0.78**	0.22	0.35

Notes. *n*=1,348 observations from external ("Master Educator") evaluators in DCPS for first-year teachers who received full evaluations in AY 2009-10 through AY 2011-12. Varimax-rotated factor loadings of 0.40 or higher are in bold. Results are from a principal-component factor analysis in which the TLF is forced to load onto more than one factor. The first factor has an eigenvalue of 4.56 and explains 51% of the variance in TLF scores; the second factor has an eigenvalue of 0.92 and explains an additional 10% of the variance.

TABLE A.3.3

Robustness of Estimates to Alternative Modeling Approaches

		Censored Growth Model with Equal Teacher Weights				Discontinuous Career Model		
		(1)	(2)	(3)		(1)	(2)	(3)
Experience	1	0.416***	0.414***	0.420***	1	0.399***	0.396***	0.398***
		(0.029)	(0.029)	(0.029)		(0.031)	(0.031)	(0.031)
	2	0.612***	0.598***	0.606***	2	0.585***	0.576***	0.583***
		(0.040)	(0.040)	(0.041)		(0.039)	(0.039)	(0.040)
	3	0.757***	0.737***	0.744***	3	0.680***	0.670***	0.681***
		(0.052)	(0.052)	(0.053)		(0.048)	(0.048)	(0.048)
	4	0.930***	0.909***	0.920***	4	0.811***	0.799***	0.813***
		(0.065)	(0.065)	(0.065)		(0.055)	(0.055)	(0.055)
Teacher FE		X	X	X		X	X	X
Year FE		X	X	X		X	X	X
School-level student characteristics			X	X			X	X
School FE				X				X
Sample Size		10,399	10,399	10,399		16,580	16,580	16,580
		Two-Stage Model				Indicator Variable Model		
		(1)	(2)	(3)		(1)	(2)	(3)
Experience	1	0.379***	0.388***	0.397***				
		(0.035)	(0.037)	(0.038)				
	2	0.558***	0.565***	0.592***	1-2	0.401***	0.398***	0.392***
		(0.044)	(0.049)	(0.052)		(0.037)	(0.037)	(0.038)
	3	0.654***	0.666***	0.706***				
		(0.054)	(0.061)	(0.066)				
	4	0.805***	0.823***	0.887***	3-4	0.559***	0.553***	0.547***
		(0.063)	(0.074)	(0.081)		(0.059)	(0.059)	(0.060)
Teacher FE		X	X	X		X	X	X
Year FE		X	X	X		X	X	X
School-level student characteristics			X	X			X	X
School FE				X				X
Sample Size		10,354	10,354	10,354		10,399	10,399	10,399

Notes. Data are teacher-by-year, using master-educator-assigned evaluation scores only. TLF scores are standardized relative to DCPS teachers' entry-year achievement. The censored growth model uses value E=15; the indicator variable model uses indicators for experience equal to 1-2, 3-4, 5-9, 10-14, 15-14, and 25+. Robust standard errors are in parentheses.

$^{***}\,p < 0.001$; $^{**}\,p < 0.01$; $^{*}\,p < 0.05$; and $^{+}\,p < 0.10$

TABLE A.3.4

Estimates of Returns to Experience from Censored Growth Models, Using All Evaluators' Scores

		(1)	(2)	(3)
Experience	1	0.538 ***	0.537 ***	0.540 ***
		(0.035)	(0.035)	(0.035)
	2	0.798 ***	0.785 ***	0.798 ***
		(0.049)	(0.049)	(0.050)
	3	0.964 ***	0.945 ***	0.959 ***
		(0.064)	(0.064)	(0.065)
	4	1.173 ***	1.153 ***	1.179 ***
		(0.078)	(0.078)	(0.079)
Teacher FE		X	X	X
Year FE		X	X	X
School-level student characteristics			X	X
School FE				X
Sample Size		10,399	10,399	10,399

Notes. Data are teacher-by-year, using master-educator- and administrator-assigned evaluation scores only. Units are standardized such that each coefficient estimate represents SD gains on the TLF relative to ME-assigned scores at experience=0. Models are censored at *E*=15. Robust standard errors are in parentheses.

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; and $^{+}p < 0.10$

TABLE A.3.5

Estimated Returns to Experience, by Analytic Sample and Subskill

		Environment	Clarity	Teach 1	Teach 2	Teach 3	Teach 4	Teach 5	Teach 6	Teach 7	Teach 8	Teach 9
		RESTRICTED SAMPLE										
Experience	1	0.577 *	0.291	0.200	0.618 *	0.283	0.522 *	0.365	0.689 **	0.259	0.525 *	0.472 $^{+}$
		(0.242)	(0.232)	(0.245)	(0.285)	(0.289)	(0.237)	(0.250)	(0.256)	(0.301)	(0.246)	(0.287)
	2	0.754 $^{+}$	0.356	0.142	0.903 $^{+}$	0.304	0.730 $^{+}$	0.598	0.854 $^{+}$	0.121	0.616	0.760
		(0.439)	(0.416)	(0.443)	(0.514)	(0.538)	(0.428)	(0.448)	(0.442)	(0.543)	(0.430)	(0.520)
	3	0.971	0.361	0.043	1.093	0.379	0.860	0.851	1.215 $^{+}$	-0.003	0.726	0.865
		(0.653)	(0.612)	(0.656)	(0.760)	(0.804)	(0.631)	(0.659)	(0.648)	(0.802)	(0.637)	(0.777)
	4	1.014	0.308	-0.130	1.258	0.230	0.871	0.746	1.563 $^{+}$	-0.050	0.613	0.889
		(0.846)	(0.788)	(0.848)	(0.982)	(1.050)	(0.816)	(0.851)	(0.830)	(1.037)	(0.815)	(0.996)
		FULL SAMPLE										
Experience	1	0.388 ***	0.223 ***	0.262 ***	0.314 ***	0.379 ***	0.280 ***	0.253 ***	0.386 ***	0.239 ***	0.309 ***	0.322 ***
		(0.035)	(0.036)	(0.038)	(0.040)	(0.037)	(0.036)	(0.038)	(0.041)	(0.039)	(0.036)	(0.037)
	2	0.525 ***	0.366 ***	0.385 ***	0.488 ***	0.504 ***	0.436 ***	0.406 ***	0.487 ***	0.350 ***	0.412 ***	0.503 ***
		(0.048)	(0.051)	(0.052)	(0.056)	(0.052)	(0.051)	(0.052)	(0.057)	(0.056)	(0.048)	(0.050)
	3	0.579 ***	0.489 ***	0.548 ***	0.596 ***	0.612 ***	0.443 ***	0.484 ***	0.598 ***	0.422 ***	0.480 ***	0.557 ***
		(0.061)	(0.066)	(0.069)	(0.072)	(0.067)	(0.066)	(0.067)	(0.073)	(0.071)	(0.061)	(0.065)
	4	0.734 ***	0.557 ***	0.631 ***	0.656 ***	0.736 ***	0.528 ***	0.564 ***	0.758 ***	0.590 ***	0.553 ***	0.725 ***
		(0.075)	(0.081)	(0.084)	(0.089)	(0.082)	(0.081)	(0.082)	(0.090)	(0.087)	(0.074)	(0.080)

Notes. Data are teacher-by-year, using master-educator-assigned evaluation scores only. Estimates from a censored growth model, using *E*=5 (restricted sample) and *E*=15 (full sample), with controls for teacher and year fixed effects, and school-average student characteristics. The full sample consists of 3,407 teachers with at least two years of continuous experience in DCPS between 2009-10 and 2015-16 who have scores in each year from master educators and school administrators and were not involuntarily separated due to their IMPACT scores (teacher-by-year *n*=10,399) and the restricted sample further requires teachers to be observed for at least five continuous years from their initial year teaching (*n* of unique teachers = 120; teacher-by-year *n*=677). Units are standard deviations of first-year teachers' overall ME-assigned TLF scores. Robust standard errors are in parentheses.

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; and $^{+}p < 0.10$

TABLE A.3.6

The Relationship Between Student Achievement and Changes in Teachers' Practice, by Sub-Score

	Math			Reading		
Teach 1	0.048 **	0.044 *	0.040 *	0.024 +	0.042 **	0.042 **
	(0.016)	(0.018)	(0.019)	(0.013)	(0.014)	(0.015)
	{0.003}	{0.014}	{0.034}	{0.067}	{0.003}	{0.005}
Teach 2	0.034 *	0.029	0.022	0.004	0.013	0.009
	(0.017)	(0.018)	(0.018)	(0.017)	(0.019)	(0.020)
	{0.043}	{0.105}	{0.224}	{0.839}	{0.506}	{0.647}
Teach 3	0.017	0.014	0.013	-0.007	-0.006	-0.006
	(0.017)	(0.017)	(0.018)	(0.016)	(0.016)	(0.017)
	{0.325}	{0.416}	{0.471}	{0.686}	{0.717}	{0.725}
Teach 4	0.021	0.015	0.011	0.005	0.013	0.015
	(0.017)	(0.020)	(0.021)	(0.018)	(0.019)	(0.020)
	{0.221}	{0.455}	{0.608}	{0.766}	{0.488}	{0.468}
Teach 5	0.018	0.011	0.013	0.003	0.016	0.017
	(0.016)	(0.018)	(0.019)	(0.016)	(0.018)	(0.019)
	{0.255}	{0.556}	{0.495}	{0.864}	{0.379}	{0.350}
Teach 6	0.009	0.001	-0.005	-0.010	-0.005	-0.004
	(0.018)	(0.018)	(0.019)	(0.016)	(0.017)	(0.018)
	{0.619}	{0.940}	{0.801}	{0.534}	{0.755}	{0.809}
Teach 7	0.035 *	0.032 +	0.031 +	-0.002	0.001	-0.001
	(0.017)	(0.018)	(0.018)	(0.018)	(0.018)	(0.019)
	{0.043}	{0.071}	{0.089}	{0.914}	{0.964}	{0.954}
Teach 8	0.023	0.019	0.017	-0.005	0.000	-0.006
	(0.015)	(0.016)	(0.017)	(0.015)	(0.016)	(0.016)
	{0.125}	{0.231}	{0.294}	{0.738}	{0.987}	{0.725}
Teach 9	0.015	0.006	0.001	-0.001	0.012	0.012
	(0.018)	(0.020)	(0.021)	(0.018)	(0.022)	(0.023)
	{0.414}	{0.756}	{0.981}	{0.945}	{0.580}	{0.608}
Teacher FE	X	X	X	X	X	X
Experience		X	X		X	X
School FE			X			X

Notes. Outcomes are averaged at the teacher level, after residualizing using a vector of student characteristics including race/ethnicity, gender, lagged absences and achievement, poverty status, special education status, grade level, and indicators for limited English proficiency and whether the student is in a new school. TLF scores are standardized relative to the distribution of all DCPS teachers' overall master-educator (ME)-assigned TLF scores. Point estimates for each subdomain are from separate regressions. Robust standard errors are in parentheses; *p*-values are in brackets. $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; and $^{+}p < 0.10$

FIGURE A.2.1. *Density of Observations at the ME/D Threshold*

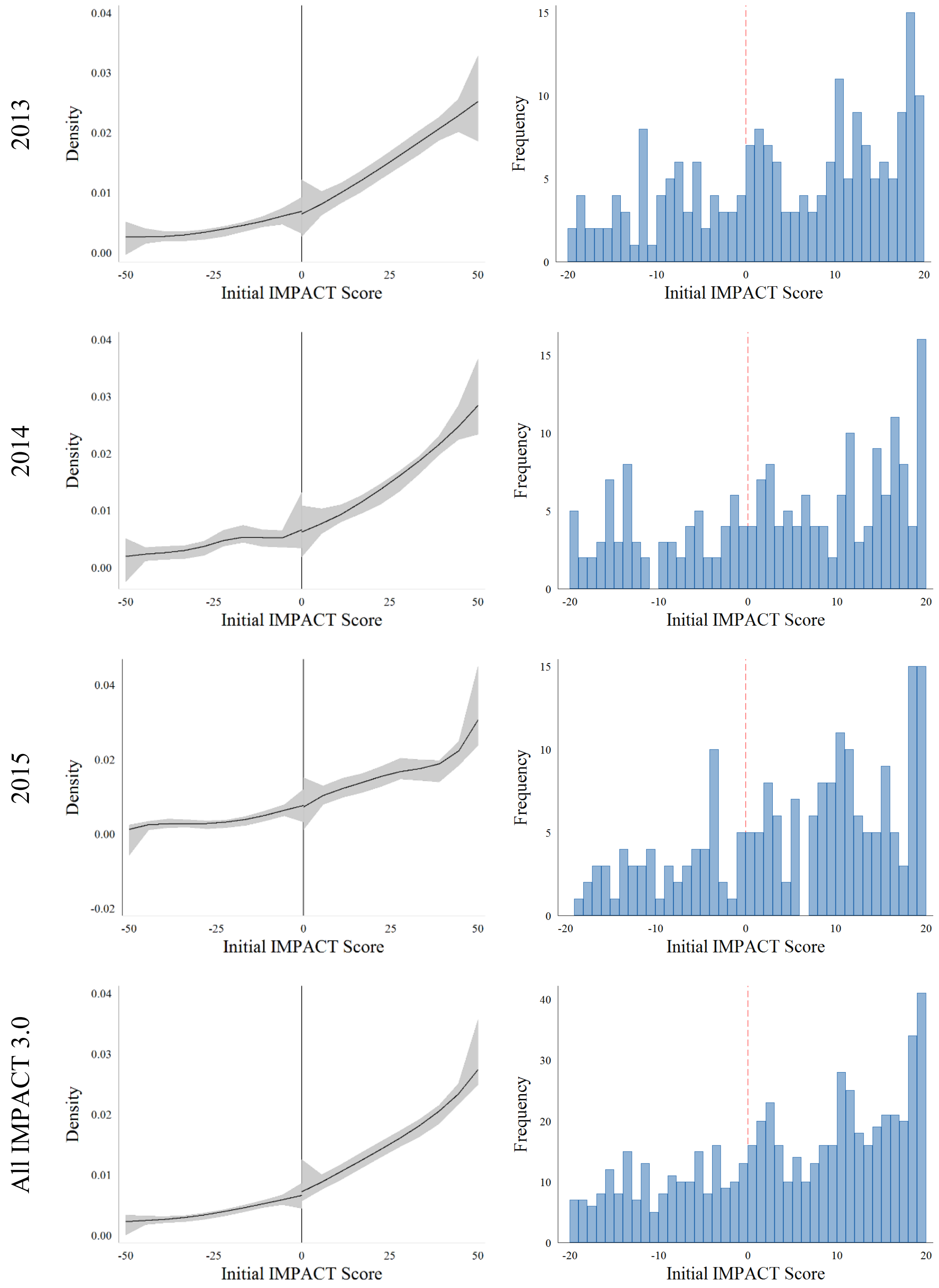
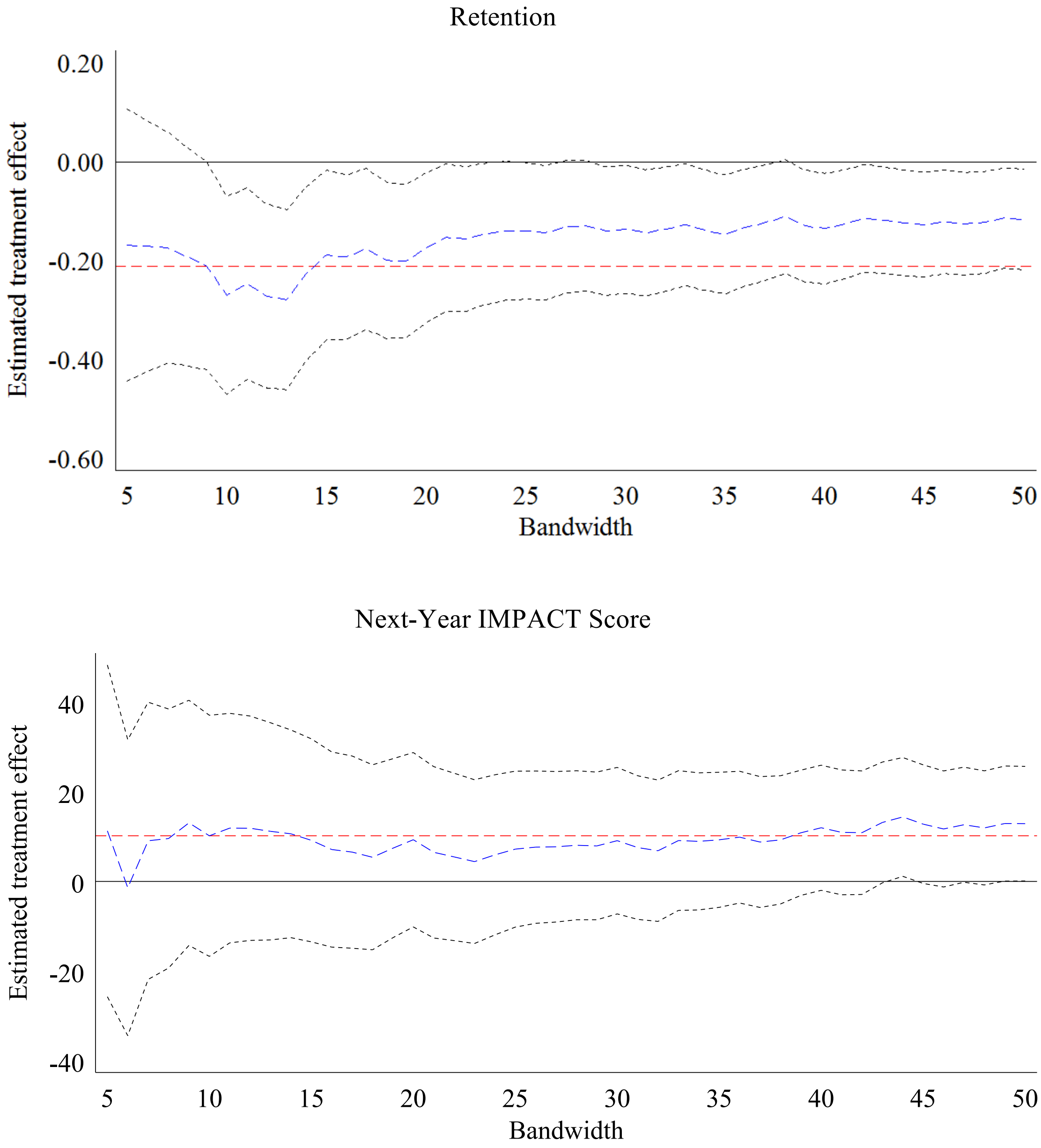
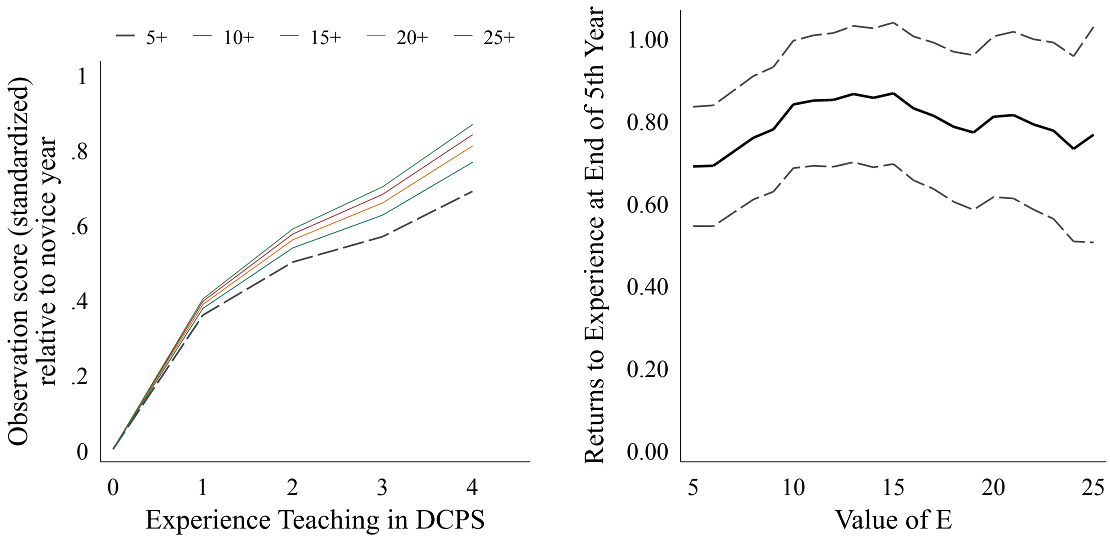


FIGURE A.2.2. *Local Linear Regressions with Varying Bandwidths*



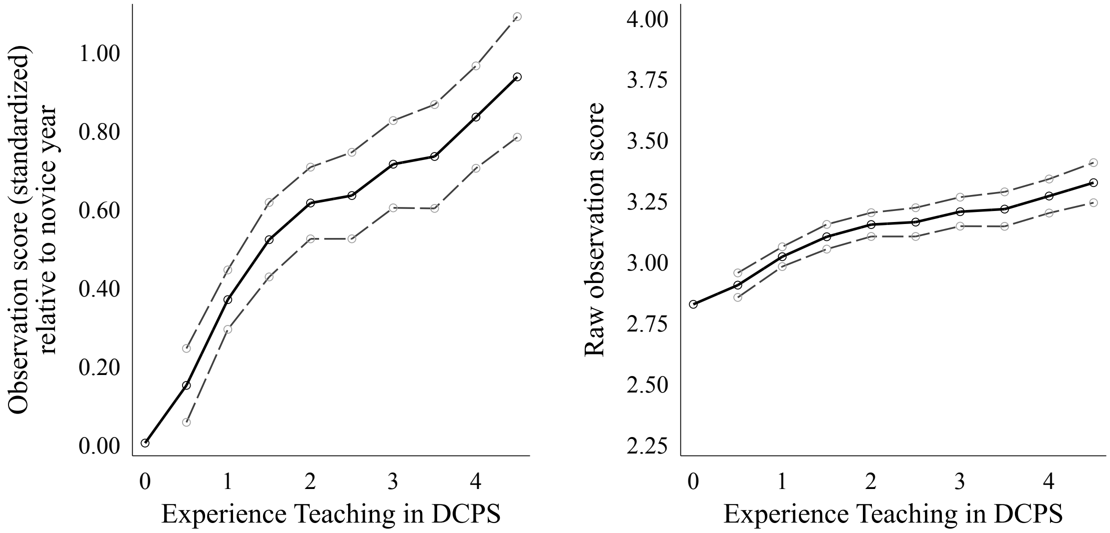
Notes. Estimates from a series of local linear regressions with varying bandwidths. Regressions use uniform kernel weights and include year fixed effects. The dashed red line represents the treatment estimate at the MSE-optimal bandwidth. The dashed blue line represents treatement effects at each of the plotted bandwidths, with dashed gray lines representing 95% confidence intervals.

FIGURE A.3.1. *Returns to Experience Estimated Using Censored Growth Model, by Value of E*



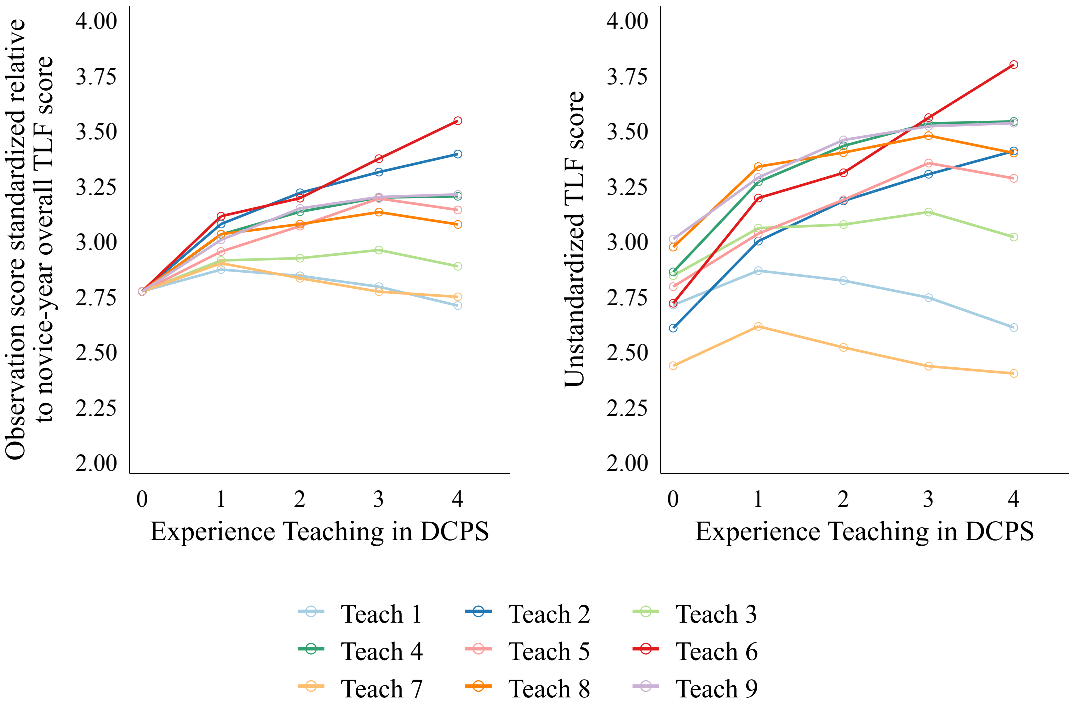
Notes. Point estimates and 95% confidence intervals for first five years obtained from fitting Equation 1 with differing values of *E* using the sample of 3,407 teachers with at least two years of continuous experience in DCPS between 2009-10 and 2015-16 who have scores in each year from master educators and school administrators and were not involuntarily separated due to their IMPACT scores. Regressions include teacher and year fixed effects, as well as school-averaged student characteristics. The dependent variable is the average teacher observation score assigned by master educators (MEs); scores are scaled to the distribution of novice teachers' overall ME-assigned TLF scores.

FIGURE A.3.2. *Returns to Experience for Novice Teachers' Practice, Estimated Within Year and Observation Window*



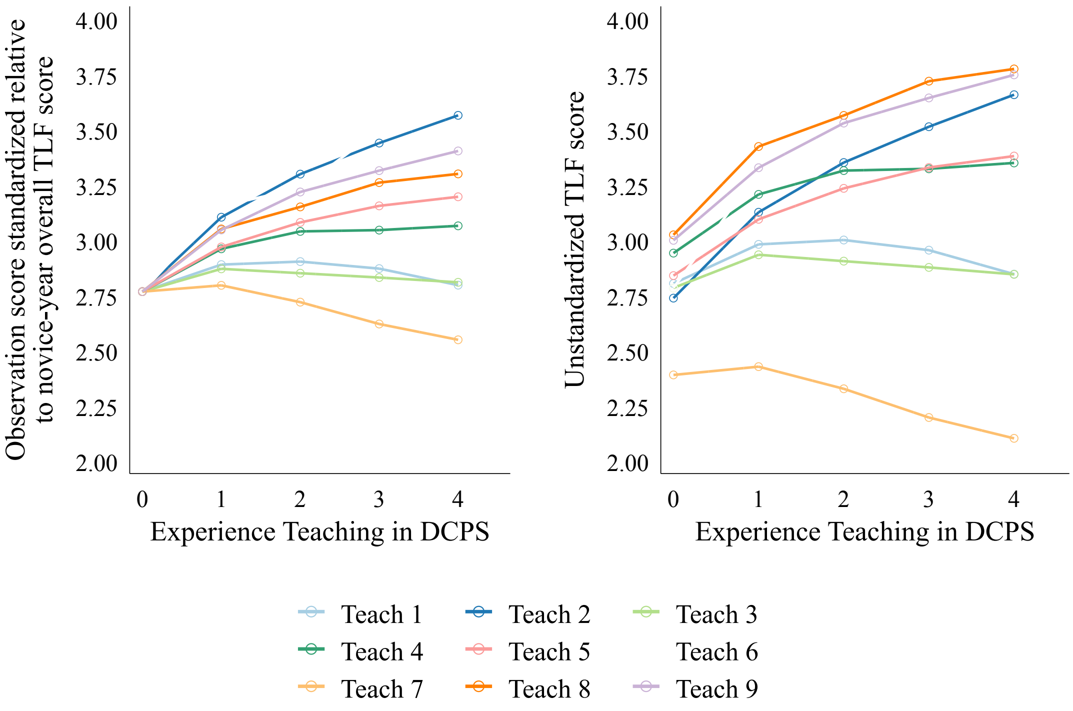
Notes. Point estimates and 95% confidence intervals for first five years obtained from fitting Equation 1 with $E = 15$ using the sample of 3,407 teachers with at least two years of continuous experience in DCPS between 2009-10 and 2015-16 who have scores in each year from master educators and school administrators and were not involuntarily separated due to their IMPACT scores. Regressions include teacher and year fixed effects, as well as school-averaged student characteristics. The dependent variable is the average teacher observation score; in the graph on the left, scores are scaled to the distribution of novice teachers' overall assigned TLF across both types of raters (master educators and school administrators); the right shows these scores in their raw (i.e., unstandardized) form. Note that each cycle-level slope does not necessarily include the same teachers; teachers can be subject to fewer evaluations as they advance on DCPS's career ladder.

FIGURE A.3.3. *Returns to Experience for Novice Teachers' Practice, by TLF Subdomain*



Notes. Point estimates and 95% confidence intervals for first five years obtained from fitting Equation 1 with $E = 5$ using the sample of 120 teachers who entered DCPS with no prior experience, continued teaching for at least five consecutive years between 2009-10 and 2015-16, have scores in each year from master educators and school administrators, and were not involuntarily separated due to their IMPACT scores. Regressions include teacher and year fixed effects, as well as school-averaged student characteristics. The dependent variable is the average ME-assigned teacher observation score; in the graph on the left, scores are scaled to the distribution of entry-year teachers' overall ME-assigned TLF; the right shows these scores in their raw (i.e., unstandardized) form.

FIGURE A.3.4. *Returns to Experience for Novice Teachers' Practice, by TLF Subdomain, Including Teachers Who Are Missing Scores for Teach 6 in Any of Their First Five Years in the Classroom*



Notes. Point estimates and 95% confidence intervals for first five years obtained from fitting Equation 1 with $E = 5$ using the sample of 302 teachers who entered DCPS with no prior experience, continued teaching for at least five consecutive years between 2009-10 and 2015-16, have scores in each year from master educators and school administrators, and were not involuntarily separated due to their IMPACT scores. Regressions include teacher and year fixed effects, as well as school-averaged student characteristics. The dependent variable is the average ME-assigned teacher observation score; in the graph on the left, scores are scaled to the distribution of entry-year teachers' overall ME-assigned TLF; the right shows these scores in their raw (i.e., unstandardized) form.

APPENDIX B

Linking Scores Across the CAS and PARCC Exams

LINKING SCORES ACROSS THE CAS AND PARCC EXAMS

Detailed descriptions of each proposed linking process used in Chapter 1 are described below, followed by a discussion of how to evaluate the quality of each score linkage and other considerations and potential drawbacks associated with each linking approach.

Propensity Score Matching

The PSM method attempts to establish pseudo-equivalent groups across the exams using observed student characteristics to create comparable testing samples. This approach requires first limiting the analytic sample to students tested in adjacent years (i.e., 2014 CAS and 2015 PARCC). Because I wish to link PARCC scores to the CAS distribution, I define the students tested in the CAS year as the treatment group and PARCC examinees as the control. To create matched samples, I first estimate a logistic regression to predict the probability ($\hat{p}_{sgi}$) of treatment (CAS_{sgi}) given a vector of pre-treatment student characteristics ($\boldsymbol{X}_{sgi}$), for student i, in subject s and grade g, where $\hat{p}_{sgi} = \Pr(CAS_{sgi} = 1 | \boldsymbol{X})$:

$$\ln\left(\frac{p_{sgi}}{1-p_{sgi}}\right) = \beta_0 + \boldsymbol{X}_{sgi}\beta_1 + \varepsilon_{sgi} \tag{B1}$$

In the DCPS administrative data, potential pre-treatment characteristics ($\boldsymbol{X}_{sgi}$) include race/ethnicity, gender, lagged achievement (applicable for grades 4 through 8 only) and lagged absences, eligibility for free or reduced-price lunch, special education status, ELL status, an indicator for whether a student is new to a school, and school fixed effects.

I then constrain the sample to the region of common support across estimated p-scores—the range of values of $\hat{p}_{sgi}$ within which both treatment and control cases are observed. Next, I divide the matched cases into a given number of equally-sized propensity strata (e.g., 10) that contain both treatment and control observations, randomly dropping treatment or control observations so that there are equal numbers of treatment and control cases within each stratum.

At this point I check for balance across each of the k covariates and across each stratum j:

$$X_{kjsgi} = \alpha_{kjsg} + \beta_{kjsg} CAS_{kjsgi} + \varepsilon_{kjsgi} \qquad \text{(B2)}$$

I also test balance using Cohen's d and variance ratio v estimates to confirm that these values are within acceptable ranges (Caliendo & Kopeinig, 2008; Steiner, Cook, Shadish, & Clark, 2010).

When there are more covariates—or covariates that are especially important for identifying differences across the two exams, such as prior ability as measured by the CAS exam—that exhibit imbalance than would be expected by chance, I iterate specifications of (B1) with interactions and higher-order polynomials, and use alternative matching models (e.g., nearest-neighbor or caliper matching in lieu of stratification, or different numbers of propensity strata). I likewise explore the robustness of results to the choice of covariates and to these matching decisions.

After iterating across specifications and methods, the method that best-produces balance across available student characteristics is the stratification approach using five propensity strata with controls for poverty, special education, and ELL status, as well as gender, race, whether the student is new to his or her school, lagged absences and its quadratic, and the interactions between poverty and gender, black and gender, and Hispanic and absences.

Once satisfied with the covariance balance across the treatment and control groups, I use an equipercentile equating approach to transform these scores, interpolating for any missing scores.[45] This method transforms each score in the treatment group to the score in the control group that corresponds to the same percentile rank, producing linked scores with a near-identical distribution to that of the linking exam. I then continuize the

[45] I choose an equipercentile transformation over a linear transformation method because, while the linear approach will produce scores with the same mean and standard deviation of the CAS distribution to which I am equating, it can also produce out-of-range scores—that is, scores that fall outside of the minimum (0) and maximum (99) possible on the CAS exam. The linear method is also sensitive to the abilities of the linking population and differences in difficulty across the two exams. By contrast, the equipercentile approach defines the transformation process in terms of students' relative position (i.e., percentile rank) in the test performance distribution. If the two exams produced distributions with the same shape, the linear and equipercentile linking functions would produce nearly identical results; the PARCC exam, however, produces a performance distribution with a decided right skew, while the CAS score distribution is slightly left-skewed.

distribution of linked scores by using a locally weighted regression of linked scores on their PARCC equivalent score, with a bandwidth of 0.10, to smooth out noise from equivalent-score estimates.

Finally, I return to the full sample of examinees and use the translation defined in the match samples for each subject and grade to adjust all students' PARCC scores in 2015 and 2016 to the CAS scale.

Using NAEP as a Benchmark

This approach first requires linearly interpolating NAEP means and standard deviations from grades 4 and 8, which are commonly-tested grades across NAEP, PARCC, and CAS, to grades 5 through 7, and extrapolating to grade 3; these are each grades that are tested on the PARCC and CAS exams, but not on NAEP. For example, for the mean district-level NAEP score ($\hat{\mu}_{sgt}^{NAEP}$) in subject s, grade g, and year t:

$$\hat{\mu}_{sgt}^{NAEP} = \hat{\mu}_{s4t}^{NAEP} + \frac{g-4}{4}\left(\hat{\mu}_{s8t}^{NAEP} - \hat{\mu}_{s4t}^{NAEP}\right), \text{ for } g \ \in \{3, 5, 6, 7\} \tag{B3}$$

$$\hat{\sigma}_{sgt}^{NAEP} = \hat{\sigma}_{s4t}^{NAEP} + \frac{g-4}{4}\left(\hat{\sigma}_{s8t}^{NAEP} - \hat{\sigma}_{s4t}^{NAEP}\right), \text{ for } g \ \in \{3, 5, 6, 7\} \tag{B4}$$

I then interpolate NAEP scores for each grade and subject in even (non-tested) years using score distributions (means and standard deviations) from odd years, as during the sample period the NAEP exam was only administered in odd years:

$$\hat{\mu}_{sgt}^{NAEP} = \frac{1}{2}\left(\hat{\mu}_{sg[t-1]}^{NAEP} - \hat{\mu}_{sg[t+1]}^{NAEP}\right), \text{ for } t \ \in \{2010, 2012, 2014, 2016\} \tag{B5}$$

$$\hat{\sigma}_{sgt}^{NAEP} = \frac{1}{2}\left(\hat{\sigma}_{sg[t-1]}^{NAEP} - \hat{\sigma}_{sg[t+1]}^{NAEP}\right), \text{ for } t \ \in \{2010, 2012, 2014, 2016\} \tag{B6}$$

Finally, I linearly transform each subject-year-grade score ($Y_{sgt}^{PARCC/CAS}$) to its corresponding score in the NAEP distribution ($\hat{Y}_{sgt}^{NAEP}$), adjusting for measurement error in the CAS and PARCC exams using subject-grade reliability estimates (see table B1) for the given test ($\hat{\rho}_{sgt}^{CAS/PARCC}$):

$$\hat{Y}_{sgt}^{NAEP} = \hat{\mu}_{sgt}^{NAEP} + \frac{Y_{sgt}^{CAS|PARCC}}{\hat{\rho}_{sgt}^{CAS|PARCC}} * \hat{\sigma}_{sgt}^{NAEP} \tag{B7}$$

Predicting CAS Achievement

This approach attempts to leverage within-student variation in achievement on the CAS to estimate the distribution of expected performance on the PARCC exam. Specifically, for each subject s, I separately regress students' CAS achievement in a given grade g and year t on their lagged performance on that exam with controls for a vector of observed student characteristics:

$$CAS_{msgti} = \beta_0 + \beta_1 CAS_{msg[t-1]i} + \boldsymbol{X}_{msgti}\beta_2 + \tau_m + \lambda_g + \varepsilon_{msgti} \qquad \text{(B8)}$$

where CAS_{msgti} is student i's CAS exam score in subject s and grade g, in year t; $\boldsymbol{X}_{sgti}$ is a vector of student covariates including a dummy for whether the student is retaking the same subject-grade test as in the lagged $[t-1]$ year; τ_m is a school fixed effect; λ_g is a grade fixed effect; and ε_{sgti} is an idiosyncratic error term.

Using the coefficients from (B8), I then estimate predicted scores on the CAS exam for the 2015 year for each student for whom lagged achievement data are available, had the test not changed to the PARCC exam:

$$\widehat{CAS}_{msg2015i} = \hat{\beta}_0 + \hat{\beta}_1 CAS_{msg2014i} + \boldsymbol{X}_{msg2015i}\hat{\beta}_2 + \hat{\tau}_m + \hat{\lambda}_g \qquad \text{(B9)}$$

I then create a crosswalk linking each observed PARCC score value Y_{sgi}^{PARCC} to the average corresponding value of $\widehat{CAS}_{sg2015i}$, which can then be used to translate the scores for all PARCC examinees (except in grade 3, where lagged scores are not available) to a CAS-linked scale.

This predicted-score approach, however, will further attenuate estimates of $\widehat{CAS}_{sg2015i}$—already a concern for the symmetry of the linking function, as discussed in the following section.

Evaluating the Quality of Linkages

Each approach is not immune to drawbacks and limitations. Dorans and Holland (2000) identify five rules that are generally considered necessary for equating tests. While the requirements for linking exams are less strict than for equating exams, violations of any of these assumptions will limit the validity of the test linkage. Each is of potential concern in this context, and is discussed individually.

Assumption 1: The exams should measure similar constructs.

In theory, this requirement should be met given that both exams purport alignment to the CCSS, though the test specifications differ considerably across exams, and the ways in which the constructs are operationalized likely affects the constructs truly being assessed. For example, PARCC and CAS have clearly distinct test formats[46], and these exams are delivered under different stakes, as teachers' value-added was not estimated for accountability purposes during the first two years of the PARCC exam in DCPS. This is also of concern for the method in which I use the NAEP exam as a moderator for linking PARCC and CAS scores, as NAEP likewise is built to different specifications than either CAS or PARCC.

There is no fool-proof test of construct similarity, but there are several statistical tests that can reveal potential violations of this assumption. One is to conduct simple correlational tests for convergent validity across the two exams (Dorans & Holland, 2000); if students' scores are not highly correlated across the tests it would suggest important differences in the constructs being assessed. Because the tests are administered in different years, however, I cannot correlate the two exams at the same time point. Instead, my best option is to correlate PARCC scores in 2015 to lagged CAS scores from 2014. If these cross-year correlations differ from those of CAS 2014 to CAS 2013 correlations, this would raise a red flag about the consistency of constructs across the tests.

Within the CAS exam, the 2014 scores are correlated with 2013 scores at $r = .80$ in both math and ELA; within the PARCC exam, 2016 scores are correlated with 2015 scores at $r = .85$ in each subject. In contrast, the PARCC-to-CAS correlations are lower, but only slightly so: $r = .77$ in math and $r = .79$ in ELA. These differences are not sufficiently large to raise concerns about construct similarity.

A second, post-hoc, test of the same-construct requirement would be to assess whether the two exams similarly sort subgroups of examinees by performance. This is

[46] One key distinction is that the CAS exam was delivered as a booklet-style paper and pencil test, while the PARCC exam was administered largely online. The transition to online testing was not without hiccups for DC and other school districts, with many low-income students with limited computer access reportedly struggling to adapt to the computer-based format of the PARCC exam. Evidence about these differences is mixed. An analysis by PARCC officials determined that while there were performance differences between students who had taken the exam online and those who took a paper and pencil version, these differences were likely attributable to variation in the population of students taking the respective versions of the exam; meanwhile, there may also be "mode" effects in certain states and districts, in which exam scores were capturing students' computer skills in addition to intended constructs in math and reading (Brown, 2016; Herold, 2016).

similar to the fifth assumption, population invariance, and is discussed in detail in the respective section below on differences in the equating function across subpopulations.

Assumption 2: The exams should have equal, and high, reliability.

This assumption is generally, though imperfectly met. Both exams have high reliability—exceeding 0.90 across grades, subjects and years; reliability coefficients are similar but not identical across the exams (see appendix table B.1).

Assumption 3: The equating function for converting scores from test A to test B should work inversely to equate test B to A.

This assumption is by definition met for the PSM and NAEP approaches, because each produces scores that are symmetrical; they can be used interchangeably, and rely on functions for which the inverse will recover the original score. The within-student regression approach, however, will likely violate this assumption, given that the regression will attenuate predicted scores due to measurement error in the test (Otis, 1922; Thorndike, 1922). Solving for $CAS_{msg2014i}$ in (B9) will not recover the same pre-linked scores; predicted CAS scores will be biased downward for high-performing students relative to their “true” CAS-aligned PARCC score, while predicted scores will be biased upward for low-performing students.

Assumption 4: It should not matter in terms of equated (linked) scores for a given student to have been tested under one test relative to the other.

This assumption, also known as Lord’s (1980) equity property of equating, states that a student tested under one exam should expect to receive the same score on the linked exam, such that the distribution of scores for an identical set of examinees on a given assessment would be identical to that of their equated scores on a different exam. This condition is nearly impossible to meet in practice, but Morris’s (1982) first- and second-order equity assumptions are relatively more feasible. These conditions are, respectively, that—after equating—the two exams produce similar expected scale scores, conditional on ability, and that the two forms’ scores produce similar standard errors of measurement.

Typically, adherence to these conditions is evaluated using item-response-theory-produced estimates of examinee ability (θ)[47], but given that I lack item-level response data, I must rely on methods for assessing equity that use only total scaled scores, substituting observed, pre-treatment characteristics for θ. I include lagged achievement in my proxy for θ, and so I do not test this assumption for grade 3. The difference in expected scale scores, conditional on ability, (D_1) should be statistically no different from 0, and the ratio of error variances of measurement (D_2) conditional on ability should be statistically no different from 1. The formulas for estimating D_1 and D_2 are below, where: SC_{CAS} is the scaled CAS exam score; SC_{PARCC} is the PARCC exam score that has been transformed through one of the linkage processes described above to the CAS scale; w_x is the weight of θ_x; and EV is the error variance of a given exam.

$$D_1 = \frac{\sum_x w_x[E(SC_{CAS}|\theta_x) - E(SC_{PARCC}|\theta_x)]}{\sum_x w_x} \qquad \text{(B10)}$$

$$D_2 = \sqrt{\frac{\sum_x w_x[(EV_{CAS}|x) - (EV_{PARCC}|\theta_x)]}{\sum_x w_x}} \qquad \text{(B11)}$$

I use a bootstrapping procedure to estimate confidence intervals for the first- and second-order estimates, drawing 1,000 independent random sub-samples from the score-linking samples used across each of the proposed linking methods. These test statistics are presented in appendix table B2; values are highlighted for the method that minimizes estimates of $D_1$and D_2. While none of these linking methods yields ideal point estimates (all D_1statistics are statistically greater than 0 and all D_2 estimates are greater than 1), propensity-score-matching with equipercentile linking performs the best of the three methods for first-order equity. This method also performs somewhat better in ELA than in math and for students in lower grade levels than for those in the upper grades used for this analysis. The PSM approach is also generally higher-performing for the second-order equity assumption, but in some ELA grades the predicted-CAS-score approach yields the lowest D_2 estimates.

Assumption 5: The equating function should not differ across subpopulations.

I test the fifth (population invariance) assumption by comparing linking results estimated separately for each of a given set of subpopulations to those estimated across

[47] See, for instance: Kim & DeCarlo, 2016; Tong & Kolen, 2005; and Lee, Lee, & Brennan, 2010.

the full population. Ideally, the linking functions used for individual subgroups of students should be similar to each other (Feuer, Holland, Green, Bertenthal, & Hemphill, 1999). If I find that the population variance assumption is not met for a given subgroup of students, it may merit establishing separate concordance functions for the respective subgroup(s).

While there are multiple ways in which one can estimate population invariance (see Huggins & Penfield, 2012), I do so by calculating the typical distance for a given set of subgroups' (e.g., male and female) linking functions from that of the overall population. The first measure of the population invariance that I estimate is the root mean square difference ($RMSD_x$), which estimates population invariance of the linking function for each pre-linked PARCC score (i.e., $Y_{sgt}^{PARCC} = x$). The second measure, the root expected mean square difference ($REMSD$) is unconditional on x, and represents overall invariance across the PARCC score distribution. For each mutually exclusive subgroup j relative to the overall population q, where $w_j$is a population weight for subgroup j, $d_j(x)$ is the difference in linked scores between that subgroup and population q at score x, $\sigma_q$is the overall population's standard deviation of CAS scores on which the PARCC exam is to be linked, and P_x is the proportion of examinees at score x:

$$RMSD_x = \frac{\sqrt{\sum_j w_j[d_j(x)]^2}}{\sigma_q} \qquad \text{(B12)}$$

$$REMSD = \frac{\sqrt{\sum_x P_x\{\sum_j w_j[d_j(x)]^2\}}}{\sigma_q} \qquad \text{(B13)}$$

I calculate $RMSD_x$ and $REMSD$ for each subject, grade, and linking method to compare the population invariance across each of the proposed linking approaches. Ideally, both values should be near zero, as larger values indicate population invariance of the linking function and perhaps that the exams are measuring different constructs across subgroups. A common bound for defining a "difference that matters" (DTM) when scores are reported in rounded, integer units, as is the case with DCPS's exams, is a half a point; any difference higher than that would result in a different score. I use the standardized DTM (STDM) to flag any linkages that result in $RMSD_x$ or $REMSD$ values that exceed conventionally accepted levels of population invariance.

$REMSD$ values for each grade level, subgroup, and linking method are presented in appendix table B.3. Here, too, none of the methods perform at ideal levels. Generally, the PSM approach yields lower $REMSD$ values than the NAEP or regression linking methods, yet it still produces $REMSD$ values that exceed the SDTM in all but one subgroup (ELL), and only in grades four and five ELA. Each method also performs somewhat worse at upper grade levels. $RMSD_x$ values—which assess population invariance across the scores distribution—indicate higher invariance in the tails of the score distribution; each method tends to produce $RMSD_x$ values below the SDTM for scores near the center of the score distribution, suggesting that low sample sizes or other statistical noise in the tails may be driving much of the observed population invariance.[48]

Additional Considerations

Each approach has potential drawbacks beyond those discussed above. For example, the PSM method is only as good as the observed covariates with which I am able to match samples across the exams. If there are important differences in examinees beyond what I can capture in X_{kjsgi}, the PSM process may still produce non-equivalent groups which could yield biased linkages. To some extent, these concerns can be assessed by the approaches discussed for assumptions 4 and 5 above.

The NAEP-as-moderator method likewise assumes that the samples of students taking the NAEP exam are randomly equivalent to the DCPS sample from which they are drawn. While NAEP intentionally selects its Trial Urban District Assessment (TUDA) samples to be representative of the given districts overall, I do not have a way to test for group equivalence between NAEP and each of the district-administered exams. One concern would be if there were non-random selection out of testing on one or both of the state exams. I test for this by estimating the probabilities of PARCC and CAS test-taking across subgroups of students and find that there is not a significant difference in the probability of participating in the two exams across subgroups, with the exception of students who qualify for free or reduced-price lunch. These students were slightly more likely to participate in the PARCC exam than CAS, although this difference may be confounded by co-occurring changes to DCPS's methods for identifying FRPL students

[48] $RMSD_x$ statistics are not shown, but available upon request.

over this period. I also compare differences in NAEP participation by subgroup across these transition years using publicly-reported TUDA subgroup participation data, finding no evidence to indicate changes in participation across these groups between the years in which I estimate score linkages.[49] The accuracy of the NAEP-as-moderator method is also contingent upon the accuracy of my assumption about the linearity of achievement gains, given that I am interpolating and extrapolating scores for non-tested grades and years.

Finally, beyond violating the symmetry assumption for linking, the most apparent drawback to the regression approach is that the predicted score distributions cannot be estimated for students who are missing lagged CAS scores or covariates from which to predict their place in the achievement distribution. This will exclude students tested in any PARCC year beyond 2015. This approach is likewise affected by the quality of covariates; failure to control for sufficient and appropriate predictors of student achievement will potentially bias linkages for certain types of students and will result in high measurement error in predicted scores.

[49] Links to participation data are available at https://www.nationsreportcard.gov/faq.aspx#q2.

TABLE B.1

Test Reliability by Exam, Grade, Subject, and Year

	CAS					PARCC	
	2010	2011	2012	2013	2014	2015	2016
			Reading / English Language Arts				
Grade 3	0.94	0.93	0.94	0.94	0.93	0.90	0.91
Grade 4	0.93	0.93	0.92	0.92	0.93	0.91	0.91
Grade 5	0.93	0.93	0.92	0.92	0.93	0.91	0.91
Grade 6	0.92	0.92	0.92	0.92	0.92	0.92	0.92
Grade 7	0.92	0.92	0.91	0.91	0.91	0.93	0.93
Grade 8	0.91	0.92	0.91	0.92	0.92	0.92	0.93
				Math			
Grade 3	0.94	0.94	0.94	0.93	0.93	0.94	0.93
Grade 4	0.93	0.93	0.93	0.94	0.94	0.94	0.93
Grade 5	0.92	0.93	0.93	0.94	0.94	0.93	0.92
Grade 6	0.94	0.94	0.94	0.93	0.94	0.94	0.93
Grade 7	0.93	0.92	0.92	0.92	0.92	0.93	0.92
Grade 8	0.91	0.92	0.92	0.91	0.93	0.91	0.91

TABLE B.2

Tests of First- and Second-Order Equity Assumptions, by Linking Method

		Math					
Linking method		PSM & equipercentile		NAEP-as-moderator		Predicted CAS	
Statistic	Grade	Estimate	95% CI	Estimate	95% CI	Estimate	95% CI
D_1: Mean difference in expected scale scores, conditional on ability	4	1.93	[1.23, 2.62]	6.66	[5.21, 8.1]	3.56	[2.76, 4.36]
	5	2.28	[1.56, 2.99]	6.82	[5.22, 8.43]	4.46	[3.49, 5.43]
	6	2.48	[1.48, 3.48]	6.34	[4.23, 8.46]	4.95	[3.92, 5.98]
	7	2.53	[1.68, 3.37]	8.57	[5.96, 11.18]	4.60	[3.71, 5.49]
	8	3.25	[1.86, 4.65]	12.31	[8.33, 16.29]	5.49	[4.32, 6.67]
	All	1.02	[0.68, 1.36]	4.32	[3.22, 5.41]	3.72	[3.29, 4.15]
D_2: Mean difference in conditional standard errors of measurement	4	5.44	[4.34, 6.53]	14.31	[12.35, 16.27]	7.15	[6.06, 8.24]
	5	5.95	[4.78, 7.12]	15.11	[12.51, 17.7]	7.54	[6.54, 8.55]
	6	6.38	[4.93, 7.84]	15.26	[12.74, 17.78]	8.27	[7.15, 9.39]
	7	7.09	[5.32, 8.85]	18.79	[15.69, 21.88]	8.26	[7.13, 9.39]
	8	8.21	[6.21, 10.22]	26.29	[22.32, 30.25]	9.86	[8.37, 11.34]
	ALL	4.54	[3.65, 5.43]	17.90	[16.17, 19.63]	7.81	[7.28, 8.33]

		ELA					
Linking method		PSM & equipercentile		NAEP-as-moderator		Predicted CAS	
Statistic	Grade	Estimate	95% CI	Estimate	95% CI	Estimate	95% CI
D_1: Mean difference in expected scale scores, conditional on ability	4	1.44	[0.89, 1.99]	7.25	[4.88, 9.62]	3.45	[2.81, 4.08]
	5	1.50	[0.98, 2.02]	9.10	[6.92, 11.28]	2.81	[2.21, 3.41]
	6	1.58	[0.93, 2.23]	9.11	[6.53, 11.69]	3.09	[2.37, 3.81]
	7	1.61	[1.04, 2.18]	11.04	[8.44, 13.64]	2.00	[1.36, 2.63]
	8	1.61	[0.99, 2.23]	8.83	[6.52, 11.14]	2.81	[2.12, 3.49]
	All	0.89	[0.61, 1.17]	8.27	[6.79, 9.75]	2.75	[2.44, 3.05]
D_2: Mean difference in conditional standard errors of measurement	4	4.85	[3.84, 5.85]	17.10	[14.36, 19.84]	6.28	[5.43, 7.14]
	5	5.81	[4.54, 7.08]	19.02	[16.07, 21.97]	5.17	[4.19, 6.16]
	6	5.11	[3.91, 6.32]	17.81	[14.78, 20.83]	6.52	[5.33, 7.71]
	7	4.97	[3.86, 6.08]	18.01	[14.62, 21.4]	5.08	[3.87, 6.28]
	8	5.03	[3.9, 6.16]	18.16	[14.74, 21.59]	6.09	[4.83, 7.34]
	ALL	3.82	[3.08, 4.56]	17.93	[16.36, 19.51]	5.99	[5.55, 6.42]

Note: CI = confidence interval. Highlighted cells indicate the lowest value of D1 or D2 across linking methods.

TABLE B.3

REMSD Estimations of Population Invariance Across Linking Methods, Subject, and Grade Levels

			Gender			Race			FRPL Status			ELL Status[a]			Special Ed Status		
Subject	Grade	SDTM	PSM	NAEP	Pred	PSM	NAEP	Pred	PSM	NAEP	Pred	PSM	NAEP	Pred	PSM	NAEP	Pred
Math	3	0.025	0.062	0.448	N/A	0.089	0.374	N/A	0.084	0.428	N/A	0.047	0.414	N/A	0.045	0.393	N/A
	4	0.026	0.037	0.410	0.034	0.084	0.375	0.130	0.072	0.458	0.159	0.049	0.399	0.055	0.053	0.431	0.160
	5	0.027	0.056	0.333	0.068	0.093	0.285	0.135	0.418	0.414	0.226	0.035	0.340	0.037	0.067	0.439	0.192
	6	0.026	0.043	0.182	0.041	0.077	0.175	0.193	0.053	0.347	0.683	0.060	0.167	0.041	0.133	0.223	0.165
	7	0.026	0.094	0.729	0.038	0.085	0.554	0.078	0.129	0.762	0.071	0.236	0.494	0.100	0.068	0.469	0.141
	8	0.028	0.101	1.825	0.076	0.114	0.945	0.071	0.069	1.640	0.068	0.082	1.420	0.062	0.140	1.404	0.183
ELA	3	0.029	0.032	0.301	N/A	0.126	0.365	N/A	0.066	0.489	N/A	0.047	--	N/A	0.057	0.414	N/A
	4	0.031	0.058	0.458	0.081	0.062	0.399	0.224	0.120	0.550	0.150	**0.024**	0.415	0.072	0.062	0.602	0.150
	5	0.036	0.129	0.462	0.092	0.103	0.290	0.127	0.107	0.501	0.130	**0.030**	--	0.081	0.126	0.641	0.224
	6	0.033	0.070	0.257	0.082	0.076	0.194	0.144	0.111	0.306	0.107	0.041	--	0.069	0.129	0.444	0.199
	7	0.034	0.163	0.261	0.051	0.100	0.279	0.064	0.118	0.353	0.079	0.045	--	0.054	0.079	0.381	0.147
	8	0.034	0.096	0.362	0.079	0.186	0.280	0.065	0.223	0.425	0.096	0.050	--	0.110	0.136	0.425	0.220

[a] There were insufficient DCPS TUDA ELL students tested on the NAEP reading exam in all but grade 4 to create within-subgroup linkages by ELL status in reading. Notes: Highlighted cells indicate the method within each grade and subgroup that minimizes the value of the root expected mean square difference (REMSD). Values in bold are below the standardized difference that matters (SDTM), which is equal to one point on the CAS scale converted to standard deviation units.