American Educational Research Journal
April 2022, Vol. 59, No. 2, pp. 252–283
DOI: 10.3102/00028312211024596
Article reuse guidelines: sagepub.com/journals-permissions

New Assessments and Teacher Accountability: Lessons for Teachers' Practice

Jessalynn James
Brown University

The transition to new assessments aligned to the Common Core State Standards was a significant turning point in the standards' implementation. Concerns about the transition led districts to suspend the use of value-added scores for evaluating teachers, but changes to other measures, such as classroom observations, were rare. Using data from the Washington, DC Public Schools, I evaluate the effect of the assessment transition on teachers' practice. I find substantial declines in instructional practice, ranging from 13% to 20% of a standard deviation, for teachers in tested grades and subjects when the new exam was introduced. These results suggest that policymakers should consider the ramifications of testing changes on a wider array of teaching outcomes than value-added scores alone.

Keywords: classroom observation, common core state standards, teacher evaluation, Partnership for Assessment of College and Career Readiness exam, PARCC exam

Introduction

In June 2010, the National Governors Association and the Council of Chief State School Officers unveiled the Common Core State Standards (CCSS). By the end of 2011, the standards were officially adopted by all but five states ("Map: Tracking the Common Core State Standards," 2015) and new assessments aligned to the standards were rolled out by the 2014–2015 academic year (AY 2015).[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

Jessalynn James, PhD, is a postdoctoral research associate at the Annenberg Institute at Brown University, 164 Angell Street, Providence, RI 02906, USA; email: *jessalynn_james@brown.edu*. Her research uses a mix of causal and descriptive techniques to evaluate policies that help identify and enhance teaching quality (e.g., through evaluation, recruitment, training, and retention), as well as the implications of these policies for ameliorating academic disparities.

proficiency on these standards was expected to differ from the teaching required for other, less-rigorous standards (Conley, 2014; Floden et al., 2017), and school districts scrambled to better equip teachers with the skills and practices necessary for student success on aligned exams (Jochim & McGuinn, 2016). The transition to the new assessments was a source of considerable stress for teachers (Jochim & McGuinn, 2016); there was a widespread expectation that the type of teaching required for student success on the new exams would differ from the status quo (Kane et al., 2016; McDuffie et al., 2017), and teachers were still learning at the point of the transition how to interpret the standards and adapt their instruction accordingly (Buzick et al., 2019; Edgerton, 2020; Jochim & McGuinn, 2016).

I use evidence from the District of Columbia Public Schools (DCPS) to explore whether the transition to these new exams influenced teachers' practice. Specifically, using a difference-in-differences model, I leverage the transition to a new assessment for teachers in tested subjects and grades versus other general-education teachers—who experienced no such transition—to estimate effects on teachers' practice, as measured by the district's classroom observation rubric at the time, the Teaching and Learning Framework (TLF). While a handful of qualitative studies have begun to explore the effect of this transition on teachers' practice (Ajayi, 2016; Stosich, 2016), there is to date no empirical evidence as to whether these exams caused teachers to alter their instructional emphasis. The qualitative literature, however, suggests that teachers had difficulty adapting to the new standards and exams—consistent with prior evidence on exam changes under accountability regimes (e.g., Hamilton et al., 2007). Localized surveys and interviews with teachers (Ajayi, 2016; Stosich, 2016; Troia & Graham, 2016) and their students (Kolluri, 2018), as well as reports of teachers' practice (e.g., Buzick et al., 2019; Schweig et al., 2020), suggested that teachers had variable success in shifting their practice, and many struggled during the transition to effectively implement conceptual learning in the classroom.

From a policy perspective, this question can provide important insight into teacher evaluation systems and the contexts in which they operate. For example, it might provide guidance to districts on which practices and skills they should focus professional development under new standards and testing regimes that emphasize conceptual knowledge. Similarly, the effects of the new assessment on teachers' practice can highlight areas where teachers' instructional skills may be sensitive to assessment changes. I begin the article with a brief overview of the education context in the lead-up to the new tests. Next, I review the prior evidence on ways that teachers adapt or adjust their practice, and discuss the ways in which the exams were expected to require meaningfully different teaching. I then describe the unique context of DCPS—both specific to the testing transition, and within the broader accountability policy context at the time. Finally, I hypothesize about whether and how we might observe effects on teachers' practice, before turning to my methodological approach and findings.

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 (Common Core State Standards Initiative, 2010; Conley, 2014). Among the goals driving development of the CCSS were for the new standards to be "fewer, clearer, and higher." These standards were designed in response to evidence of a core set of skills ("fewer") required for success in 2-year college, regardless of program path. The CCSS developers 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). Critically, 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 increased 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; Doorey & Polikoff, 2016; Peterson et al., 2016; Yuan & Le, 2012).

New Standards, New Tests

In conjunction with the standards, the creators of the Common Core 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, the Partnership for Assessment of College and Career Readiness (PARCC) and the Smarter Balanced Assessment Consortium, were convened to address this goal, each developing its own Common-Core-aligned assessment to be used across participating states.

These assessments are considered to be well aligned to the Common Core math and ELA content standards across grades, and good matches to the depth of learning prescribed by the new standards (Doorey & Polikoff, 2016; Schultz et al., 2016). Although states establish their own proficiency levels even across common assessments, an analysis that compared proficiency standards—before and after the transition to Common-Core-aligned assessments—to a rigorous, nationally recognized benchmark found that most states, including Washington, DC, significantly raised their expectations for students' proficiency (Peterson et al., 2016).

Perhaps recognizing that states and districts might only superficially adopt the Common Core, the developers 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. Indeed, evidence suggests that teachers altered their instruction in response to the new exams. One educator, for example, described "reverse engineering" the assessments to align his teaching with the Common Core (Cunningham, 2014). More broadly, surveys of teachers across the testing transition suggest that teachers used the exams to guide their instruction. Survey responses indicate that teachers expected to alter their instruction (Kane et al., 2016; McDuffie et al., 2017; Troia & Graham, 2016) and self-reported practice logs suggest that teachers ultimately changed the way that they taught in response to the new exams (Buzick et al., 2019). What these studies on instructional effects of the testing transition do not clarify, however, is the distinction between whether the instructional influences consisted of changes to teachers' practice (i.e., *how* teachers instruct their students) versus the *content* of their instruction (cf. Edgerton & Desimone, 2018). The consistent implementation of the TLF over the course of the transition allows a direct test of the ways in which adopting a new, Common-Core–aligned assessment influenced teachers' practice.

Teaching Under the Common Core Exams

The Malleability of Teaching Practices

This article explores the possibility that teachers changed their practice in response to the change in assessment in DCPS. For that to be true, we must also believe that teachers' practice is malleable and that when teachers can change their practice in ways that are intentional or strategic. We have compelling evidence of both. First, a nascent body of literature demonstrates that the "returns to experience" that have been widely demonstrated for teachers' effects on student achievement (Atteberry et al., 2015; Rivkin et al., 2005; Rockoff, 2004) also occur for the teaching practices measured by classroom observation instruments; teachers demonstrate substantial improvements to their performance on classroom observations over their early careers (Bell et al., 2021; Kraft et al., 2020; Papay & Laski, 2020). We also know that improvements in teachers' effects on their student learning continue to occur, albeit at a slower pace, as teachers advance further into their careers (Papay & Kraft, 2015). The extent to which these overall improvements reflect skill development acquired through opportunities for repeated practice, supportive working conditions (e.g., Kraft & Papay, 2014), professional development interventions (e.g., Allen et al., 2011; Papay et al.,

2020), or strategic decisions on the part on teachers, is unclear. Multiple meta-analyses have demonstrated, however, that coaching programs (Kraft et al., 2018) and other purposeful teaching interventions (Garrett et al., 2019) can meaningfully move teachers' practice.

There is also reason to believe that high-stakes evaluations can influence teachers' *strategic* focus on their practice. Evidence from prior to the testing transition, for example, suggests that individual teachers' strategic improvements to practice may be at play in DCPS, which is a uniquely high-stakes setting for teacher evaluation. Teachers can lose their jobs if they fail to meet performance thresholds on the overall evaluation measure, and classroom observations are the heaviest weighted of these measures. Meanwhile, teachers who perform exceptionally well are eligible for substantial financial rewards. Multiple studies have demonstrated that this incentive structure leads to improved overall performance for teachers on both ends of the performance distribution, with measurable gains on the TLF (Adnot, 2016; Dee et al., 2021; Dee & Wyckoff, 2015). Phipps and Wiseman (2021) looked specifically at the observation component of the evaluation system, and found that the expectation of a classroom observation—proxied by the number of days remaining in the observation window at the point of observation—was associated with better teaching performance. Adnot (2016) took a more nuanced look at incentive effects by examining teachers' improvements across each of the rubric standards; teachers who were at risk of involuntary separation made meaningful improvements, but gains were concentrated on the most-prescriptive and the least-difficult teaching domains.

Instructional Practice and Common-Core–Aligned Assessment

Beyond high-stakes evaluation settings, there is some evidence that assessments are themselves important drivers of teachers' practice decisions (Buzick et al., 2019; Cunningham, 2014; Floden et al., 2017; Hamilton et al., 2007; Jennings & Lauen, 2016; Troia & Graham, 2016). Full implementation of the CCSS may additionally be most likely to occur when assessments are well aligned with student-learning standards, with an effect strengthened by accountability systems (Coburn et al., 2016; Hamilton et al., 2007). Indeed, in a survey of educators across five states that adopted the PARCC and Smarter Balanced 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 et al., 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 Common-Core–aligned assessments to influence their instruction (McDuffie et al., 2017).

The limited literature that has looked at the question of these new exams' effects on teaching practice after adoption so far (e.g., Ajayi, 2016; Stosich, 2016; Kolluri, 2018) suggests that there may be differences in performance across TLF-defined practices during the transition in DCPS. 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—particularly, if they feel uncertain about how to align their instruction accordingly (e.g., Gwynne & Cowhy, 2017; Troia & Graham, 2016). On the other hand, there may be performance gains for CCSS-aligned practices if teachers can successfully shift from a procedural to conceptual instructional emphasis.

Implementing the Common Core in DCPS

This analysis examines the implications of the assessment shift on teachers' practice through a focus on the contrast between the implementation of the PARCC assessment, which is currently used by DCPS to assess Grade 3 through 10 math and ELA achievement, and the preceding years in which DCPS used its own assessment, the DC CAS. While DCPS formally adopted the CCSS in AY 2012, there is reason to believe that teachers may not have adapted their teaching to the new standards before they began using a national CCSS-aligned assessment (e.g., Coburn et al., 2016).[2] While CAS items were gradually developed to be in alignment with the new standards, the transition to PARCC represented a much more substantial shift.

CAS and PARCC were different in ways that would be readily apparent to teachers, both in terms of structure and format.[3] The old exam asked questions in ways that could be accurately answered simply through content knowledge (e.g., the procedure for cross-multiplication), without understanding the reasons behind an answer. Based on the design of the exams, it should be more difficult to score high on PARCC without demonstrating conceptual knowledge on top of content knowledge. For example, CAS was similar to what one might picture as a traditional standardized exam, consisting almost entirely of selected-response items. The PARCC exam, in contrast, often asks students to provide justification for answers and poses scaffolded questions to demonstrate a student's thinking. It contains a high share of complex item types, including performance tasks, which place a higher emphasis on more cognitively complex skills (Darling-Hammond & Adamson, 2014; Doorey & Polikoff, 2016; Schultz et al., 2016).

Teacher Evaluation in DCPS

IMPACT

DCPS teachers may have a uniquely strong instructional response to the testing transition relative to other school districts nationally, given that its teachers are subject to high-stakes performance evaluations in which the

TLF is the dominant measure. Specifically, every school-based employee in DCPS is subject to the district's performance-evaluation system, IMPACT. All teachers are assigned annual IMPACT scores using both inputs (e.g., their instructional practice) and outputs (e.g., student achievement) of their teaching, which are then combined with other measures and weighted to determine an overall "IMPACT" score (see Table 1). Teachers in tested grades and subjects, known as "Group 1," are given value-added scores, which compare how well the teacher's students improve on the district's standardized assessment relative to similar students (in terms of prior achievement and demographics) with other teachers.[4] All other general-education teachers ("Group 2") are evaluated against student performance targets they set at the start of the school year, with approval from the school principal on both the selected measure and the teacher-developed goals. While the specific measures and component-weights used to evaluate a given teacher will vary depending on the teaching assignment, the TLF comprises a plurality all general-education teachers' final scores.

Scores are then segmented into performance levels which determine teachers' retention and advancement eligibility. The lowest performing teachers (i.e., those rated Ineffective) are subject to immediate dismissal, while other low-performing teachers must improve within one (Minimally Effective) or two (Developing) years to retain their positions. The highest performing teachers (Highly Effective) can advance multiple steps and lanes on the career ladder, and are eligible for sizeable (as much as $25,000) annual bonuses (see Dee & Wyckoff, 2015, and Dee et al., 2021, for more detail on IMPACT's design and effects).

Changes to IMPACT During PARCC Implementation

While the core structure of IMPACT did not change during the transition to the PARCC exam, the district made several related changes out of concern that it would be inappropriate to assign teachers value-added scores on a new assessment and logistical concerns about the timing of score availability (DCPS, 2014, 2015). Specifically, the district took a 2-year hiatus from estimating teachers' value-added to student achievement. In 2015 and 2016, Group 1 teachers—those in tested grades and subjects for whom value-added scores would have previously comprised 35% of their IMPACT score—saw that weight shifted to the TLF instead (Table 1). This shift in score weights could be considered a shift in incentives toward performance on the TLF.

Hypotheses

Given evidence from DCPS and other settings that teaching practice is malleable and that teachers improve in response to high-performance stakes, it is reasonable to hypothesize that the transition to new exams might cause teachers to alter their practice. This could manifest negatively, with temporary

Table 1

IMPACT Score Components and Weights, AY 2009–2010 to AY 2015–2016[a]

	CAS				PARCC
	2009–10 to 2011–12		2012–13 to 2013–14		2014–15 to 2015–16
IMPACT Components	Group 1	Group 2	Group 1	Group 2	Groups 1 and 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 downward for "Core Professionalism" (CP) violations reported by principals. Group 1 teachers did not have IVA calculated during the first 2 years of the PARCC exam (academic years [AY] 2014–2015 and 2015–2016); in those years, Group 1 teachers had the same score components and weights as Group 2 teachers.

[a]All values are in percentage.

declines in teaching performance as teachers build the skills to apply their practices to the new exam context, or it could manifest positively—with strategic improvements to the practices where teachers expect to have higher returns under the new exam.

If teachers strategically alter their practice in response to the new exam, the question of *which* practices might be affected is open. Teachers in tested grades and subjects might shift their teaching emphasis toward practices where they expect higher returns on the new exam. Given the Common Core's focus on high standards, deeper, conceptual understanding, and critical thinking—and the PARCC exam's development around these goals—the TLF standards that specifically reference deeper and conceptual understanding might, for example, be where teachers focus their instruction following the transition to PARCC.

Informed by both theory and the guidance available to teachers at the time (Achieve the Core, n.d.; Berlin & Cohen, 2020; Hiebert & Grouws, 2007; National Association of Secondary School Principals, 2013), there are four "Teach" standards on the TLF that I identify a priori as being well aligned to the CCSS and therefore potentially more important in teachers' minds as they prepare their students for the PARCC exam. The first, Teach 4 (*provide students multiple ways toward mastery*), emphasizes that teachers engage students "through a variety of learning styles, modalities[. . .], and intelligences[. . .]" while developing students' deep understanding of the content. In addition, Teach 5, 6, and 7 each highlight teaching that elicits depth of understanding. 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*) defines effective teaching as posing increasingly complex questions, following up with strategies to support understanding, and eliciting meaningful responses from students. These four standards, however, describe relatively complex teaching practices. Adnot (2016) found that teachers' performance on three of these standards (Teach 4, Teach 5, and Teach 6) was responsive to IMPACT's incentives. However, if teachers do not have the skill to improve their teaching, especially given that teachers in DCPS are already uniquely incentivized to perform at their best, their performance on these and possibly other teaching standards might remain unchanged or even decline.

The possibility of declines to teaching performance is informed by evidence—largely from the economics literature—about teachers' task-specific human capital. Teachers experience negative shocks to their performance, as measured by teacher effects on student achievement, when their teaching contexts change. This includes changes to the subjects and grades taught (Blazar, 2015; Cook & Mansfield, 2017; Ost, 2014), as well as changes

to the composition of peer teachers resulting from turnover (Ronfeldt et al., 2013). The magnitude of such "disruption" effects on teachers' practice (i.e., as measured by classroom observation rubrics) has not yet been documented, but anxiety among educators about the testing transition was high (Jochim & McGuinn, 2016) and could in theory have spurred (temporary) declines in teaching quality as teachers acclimated to the new exam and learned to adapt their instruction accordingly.[5] It may also be that teaching aligned with the new exam is more difficult to perform, leading to declines in TLF performance as teachers attempt to impart more conceptual learning.

Data

To answer whether the transition to new exams might cause teachers to alter their practice, I use administrative data containing DCPS teachers' demographic data (e.g., race/ethnicity, gender, age, and teaching experience), as well as their performance across several evaluation measures, including the TLF and its nine subcomponents, the "Teach" standards. While broad teaching assignments are defined in order to determine the specific evaluation measures used for a given teacher (their IMPACT group), the data do not reliably identify the grade levels or subjects in which an educator is teaching. The data include all teachers in instructional roles during the period beginning with IMPACT's initial implementation (AY 2010) through the last year that DCPS used the TLF (AY 2016). In Fall 2016, DCPS transitioned to a new rubric that was designed to be better aligned with the CCSS and with student-centered instruction (DCPS, 2017). Table 2 provides summary statistics on the analytic sample.

The Teaching and Learning Framework

The TLF is a standards-based classroom observation rubric which assesses teachers' practice across nine "Teach" domains (see Supplemental Appendix Table A1 in the online version of the journal for detailed descriptions of each teaching standard). Teachers in DCPS are evaluated up to five times per year—three times by their school administrator and twice by a "Master Educator" with content- and grade-level expertise who is external to the school.[6] Teachers are assigned raw scores for each domain ranging from 1 (lowest) to 4 (highest); for accountability purposes, scores are first converted to an observation-level average (i.e., within observation cycle and rater) and then averaged across cycles to create an overall TLF score. The TLF has reliability levels comparable to other commonly used observation measures, and is correlated with student achievement at similar levels to those reported for other classroom observation instruments (Cantrell & Kane, 2013; Gill et al., 2016; James, 2020; Meyer, 2016). In contrast to many other classroom observation systems in practice (Kraft & Gilmour, 2017), scores are

Table 2
Analytic Sample of DCPS Teachers

		Group 1		Group 2	
Teacher Characteristic	All	CAS	PARCC	CAS	PARCC
TLF score					
All observers	3.11 (0.46)	3.10 (0.49)	3.09 (0.49)	3.11 (0.47)	3.13 (0.44)
Administrators only	3.19 (0.52)	3.17 (0.53)	3.20 (0.51)	3.19 (0.53)	3.21 (0.48)
Master educators only	3.01 (0.51)	3.01 (0.53)	2.96 (0.57)	3.00 (0.51)	3.03 (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 (years)					
0–3	0.29	0.34	0.27	0.29	0.28
4–9	0.28	0.03	0.36	0.25	0.31
10+	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

Note. Statistics from analytic sample of teacher-by-year observations in DCPS between the 2009–2010 and 2015–2016 academic years. 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 the 2013–2014 academic year, 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. DCPS = District of Columbia Public Schools.

observed with frequency across the full range of the TLF. More detail on the observation timing and scoring process is provided in Appendix B in the online version of the journal.

For analysis purposes, I forego yearly averages and rely on item-by-observation-level data. Given the large number of intercorrelated subscores (see Supplemental Appendix Table B2 in the online version of the journal), I conduct a principal components analysis (PCA) to reduce the number of dimensions from nine to two. Identifying more than one distinct practice enables the estimation of variable effects across practices and allows for more precise identification of effects, while also according to a factor structure that aligns with what would have been expected a priori. This process was conducted with the expectation from a face reading of the rubric that the TLF measured items across at least two dimensions, which might include instructional practice, classroom environment, or classroom management—factors which

have been identified by researchers and practitioners on comparable classroom observation instruments (Archer et al., 2015; Ferguson & Danielson, 2015; Garrett et al., 2019; Gill et al., 2016; Lockwood et al., 2015). Empirically, I first used a training data set to conduct a PCA across the nine items using item- and observation-level data (see Section 2 of Supplemental Appendix B in the online version of the journal for more detail on the data reduction strategy). This procedure produces a dominant factor that is highly correlated with the first seven, *instruction*-oriented, TLF practices; this dimension includes the Teach standards where I have hypothesized that teachers might focus their practice in the context of the new exam. The secondary factor captures the TLF components (Teach 8 and Teach 9) that address the *classroom environment*. While this factor structure aligns with what has been observed in other—and similar—observation rubrics (e.g., Ferguson & Danielson, 2015; Garrett et al., 2019; Gill et al., 2016; Hafen et al., 2015; Kane & Staiger, 2012; Lockwood et al., 2015), as well as what might be expected given the rubric definitions for each subscore, I follow with a confirmatory factor analysis on the full analytic data set to ensure that the estimated factor structure fits the data. I use the standardized factor scores generated from the PCA in my subdimension analysis, rather than a simple average of each set of observed TLF scores, to address overlap in the dimensions.[7] Though my subskill analyses rely primarily on the two TLF factors generated by the PCA, I include results for the full set of TLF subscores in Supplemental Appendix Table A2, in the online version of the journal.

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 (Gill et al., 2016; Ho & Kane, 2013; Meyer, 2016; Whitehurst et al., 2014); the external evaluators' scores are likewise less subject to ceiling effects, as administrators tend to rate teachers' performance more highly than the master educators. More detail about the comparability of scoring processes and practices across both types of raters is provided in Supplemental Appendix B in the online version of the journal.

Method

To understand the effects of the PARCC exam on teachers' practice, I employ a standard difference-in-differences model, leveraging the transition to the new assessment for teachers in tested subjects and grades (referred to in DCPS as "Group 1" teachers) versus other general-education teachers in DCPS (Group 2)—who experienced no such transition—to estimate effects on teachers' practice. The model takes the following form:

$$TLF_{ktm} = \beta_0 + \beta_1 Group1_{ktm} + \beta_2 PARCC_{ktm} + \beta_3 Group1_{ktm} * PARCC_{ktm} + \boldsymbol{X}_{ktm}\beta_4 + \tau_k + \varepsilon_{ktm}. \quad (1)$$

In this model, TLF_{ktm} is teacher k's TLF score in year t at school m; $Group1_{ktm}$ is an indicator for the Group 1 status (i.e., teaching math or ELA in a tested grade); and $PARCC_{ktm}$ is an indicator for whether the teacher is teaching in a PARCC-exam year. β_0 can be interpreted as the conditionally expected value of TLF scores for Group 2 (all other general-education) teachers in the lead-up to PARCC, while β_1 provides an estimate of the difference between the two teaching groups' performance before PARCC. β_2 represents the change in scores for Group 2 teachers following the adoption of the PARCC exam. Finally, the coefficient of interest, β_3, represents the treatment effect—that is, the additional effect on teaching practice for teachers in a tested grade and subject associated with teaching in a PARCC exam year relative to contemporaneous Group 2 teachers. Preferred models control for a vector of time-varying, pretreatment teacher characteristics ($\boldsymbol{X}_{ktm}$) and teacher fixed effects (τ_k), though I also test for robustness to a school's estimated level of departmentalization and school fixed effects.

The teacher fixed effects in this model reduce potential for bias associated with characteristics of individual teachers, such as prior training, fixed teaching preferences, and underlying adaptability to new contexts that might sort teachers to different teaching assignments across the transition and otherwise be conflated with the consequences of the new exam. The trade-off with this approach, however, is that by estimating score changes within individual teachers, it only identifies effects off the teachers who were present in both the pre- and postperiod, which may limit the statistical power of the corresponding point estimates.

It should be noted that treatment in this context encompasses not simply the shift to the new exam but also a shift in the components making up those teachers' evaluation scores. As described above, DCPS is a uniquely high-stakes teaching context. Because of concerns about teachers making the transition under such high-stakes circumstances, DCPS shifted the weight of student achievement on standardized exams for teachers subject to value-added 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 changes in the probability of being a teacher in a tested grade and subject (i.e., in Group 1) in a PARCC (versus CAS) exam year are as good as random. This assumption would be violated if, for example, teachers with higher teaching

Table 3
Difference-in-Differences Covariate Balance

Teacher Characteristic	Point Estimate (Standard Error)
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)

$^{\dagger}p < .10$. $^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

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 Equation 1 with pretreatment teacher characteristics. Table 3 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.[8]

The second assumption is that of common pretreatment trends in TLF scores for teachers in tested grades and subjects (Group 1) and their general-education peers (Group 2); the relationship in scores over time for the two groups should be parallel. If trends from AY 2010 (the start of IMPACT) through 2014 (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 demonstrates that these trends are generally parallel, albeit noisy. Overall TLF scores moved in the same direction for both groups before DCPS adopted the PARCC exam. The same is true for the *instruction* and *classroom environment* subdomains, where TLF scores across the groups rise and fall roughly in tandem. Any visually detectable deviations in trends are insufficiently small to explain the large negative effects that are observed, for example, in *instruction*. The magnitude and interpretation of these effects are described in more detail in the "Results" and "Discussion and Conclusion" sections.

While visual evidence indicates generally parallel trends, a potential threat to the internal validity of these estimates is a concurrent trend toward departmentalization in DCPS, given that DCPS expected the specialization associated with departmentalization to make it such that teachers would be

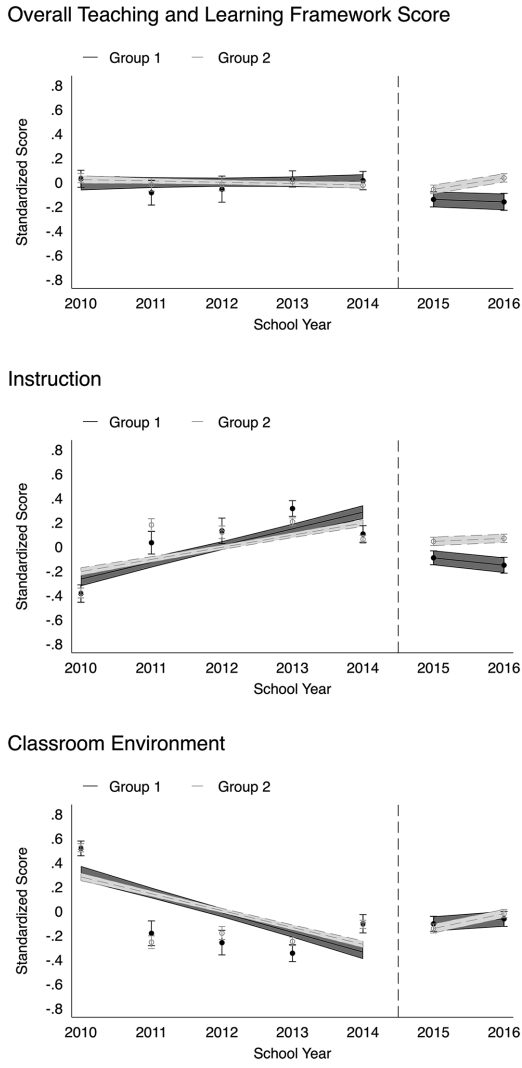

Figure 1. **Difference-in-differences of teachers' practice across the transition to PARCC.**

Note. The *Overall* score in the top panel is composed of a simple average of the nine Teach domains. The *Instruction* and *Classroom Environment* scores are factor scores from a principal-components factor analysis (see Section 2 of online Supplemental Appendix B in the online version of the journal). For each, scores are converted to standard deviation units. PARCC = Partnership for Assessment of College and Career Readiness exam.

"better able to craft rigorous and engaging lessons for students" (DCPS, 2016) and there is some evidence that specialization might influence teaching quality (e.g., Bastian & Fortner, 2020). Data on which teachers were departmentalized or departmentalization rates within schools were not collected or maintained by DCPS, so to control for departmentalization effects I instead estimate the school- and year-level 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 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.

Other factors occurring before the transition, however, might influence the scores of teachers in tested grades and subjects (Group 1) differently from their other general education (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 AY 2013, where the weight of value-added scores on Group 1 teachers' overall IMPACT scores was decreased from 50% to 35%, with some of that reallocation going toward TLF scores; the TLF weight for these teachers shifted from 35% in the 2011–2012 academic year to 40% in 2012–2013.[9] 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 Group 2 teachers relative to the preceding years, I run alternative specifications that omit the first 3 years of IMPACT.

Similarly, Figure 1 suggests that teachers in both groups scored differently in *instruction* and *classroom environment* in the first year of IMPACT 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 subscores; the rubric was substantially revised and streamlined in the second year of IMPACT, including the collapse of these subscores 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 Group 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 test for common trends empirically with a nonparametric event-study model. Specifically, I regress teachers' TLF scores on interactions between each year and teachers' Group 1 status (i.e., whether they teach in grades and subjects for which value-added scores can be estimated, versus in other general education classrooms), omitting the last pretreatment year (2014), and including year fixed effects along with the other covariates from my primary specification, as below:

$$TLF_{ktm} = \sum_{t \neq 2014} \beta_t Group1_{ktm} * PARCC_{ktm} + \boldsymbol{X}_{ktm}\beta + \tau_k + \theta_t + \varepsilon_{ktm}. \quad (2)$$

The values of β_t for Years 2010 through 2013 (the years leading up to the final pretreatment year), represent the change in TLF scores associated with being a Group 1 teacher, from that year to 2014—just before the implementation of the PARCC exam. 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, while post-PARCC effects should be significant. Figure 2 plots the results of these tests, and suggests that, while the trends in *instruction* scores did not appear to significantly 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, including a comparative interrupted time series (CITS) model which adds to Equation (1) 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 that the change in level and trend in other general education (Group 2) teachers is the change in level and trend in TLF scores we would expect to observe had the teachers subject to value-added-score estimation (Group 1) not transitioned to PARCC. CITS models also allow for a test of differential trends before the PARCC transition; these tests yield nonsignificant differences in Group 1 and Group 2 teachers' pretreatment trends.

A final potential threat to the validity of results includes the failure of measurement noninvariance—what might be considered a confounding instrumentation effect—if there are shifts in the constructs being measured by the TLF over time (e.g., Oort et al., 2005; Widaman et al., 2010). This might occur if raters reconceptualized the constructs behind the TLF in the presence of shifting expectations for student learning. Failure of measurement noninvariance is not a threat to causal identification on its own; however, if there were different response shifts for teachers in tested grades and subjects than there were for other general-education teachers over the transition, results would

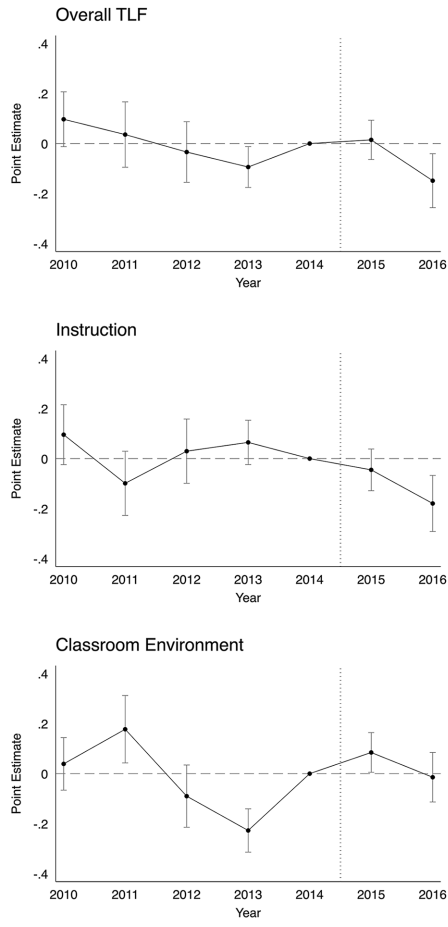

Figure 2. **Test for common trends in teachers' practice across the transition to PARCC.**

Note. Point estimates from regression of Teaching and Learning Framework (TLF) scores on interactions between group and year of evaluation; the last pretreatment year (2014) is the reference point. Group 1 (treatment) consists of teachers in tested grades and subjects; Group 2 (control) consists of all other general-education teachers. Effects are estimated within teacher. Confidence intervals are at the 95% level. The *Overall* score in the top panel is composed of a simple average of the nine Teach domains. The *Instruction* and *Classroom Environment* scores are domain scores from a principal-components factor analysis (see Section 2 of online Supplemental Appendix B in the online version of the journal). For each, scores are converted to standard deviation units. PARCC = Partnership for Assessment of College and Career Readiness exam.

be confounded by these changing constructs. That is, 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. Raters might, for example, use changing benchmarks for what constitutes "depth of understanding" when implementing the rubric during the testing transition. This risk for my analysis is likely mitigated by (a) a reliance primarily on the scores assigned by master educators and (b) the district's use of *Align*, a rater training and calibration system that is standardized across raters and over time (see Supplemental Appendix B in the online version of the journal for more detail on the calibration process). That being said, it is not possible to directly assess the presence or extent of this phenomenon in the data, and this risk remains a potential limitation.

Ultimately, it is difficult to determine with certainty whether trends are truly parallel across the two groups. Conversations with DCPS central office administrators, however, indicated no additional likely contributors to differential changes in teachers' practice over the pre-PARCC period beyond those addressed above, and statistical tests for differences do not provide evidence in support of diverging trends across the two groups of teachers.

Results

The general understanding at the time was that the new Common-Core–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 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. So what was the effect on teachers' practice?

Overall TLF Effects

Table 4 presents the results from my main difference-in-differences specifications, the first column of which displays estimates from a model controlling only for the level of departmentalization within a teacher's school.[10] This model indicates a decline in teachers' overall practice of roughly 15% of a standard deviation, much of which appears to be driven by the *instruction* domain of the TLF, where teachers' performance declines by 18% of a standard deviation when they switch to PARCC. Overall-TLF effects are robust to the inclusion of teacher controls and school fixed effects (columns 2 and 3). However, when overall-TLF effects are estimated within individual teachers (i.e., with teacher fixed effects; column 4), while the direction of change is still negative, the magnitude of the effect is no longer distinguishable from zero.

Table 4
Difference-in-Differences Estimates of PARCC Effects on Teachers' Practice

Models	(1)	(2)	(3)	(4)
Overall TLF	−0.146*** (0.042)	−0.163*** (0.041)	−0.166*** (0.038)	−0.054 (0.039)
Factor 1: *Instruction*	−0.180*** (0.038)	−0.189*** (0.037)	−0.199*** (0.036)	−0.133*** (0.043)
Factor 2: *Classroom Environment*	−0.009 (0.040)	−0.024 (0.039)	−0.017 (0.038)	0.081* (0.041)
Control for level of departmentalization	X	X	X	X
Teacher controls		X	X	X
School FE			X	
Teacher FE				X
N	22,790	22,790	22,790	22,790

Note. The outcome variable is the Teaching and Learning Framework (TLF) score assigned by master educators, standardized relative to the overall mean and standard deviation of master-educator–assigned TLF scores across the years of analysis (2009–2010 to 2015–2016). The *Overall TLF* score in the top panel is composed of a simple average of the nine Teach domains. The *Instruction* and *Classroom Environment* scores are factor scores from a principal-components factor analysis (see Section 2 of Supplemental Appendix B in the online version of the journal). For each, scores are converted to standard deviation units. Teacher controls include education level, race, gender, and experience. Robust standard errors, clustered at the teacher level, are in parentheses. FE = fixed effects.
$^{\dagger}p < .10$. $^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

Estimates from alternative specifications that include administrator-assigned scores or exclude earlier years of IMPACT from the analysis (see Supplemental Appendix Table A3 in the online version of the journal, columns 2 and 3–4, respectively) are likewise null. CITS estimates produce similar results to those from the first three models shown in Table 4 (columns 1 through 3), but imply that the null effect from the difference-in-differences analysis with teacher fixed effects (column 4) may mask heterogeneity across years. Specifically, results from the linear CITS specification with teacher fixed effects produce mixed effects directionally across the first 2 years of PARCC (see Supplemental Appendix Table A4 in the online version of the journal, column 1), and estimates from a specification that allows for nonlinear (i.e., quadratic) trends in the pre-PARCC years further indicate large negative effects on teachers' overall practice, predominantly in the second year of the exam (column 2).

Effects on Domains of Teachers' Practice

At the subdimension level, the negative overall effects appear to be concentrated within the *instruction* domain. Difference-in-difference (Table 4) estimates are consistent and precisely estimated across specifications, ranging from 13% (Column 4) to 20% (column 3) of a standard deviation decline in teachers' instructional performance. Similar-magnitude declines are estimated with the CITS approach (see Supplemental Appendix Table A4 in the online

version of the journal). 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.

To put these effect sizes in context, they are similar in magnitude to the effects resulting from the high-stakes consequences of teacher performance in DCPS. Dee and Wyckoff (2015) find for teachers on both the high and low ends of the distribution that the incentives built in to IMPACT cause teachers to improve their overall TLF scores by approximately 0.10 rubric points, or roughly 20% of a standard deviation. Adnot (2016) estimates similar-magnitude effects on the overall measure, with point estimates on the subdomains that comprise the *instruction* domain ranging from an imprecisely estimated 0.05 (Teach 1) to a statistically significant 0.21 ($p < .001$; Teach 5). In short, the instructional losses associated with the PARCC transition are comparable in size to the gains in practice induced by IMPACT's high stakes.

Meanwhile, results for the second dimension of the TLF, *classroom environment*, provide at best only suggestive evidence of positive effects on *classroom environments* for teachers who transitioned to PARCC, and these effects are not robust to other estimation decisions. Models that do not fully account for changes in teacher characteristics across the transition are null, while results from the specification with teacher fixed effects are positive and of modest magnitude (0.08 standard deviations, in column 4 of Table 4). Effects are similarly positive but underpowered when estimated with both internal and external evaluators' TLF scores, though larger and estimated with high precision when restraining pretreatment years (columns 3 and 4 of Supplemental Appendix Table A3 in the online version of the journal). A comparative interrupted time series approach (see Supplemental Appendix Table A4 in the online version of the journal) similarly points to potentially positive effects in the first year of PARCC, though point estimates are sensitive to model specification, where a quadratic approach (column 2) 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 PARCC transition effects on *instruction*, though an analysis that separately explores each of the original nine standards (see Supplemental Appendix Table A2 in the online version of the journal) suggests that the largest instruction declines may be due to teachers struggling to adequately *respond to student understanding* (Teach 6). As an illustrative example, exemplary practice on this standard includes the teacher anticipating common misunderstandings or recognizing "a student response as a common misunderstanding and sharing it with the class to lead all students to a more complete understanding" (see Supplemental Appendix Table A1 in the online version of the journal). This

is a complex skill that may be more difficult to demonstrate when the content being taught is itself more difficult. For example, if an educator were teaching a given lesson (e.g., on cross-multiplication) before PARCC, she might have relied primarily on building students' procedural knowledge to meet the learning standard and for the students to score well on CAS. It would be easier for the educator to score well on Teach 6 when a "complete understanding" of the content was purely *how* to cross-multiply. However, during the PARCC transition, if the educator were attempting to adapt her instruction to develop the more-conceptual knowledge the test was meant to measure, she might not have the depth of skills or knowledge herself to anticipate students' misunderstanding—particularly when the student is still able to arrive at the correct answer using procedural tools—and to attend to the conceptual misunderstanding when it occurs. Without experience teaching in this new context, teachers may have lacked the skills to effectively implement their instruction in the ways called for by the new exam and as defined by the TLF.

Discussion and Conclusion

The transition from traditional achievement tests to more rigorous, CCSS-aligned exams provides an opportunity to explore the implications of major transitions for teachers' practice. I find that the transition to the new exam may have altered the quality of teachers' practice—even if only temporarily. Teachers transitioning to PARCC experienced large declines in instruction. While I cannot determine whether these effects would have persisted beyond the first 2 years of the transition, at minimum the relative decline in these 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 (Desimone et al., 2019; Kane et al., 2016; Polikoff, 2012) and for their practice in general (Charalambous et al., 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 educators' ability to effectively design or otherwise implement new curricula.

Meanwhile, individuals at multiple levels of the education system struggled with inadequate guidance on how to align instruction to new learning standards well after the Common Core was officially adopted (Edgerton, 2020; Edgerton & Desimone, 2018; Gwynne & Cowhy, 2017; Pak & Desimone, 2019). While teachers felt considerable pressure to adapt their

instructional materials to their new testing environment (Kane et al., 2016), few teachers felt well prepared to help their students perform well on new exams like PARCC (Kane et al., 2016; Perry et al., 2015; Troia & Graham, 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." In short, the transition was a source of anxiety and stress for many teachers. This may have been particularly true in the uniquely high-stakes context of teaching in DCPS.

Declining outcomes following significant transitions, such as I observe here, are a common phenomenon. Fullan (2001) defines these "implementation dips" as "literally a dip in performance and confidence as one encounters an innovation that requires new skills and new understandings" (pp. 40–41). I am unable, due to other subsequent changes to the evaluation and professional development system that occurred in 2016, to determine whether teachers might have begun to improve again in the years that follow. However, across a variety of education interventions, researchers have observed accordingly null or negative effects in the early stages of the implementation before benefits begin to manifest. Such patterns have been observed, for example, following school turnaround reforms (e.g., Borman et al., 2003; de la Torre et al., 2013; Player & Katz, 2016; Sun et al., 2017), reading interventions (e.g., Borman et al., 2007), and the adoption of the online version of the PARCC assessment for student test scores (Backes & Cowen, 2019). In fact, the same patterns occurred for DCPS in the earliest year of its teacher evaluation program, IMPACT. Initial effects were null, but by the second year effects were large and statistically significant, and have persisted over time (Dee et al., 2021; Dee & Wyckoff, 2015). Temporary loss of human capital has likewise been observed for another measure of teaching quality—their value added to student achievement—when teachers switch subject and grade assignments (Blazar, 2015; Cook & Mansfield, 2017; Ost, 2014). It's quite feasible, therefore, to imagine that teachers would also experience implementation dips for their own teaching practice when the contexts within which they are teaching are subject to change.

While—as with any intervention—we might expect that it would take time for positive effects to reveal themselves, the by-year performance patterns demonstrated by the event-study analysis (Figure 2) and the comparative interrupted time series analysis (see Supplemental Appendix Table A4 in the online version of the journal) are consistent with some of this being a learning story. In 2015, teachers were preparing their students for the exam, but did not yet have materials available in order to adapt their instruction, nor did they have information with which to benchmark how well they were doing at teaching their students to perform well on the PARCC exam. It was not until the 2016 school year that teachers were able to observe their students' performance on the CCSS-aligned standards and acquire more concrete

understandings of how the standards were being operationalized by PARCC. Throughout this process, teachers were no doubt learning and adapting their instruction—and with time they were able to acquire more information about the standards, what student achievement on these standards might look like, and which teaching practices might be more effective in their classrooms in light of the new assessment. These declines may very well reflect teachers' presence in a building phase of implementing better-aligned instruction.

The transition to new exams during the No-Child-Left-Behind accountability roll-out roughly a decade earlier is illustrative. In their study of teacher experiences during the testing regime transition across three states in 2004 and 2005, Hamilton et al. (2007) found that a clear majority of surveyed teachers reported seeking out more effective teaching methods in response to the new tests, with the share of teachers engaging in this activity declining over time as they finessed the adaptations to their instruction. In Chicago, surveys of teachers implementing the Common Core likewise suggest that teachers had gained confidence in their preparedness to teach to the new standards in 2016 relative to 2015, while consistently engaging in collaboration and learning alongside their colleagues (Gwynne et al., 2018). While the change in observation rubrics in 2016–2017 prohibits the study of longer term trends of teachers' practice in DCPS, consistent improvements in student achievement on the PARCC exam over time suggest that students' math and ELA skills were growing substantially during the transition years, and continued to improve at least through 2019.[11]

At the same time, DCPS recognized the efforts that its teachers were making toward adapting their practice, in addition to acknowledging the necessity for helping its teachers implement effective, aligned instruction. In the 2016–2017 academic year, following receipt of results from the first years of PARCC testing, DCPS introduced a major reform to its approach to professional development (Toch, 2018). In part out of concern that its teachers were struggling to align their teaching with the Common Core, DCPS launched "LEarning together to Advance our Practice" (LEAP), an intensive professional development program that provides grade- and subject-specific coaching and content support to all its teachers on a weekly basis (Cohen et al., in press). 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 professional development, and included scores on Common-Core–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 that professional development such as LEAP may help teachers in tested grades and subjects develop strategies to recover the practices upon which they struggled 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 Common

Core and assessed by PARCC. DCPS also transitioned to a new classroom observation rubric, Essential Practices, which was "designed to mirror the rigor and shifts of the Common Core State Standards [. . .]" (DCPS, 2017).

Transitions in standards and assessments are not uncommon (Backes et al., 2018), and the research presented in this article provides insights into what other districts might expect in future assessment transitions, particularly 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 about teachers' value-added, districts may want to additionally consider the fairness of other measures—even those not directly linked to student achievement—when significant changes are made to state assessments. Indeed, in any period of substantial transition, districts may need to consider the supports available and consequences for teacher practice, especially when measures of practice are used for high-stakes evaluation.

DCPS, like many districts during the transition to CCSS-aligned tests, temporarily shifted the weight of its evaluation measures away from student-achievement–based outcomes and toward classroom observation during the exam implementation period. 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.

ORCID iD

Jessalynn James https://orcid.org/0000-0001-6765-2642

Notes

The opinions expressed are those of the author and do not represent views of the Institute or the U.S. Department of Education. I am grateful to the District of Columbia Public Schools for supplying the data employed in this research and to Scott Thompson, Luke Hostetter, Lauren Norton, and Liz Kim for answering my many questions. I appreciate feedback on earlier versions of this article from discussants and participants at the Association of Education Finance and Policy and the Association for Public Policy and Management annual meetings, and from colleagues at the University of Virginia and the Annenberg Institute at Brown University; I am especially grateful for the immensely helpful feedback and suggestions I have received from Jim Wyckoff, Julie Cohen, Patrick Meyer, Eric Taylor, Peter Youngs, and the editors at AERJ along with three anonymous reviewers. The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant No. R305B140026 to the Rectors and Visitors of the University of Virginia.

[1]Membership in the CCSS has since evolved as states withdrew or revised the standards.

[2]I test for changes in teachers' practice at the point of transition to the CCSS but before the transition to PARCC to test whether this assertion holds empirically; I find no meaningful

changes in TLF performance across the standards-based transition. Similarly, results are substantively unchanged when limiting analysis to post standards-adoption years.

[3]CAS technical manuals can be found at https://osse.dc.gov/publication/dc-cas-technical-reports. CAS blueprints and resources guides for the Spring 2013 administration, with sample item stems, are available at https://osse.dc.gov/service/dc-cas.

[4]Teachers rostered to students taking math or ELA exams constitute Group 1. However, this excludes third-grade teachers (the lowest grade tested) because value-added estimation requires lagged scores. While value-added scores were not estimated for teacher accountability in 2015 or 2016, IMPACT group assignments remained aligned to this definition.

[5]An instructive example includes states' transitions to standards-based assessments under No Child Left Behind a decade earlier, when high shares of teachers reported searching for more effective teaching methods in response to the transition (Hamilton et al., 2007).

[6]The rubric is modeled closely off the Danielson (2007) Framework for Teaching, with additional elements drawn from other measures including the University of Virginia's *Classroom Assessment Scoring System* (CLASS; Pianta et al., 2007) and Wiggins and McTighe's (2005) *Understanding by Design*. Teachers are subject to fewer formal observations as they advance on the district's performance-based career ladder, but in the years studied were subject to a minimum of one observation from each category of rater (i.e., internal school administrator or external Master Educator) each year, and the typical teacher received three to four formal observations per year. See Supplemental Appendix B in the online version of the journal for further details on the scoring process in DCPS.

[7]Specifically, I test a factor structure where *instruction* is the latent variable for Teach Standards 1 through 7 and *classroom environment* is the latent variable for Teach Standards 8 and 9. Evidence points to adequate-to-good fit for this factor structure. See Supplemental Appendix B in the online version of the journal for details and goodness-of-fit statistics. Effects are similar in terms of magnitude, sign direction, and statistical significance when using PCA factor scores (for the overall TLF, as well as for subdimensions) to those from CFA factor scores, as well as simple averages of subscores at the overall-TLF and subdimension level. Results using the CFA and simple-average *instruction* and *classroom environment* domain factors are not shown but are available on request.

[8]Ideally, I would control for the individuals assigning TLF scores, as effect estimates could be biased if raters were differentially assigned to Group 1 teachers at the transition and differed in their operationalization of the rubric. Unfortunately, I am unable to identify unique raters over time. This concern is mitigated, however, by DCPS's use of a rater training system, *Align*, to calibrate classroom observers. If the rater alignment process were working perfectly, it should not matter who rates a given teacher. However, it is still possible that different raters might operationalize the TLF differently from each other, or over time.

[9]See Dee et al. (2021) for a summary of the changes that went into place in AY 2013.

[10]A naïve model is shown in the first column of Supplemental Appendix Table A3 in the online version of the journal. Although proponents of departmentalization argue that it might improve teaching by allowing educators to specialize, 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.

[11]PARCC student achievement scores are available through DC's Office of the State Superintendent of Education, at https://osse.dc.gov/parcc.

References

Achieve the Core. (n.d.). *Instructional practice toolkit (IPT) & classroom videos*. https://achievethecore.org/category/1193/instructional-practice-toolkit-ipt-classroom-videos

Adnot, M. (2016). *Effects of incentives and feedback on instructional practice: Evidence from the District of Columbia Public Schools' IMPACT teacher evaluation system* [Working paper].

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. https://doi.org/10.1007/s10671-015-9174-3

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. https://doi.org/10.1126/science.1207998

Atteberry, A., Loeb, S., & Wyckoff, J. (2015). Do first impressions matter? Predicting early career teacher effectiveness. *AERA Open*, *1*(4), 1–23. https://doi.org/10.1177/2332858415607834

Archer, J., Cantrell, S., Holtzman, S. L., Joe, J. N., Tocci, C. M., & Wood, J. (2015). *Seeing it clearly: Improving observer training for better feedback and better teaching*. Bill and Melinda Gates Foundation. https://usprogram.gatesfoundation.org/-/media/dataimport/resources/pdf/2016/11/met-seeing-it-clearly-v2.pdf

Backes, B., & Cowan, J. (2019). Is the pen mightier than the keyboard? The effect of online testing on measured student achievement. *Economics of Education Review*, *68*(1), 89–103. https://doi.org/10.1016/j.econedurev.2018.12.007

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. https://doi.org/10.1016/j.econedurev.2017.10.004

Bastian, K. C., & Fortner, C. K. (2020). Is less more? Subject-area specialization and outcomes in elementary schools. *Education Finance and Policy*, *15*(2), 357–382. https://doi.org/10.1162/edfp_a_00278

Bell, C., James, J., Taylor, E., & Wyckoff, J. (2021). *Measuring improvement in true teaching performance using classroom observations* [Working paper].

Berlin, R., & Cohen, J. (2020). The convergence of emotionally supportive learning environments and college and career ready mathematical engagement in upper elementary classrooms. *AERA Open*, *6*(3), 1–20. https://doi.org/10.1177/2332858420957612

Bill and Melinda Gates Foundation. (2010). *Fewer, clearer, higher: Moving forward with consistent, rigorous standards for all students*. https://docs.gatesfoundation.org/documents/fewer-clearer-higher-standards.pdf

Blazar, D. (2015). Grade assignments and the teacher pipeline: A low-cost lever to improve student achievement? *Educational Researcher*, *44*(4), 213–227. https://doi.org/10.3102/0013189X15580944

Borman, G. D., Hewes, G. M., Overman, L. T., & Brown, S. (2003). Comprehensive school reform and achievement: A meta-analysis. *Review of Educational Research*, *73*(2), 125–230. https://doi.org/10.3102/00346543073002125

Borman, G. D., Slavin, R. E., Cheung, A. C. K., Chamberlain, A. M., Madden, N. A., & Chambers, B. (2007). Final reading outcomes of the national randomized field trial of success for all. *American Educational Research Journal*, *44*(3), 701–731. https://doi.org/10.3102/0002831207306743

Buzick, H. M., Rhoad-Drogalis, A., Laitusis, C. C., & King, T. C. (2019). Teachers' views of their practices related to Common Core State Standards-aligned assessments. *ETS Research Report Series*, *2019*(1), 1–18. https://doi.org/10.1002/ets2.12277

Cantrell, S., & Kane, T. J. (2013). *Ensuring fair and reliable measures of effective teaching: Culminating findings from the MET Project's three-year study*. Bill and Melinda Gates Foundation. https://usprogram.gatesfoundation.org/-/media/dataimport/resources/pdf/2016/12/met-ensuring-fair-and-reliable-measures-practitioner-brief.pdf

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. https://doi.org/10.1080/00220272.2012.716974

Coburn, C. E., Hill, H. C., & Spillane, J. P. (2016). Alignment and accountability in policy design and implementation: The Common Core State Standards and implementation research. *Educational Researcher*, *45*(4), 243–251. https://doi.org/10.3102/0013189X16651080

Cohen, J. C., Wyckoff, J., Katz, V., Boguslav, A., Sadowski, K., & Wiseman, E. A. (in press). Implementing targeted professional development at scale in the District of Columbia Public Schools. *American Educational Research Journal*.

Common Core State Standards Initiative. (2010). *Common Core State Standards*. http://www.corestandards.org/read-the-standards/

Conley, D. T. (2014). *The Common Core State Standards: Insight into their development and purpose*. 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*, 51–72. https://doi.org/10.1016/j.jpubeco.2016.04.001

Cunningham, E. (2014). Opportunity costs of the Common Core in high school ELA. *English Journal*, *104*(2), 34–40. https://www.jstor.org/stable/24484404

Danielson, C. (2007). *Enhancing professional practice: A framework for teaching* (2nd ed.). Association for Supervision and Curriculum Development.

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

de la Torre, M., Allensworth, E., Jagesic, S., Sebastian, J., Salmonowicz, S., Meyers, C., & Gerdeman, R. D. (2013). *Turning around low-performing schools in Chicago*. UChicago Consortium on School Research. https://consortium.uchicago.edu/publications/turning-around-low-performing-schools-chicago-full-report

Dee, T. S., James, J. K., & Wyckoff, J. (2021). Is effective teacher evaluation sustainable? Evidence from DCPS. *Education Finance and Policy*, *16*(2), 313–346. https://doi.org/10.1162/edfp_a_00303

Dee, T. S., & Wyckoff, J. (2015). Incentives, selection, and teacher performance: Evidence from IMPACT. *Journal of Policy Analysis and Management*, *34*(2), 267–297. https://doi.org/10.1002/pam.21818

Desimone, L. M., Stornaiuolo, A., Flores, N., Pak, K., Edgerton, A. K., Nichols, T. P., Plummer, E., & Porter, A. (2019). Successes and challenges of the "new" college- and career- ready standards: Seven implementation trends. *Educational Researcher*, *48*(3), 167–178. https://doi.org/10.3102/0013189X19837239

District of Columbia Public Schools. (2014). *2014-15 IMPACT guidebook: Group 1: General education teachers with individual value-added student achievement data*.

District of Columbia Public Schools. (2015). *2015-16 IMPACT guidebook: Group 1: General education teachers with individual value-added student achievement data*.

District of Columbia Public Schools. (2016). *FY2016 performance accountability report*. https://oca.dc.gov/sites/default/files/dc/sites/oca/publication/attachments/DCPS_FY16PAR.pdf

District of Columbia Public Schools. (2017). *2016-17 IMPACT guidebook: Group 1: General education teachers with individual value-added student achievement data*.

Doorey, N., & Polikoff, M. (2016). *Evaluating the content and quality of next generation assessments*. Thomas B. Fordham Institute. https://fordhaminstitute.org/national/research/evaluating-content-and-quality-next-generation-assessments

Edgerton, A. K. (2020). Learning from standards deviations: Three dimensions for building education policies that last. *American Educational Research Journal*, *57*(4), 1525–1566. https://doi.org/10.3102/0002831219876566

Edgerton, A. K., & Desimone, L. M. (2018). Teacher implementation of college- and career- readiness standards: Links among policy, instruction, challenges, and resources. *AERA Open*, *4*(5), 1–22. https://doi.org/10.1177/2332858418806863

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, & R. C. Pianta (Eds.), *Designing teacher evaluation systems: New guidance from the measures of effective teaching project* (pp. 98–143). Jossey-Bass.

Floden, R. E., Richmond, G., & Andrews, D. C. (2017). Responding to the challenge of new standards. *Journal of Teacher Education*, *68*(3), 236–238. https://doi.org/10.1177/0022487117702380

Fullan, M. (2001). *Leading in a culture of change*. Jossey-Bass.

Garrett, R., Citkowicz, M., & Williams, R. (2019). How responsive is a teacher's classroom practice to intervention? A meta-analysis of randomized field studies. *Review of Research in Education*, *43*(1), 106–137. https://doi.org/10.3102/0091732X19830634

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)*. U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Mid-Atlantic. https://ies.ed.gov/ncee/edlabs/projects/project.asp?projectID=4474

Gwynne, J. A., & Cowhy, J. R. (2017). *Getting ready for the Common Core State Standards: Experiences of CPS teachers and administrators preparing for the new standards*. UChicago Consortium on School Research. https://consortium.uchicago.edu/publications/getting-ready-common-core-state-standards-experiences-cps-teachers-and-administrators

Gwynne, J. A., Cowhy, J. R., & Quispe, R. S. (2018). *Getting ready for the Common Core State Standards: A 2016 update*. UChicago Consortium on School Research. https://consortium.uchicago.edu/publications/getting-ready-common-core-state-standards-experiences-cps-teachers-and-administrators

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. *Journal of Early Adolescence*, *35*(5–6), 651–680. https://doi.org/10.1177/0272431614537117

Hamilton, L. S., Stecher, B. M., Marsh, J. A., McCombs, J. S., Robyn, A., Russell, J., Naftel, S., & Barney, H. (2007). *Standards-based accountability under No Child Left Behind: Experiences of teachers and administrators in three states*. RAND Corporation. https://www.rand.org/pubs/monographs/MG589.html

Hiebert, J., & Grouws, D. (2007). *Effective teaching for the development of skill and conceptual understanding of number: What is most effective?* National Council of Teachers of Mathematics. https://web.archive.org/web/20120420095358/http://www.nctm.org/news/content.aspx?id=8448

Ho, A. D., & Kane, T. J. (2013). *The reliability of classroom observations by school personnel*. Bill and Melinda Gates Foundation. https://usprogram.gatesfoundation.org/-/media/dataimport/resources/pdf/2016/12/met-reliability-of-classroom-observations-research-paper.pdf

James, J. (2020). *Measuring teacher quality: Does the test make a difference?* [Working paper].

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*(5), 220–241. https://doi.org/10.7758/RSF.2016.2.5.11

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. https://www.educationnext.org/the-politics-of-common-core-assessments-parcc-smarter-balanced/

Kane, T. J., Owens, A. M., Marinell, W. H., Thal, D. R., & Staiger, D. O. (2016). *Teaching higher: Educators' perspectives on Common Core implementation*. Center for Education Policy Research. https://cepr.harvard.edu/publications/teaching-higher-educators-perspectives-common-core-implementation

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains*. Bill and Melinda Gates Foundation. https://files.eric.ed.gov/fulltext/ED540960.pdf

Kolluri, S. (2018). Student perspectives on the Common Core: The challenge of college readiness at urban high schools. *Urban Education*. Advance online publication. https://doi.org/10.1177/0042085918772630

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. https://doi.org/10.3102/0034654318759268

Kraft, M. A., & Gilmour, A. (2017). Revisiting *The Widget Effect*: Teacher evaluation reform and the distribution of teacher effectiveness. *Educational Researcher*, *46*(5), 234–249. https://doi.org/10.3102/0013189X17718797

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. https://doi.org/10.3102/0162373713519496

Kraft, M. A., Papay, J. P., & Chi, O. L. (2020). Teacher skill development: Evidence from performance ratings by principals. *Journal of Policy Analysis and Management*, *39*(2), 315–347. https://doi.org/10.1002/pam.22193

Lockwood, J. R., Savitsky, T. D., & McCaffrey, D. F. (2015). Inferring constructs of effective teaching from classroom observations: An application of Bayesian exploratory factor analysis without restrictions. *Annals of Applied Statistics*, *9*(3), 1484–1509. https://doi.org/10.1214/15-AOAS833

Map: Tracking the common core state standards. (2015). *Education Week*, *34*(36). https://www.edweek.org/teaching-learning/map-tracking-the-common-core-state-standards

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. https://doi.org/10.1086/673163

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. https://doi.org/10.1177/0895904815586850

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.

National Association of Secondary School Principals. (2013). *Implementing the Common Core State Standards: The role of the secondary school leader*. https://files.nassp.org/documents/Content/158/RevisedSecondaryActionBrief_Final_Feb.pdf

Oort, F. J., Visser, M. R., & Sprangers, M. A. (2005). An application of structural equation modeling to detect response shifts and true change in quality of life data from cancer patients undergoing invasive surgery. *Quality of Life Research*, *14*(3), 599–609. https://doi.org/10.1007/s11136-004-0831-x

Ost, B. (2014). How do teachers improve? The relative importance of specific and general human capital. *American Economic Journal: Applied Economics*, *6*(2), 127–151. https://doi.org/10.1257/app.6.2.127

Pak, K., & Desimone, L. (2019). How do states implement college- and career-readiness standards? A distributed leadership analysis of standards-based reform. *Educational Administration Quarterly*, *55*(3), 447–476. https://doi.org/10.1177/0013161X18799463

Papay, J. P., & Kraft, M. A. (2015). Productivity returns to experience in the teacher labor market: Methodological challenges and new evidence on long-term career improvement. *Journal of Public Economics*, *130*, 105–119. https://doi.org/10.1016/j.jpubeco.2015.02.008

Papay, J. P., & Laski, M. (2020). *Understanding the dynamics of teacher productivity development: Evidence on teacher improvement in Tennessee* [Working paper].

Papay, J. P., Taylor, E. S., Tyler, J. H., & Laski, M. E. (2020). Learning job skills from colleagues at work: Evidence from a field experiment using teacher performance data. *American Economic Journal: Economic Policy*, *12*(1), 359–388. https://doi.org/10.1257/pol.20170709

Perry, R. R., Finkelstein, N. D., Seago, N., Heredia, A., Sobolew-Shubin, S., & Carroll, C. (2015). Taking stock of Common Core math implementation: Supporting teachers to shift instruction. *WestEd*. https://www.wested.org/resources/taking-stock-common-core-math-implementation/#

Peterson, P. E., Barrows, S., & Gift, T. (2016). After Common Core, states set rigorous standards. *Education Next*, *16*(3), 9–15. https://www.educationnext.org/after-common-core-states-set-rigorous-standards/

Phipps, A. R., & Wiseman, E. A. (2021). Enacting the rubric: Teacher improvements in windows of high-stakes observation. *Education Finance and Policy*, *16*(2), 283–312. https://doi.org/10.1162/edfp_a_00295

Pianta, R. C., LaParo, K. M., & Hamre, B. K. (2007). *Classroom Assessment Scoring System (CLASS)*. Brookes.

Player, D., & Katz, V. (2016). Assessing school turnaround: Evidence from Ohio. *Elementary School Journal*, *116*(4), 675–698. https://doi.org/10.1086/686467

Polikoff, M. S. (2012). Instructional alignment under No Child Left Behind. *American Journal of Education*, *118*(3), 341–368. https://doi.org/10.1086/664773

Polikoff, M. S. (2015). How well aligned are textbooks to the Common Core Standards in mathematics? *American Educational Research Journal*, *52*(6), 1185–1211. https://doi.org/10.3102/0002831215584435

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, *73*(2), 417–458. https://doi.org/10.1111/j.1468-0262.2005.00584.x

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review*, *94*(2), 247–252. https://doi.org/10.1257/0002828041302244

Ronfeldt, M., Loeb, S., & Wyckoff, J. (2013). How teacher turnover harms student achievement. *American Educational Research Journal*, *50*(1), 4–36. https://doi.org/10.3102/0002831212463813

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*. Human Resources

Research Corporation. https://www.humrro.org/corpsite/wp-content/uploads/old_files/HQAP_HumRRO_High_School_Study_Final%20Report.pdf

Schweig, J. D., Kaufman, J. H., & Opfer, V. D. (2020). Day by day: Investigating variation in elementary mathematics instruction that supports the Common Core. *Educational Researcher*, *49*(3), 167–187. https://doi.org/10.3102/0013189X20909812

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. https://doi.org/10.3102/0002831216675403

Student Achievement Partners. (2013). Introduction to the ELA/literacy shifts of the Common Core State Standards [PowerPoint slides]. *Achieve the Core*. https://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]. *Achieve the Core*. https://achievethecore.org/page/399/introduction-to-the-math-shifts

Sun, M., Penner, E., & Loeb, S. (2017). Resource- and approach-driven multidimensional change: Three-year effects of school improvement grants. *American Educational Research Journal*, *54*(4), 607–543. https://doi.org/10.3102/0002831217695790

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

Troia, G. A., & Graham, S. (2016). Common Core writing and language standards and aligned state assessments: A national survey of teacher beliefs and attitudes. *Reading and Writing*, *29*(9), 1719–1743. https://doi.org/10.1007/s11145-016-9650-z

Whitehurst, G. J. R., Chingos, M. M., & Lindquist, K. M. (2014). *Evaluating teachers with classroom observations*. Brookings Institution. https://www.brookings.edu/research/evaluating-teachers-with-classroom-observations-lessons-learned-in-four-districts/

Widaman, K. F., Ferrer, E., & Conger, R. D. (2010). Factorial invariance within longitudinal structural equation models: Measuring the same construct across time. *Child Development Perspectives*, *4*(1), 10–18. https://doi.org/10.1111/j.1750-8606.2009.00110.x

Wiggins, G. P., & McTighe, J. (2005). *Understanding by design* (2nd ed.). ACSD.

Yuan, K., & Le, V. (2012). *Estimating the percentage of students who were tested on cognitively demanding items through the state achievement tests*. RAND Corporation. https://www.rand.org/pubs/working_papers/WR967.html

Manuscript received January 9, 2020
Final revision received April 23, 2021
Accepted May 17, 2021