


The Change We Cannot See: Instructional Quality and Classroom Observation in the Era of Common Core

Educational Policy
2022, Vol. 36(6) 1261–1287
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0895904820951114
journals.sagepub.com/home/epx



Julie Cohen¹ , Ethan Hutt², Rebekah Berlin³,
and Emily Wiseman¹

Abstract

The adoption of “College and Career Ready” standards—including Common Core State Standards—aims to raise academic expectations for students nationwide. Meeting these outcomes requires shifts in teaching, which, in turn, requires developing measures for the observation, assessment, and support of new kinds of instruction. This essay focuses on our efforts to develop such measures in a research project conducted in the District of Columbia Public Schools, which raised fundamental questions about whether existing measures can meet this challenge. By emphasizing observable elements of individual lessons, current measures produce a restricted view of instructional quality, omitting crucial elements of instruction called forth by new standards. Having identified this disconnect, we offer suggestions for developing multi-measure systems to capture a fuller picture of standards aligned teaching.

¹University of Virginia, Charlottesville, VA, USA

²University of North Carolina, Chapel Hill, NC, USA

³Deans for Impact, Austin, TX, USA

Corresponding Author:

Julie Cohen, University of Virginia, Bavaro Hall-Emmet Street South, Charlottesville, VA 22903-1738, USA.

Email: jjc7f@virginia.edu

Keywords

classroom instruction, educational reform, evaluation and assessment, teacher quality

For decades, American politicians and policy makers have focused educational policy on setting more ambitious goals for students. Whether driven by dissatisfaction with “old” approaches or fear of falling behind international competitors, each decade has brought a new statement about what we want from our schools. Most recently, this impulse has given rise to the widespread adoption of “College and Career Ready” standards—including, but not limited to, the Common Core State Standards (CCSS; Council of Chief State School Officers, 2010).

Though policy makers have found it easy to ask for more from our schools and our students, figuring out how to secure these more ambitious goals—to leave “no child behind,” or have “every student succeed” or be “college and career ready”—has been an endemic challenge for American educators and an enduring focus for educational researchers. A major source of this challenge is that policies invariably articulate discrete learning goals for students but not specific curricular or pedagogical choices for teachers. As in the case of the newly adopted “College and Career Ready” standards, the standards *imply* a certain kind of teaching and a certain set of pedagogical choices but stop short of explicitly spelling out for teachers the particular facets of instructional quality that would facilitate students’ meeting those standards (Cohen & Hill, 2008; Hiebert & Stigler, 2017). For example, the first CCSS standard for mathematical practice asserts that students should be able to “make sense of problems and persevere in solving them,” but does not articulate what kind of instruction would support students in sense making or persevering when solving mathematical problems. The facets of instructional quality and the tools that would support the high-quality teaching that would facilitate these outcomes are left to the imagination.

Whatever the motivation for doing so, the decision to avoid direct discussion of instructional quality in the context of new standards creates the possibility for wildly divergent outcomes: productive variation in pursuit of these standards at one extreme and, at the other extreme, aberrant, counter-productive interpretations that undermine the goal of uniform standards. In any case, this decision places a particular emphasis on the systems developed by individual schools or districts to support teachers as they adapt to new curricula, develop new pedagogies, and create new modes of instruction in light of new learning goals. As a result, the ability for teachers to fulfill new standards becomes a function of these support systems.

The goal of this paper is to highlight what we consider to be the incompatibility between the vision of ambitious instruction embedded in, if not articulated by, our current “next generation” standards and our current support systems and tools for facilitating instructional quality. At the heart of this incompatibility between standards for students and tools to support teachers is a misalignment around the construct of instructional quality. By instructional quality, we mean the learning experiences teachers provide students. These include the ways individual teachers support students in classroom interactions (Hill et al., 2008; Pianta & Hamre, 2009), as well as the ways that teachers within a school coordinate those experiences across subjects and grade levels (Johnson, 2015). It also includes the tasks, texts, and learning activities a teacher selects, and the degree to which curricular resources align with grade-level expectations for students (Pepin & Haggarty, 2001; Remillard & Heck, 2014; Stein & Kim, 2011; Walkowiak et al., 2014). Thus, when we talk about instructional quality, we are discussing the nexus of standards-based reform and increased student learning.

In developing our argument, we provide a fundamentally theoretical account of these issues. By theoretical we mean that our goal is not to provide a description of any specific district, system, or protocol, but instead to explore the issues we see as inherent to the work of developing systems to support instructional quality to meet ambitious learning standards. Though theoretical, our account is not abstract or hypothetical. It is rooted in the concrete experiences of a research project focused on Washington, DC Public Schools’ (DCPS) efforts to support teachers in implementing teaching practices that would help students in meeting the ambitious learning goals delineated in the CCSS and associated student assessments.

We provide an account of these challenges by asking questions that we think are foundational to this kind of work: what constitutes instructional quality? Is it a characteristic of an individual teacher? A set of teachers? An entire school? What is the relationship among instructional quality, curricula, and/or instructional materials? To what extent is instructional quality a visible characteristic of classroom interactions? Are there invisible, but crucial, aspects of instructional quality that could be supported in ways other than classroom observations?

In particular we focus on the use of classroom observation protocols, which are used widely in DCPS and districts across the country to provide teachers with clarity about performance goals and targeted feedback for improving practice (Hill & Grossman, 2013; Kane & Staiger, 2012). We do so because, as official tools, these rubrics convey important messages about how instructional quality has been operationalized in a particular context (Cohen & Goldhaber, 2016; Bell et. al., 2012; Hill & Grossman, 2013). As

one of the documents that attempts to translate learning goals into instructional practices, these rubrics play a key role in the reform cycle and warrant careful study, as they can only be as effective at directing teacher practice as their interpretive choices.

In examining these issues, we argue that existing observational tools identify some critical aspects of teaching, but they also make several assumptions about the relationship between instructional practice and outcomes that do not address key aspects of the broader terrain instructional quality suggested by the newest generation of widely adopted learning standards. These assumptions include: (1) that an individual lesson is the appropriate unit of analysis for instructional quality, and (2) that instructional quality is a function of visible elements within a lesson. Indeed, as we illustrate in the context of our work in DCPS, the observation of individual lessons—the foundation of virtually all school districts' summative instructional evaluation and formative support routines (Cohen & Goldhaber, 2016; Goldring et al., 2015)—provide a view of instructional quality that obscures many central facets that conceptually undergird new standards. Moreover, the curricular materials that teachers employ, such as whether teachers were working with grade level texts in language arts lessons, were also outside the aperture of most observational protocols.

Ultimately, we argue that classroom observations can provide important but insufficient information about the broader terrain of instructional quality that is conceptually aligned with new standards. While some of the challenges we encountered could be mitigated through changes in the implementation of observation systems, such as observing teachers in sequences of lessons rather than individual lessons spaced over time, others were more intractable problems. Some elements of instructional quality, such as collaboration among colleagues across grade levels, are simply not visible or measurable, nor could they likely be made so (Jones et al., 2019; Ronfeldt et al., 2015). If we intend to do more than pay lip-service to these new standards and provide teachers with formative support in meeting those standards, we will have to address the general incongruence between our existing understandings, tools, and routines for promoting instructional improvement in an era of ambitious standards for students.

We develop this argument in three parts. First, we provide some background research on instructional reform, highlighting how current systems emphasize readily observable aspects of teaching (Kane & Staiger, 2012). We then provide a brief account of efforts aimed at instructional reform in the context of DCPS. Finally, we conclude with a description of what a system that supports instructional quality in the Common Core era might entail. Though our proposed system is hypothetical, we identify elements that we believe may be achievable by school districts in the short-term,

with requisite resources, as well as those that could operate as medium and long-term goals for the future. We draw on examples from DCPS not because we want to make systematic or generalizable claims about DCPS's experience, but because in highlighting these authentic dilemmas we can spur more precise thinking about the challenges districts face in developing systems and policies to support and improve instructional quality. We suspect that these practice-based challenges will sound familiar to scholars and practitioners across the country, as will the need to develop more robust responses to these issues.

A Framework for Instructional Improvement in an Era of Assessment

Focusing on the formal, observable aspects of teaching has been the defining feature of American schooling for more than a century. Whether it was school exhibitions in the 19th century (Reese 2013), the “walk-through” in the mid-20th century (Brophy & Good, 1986), or the observation and evaluation protocols of today (Cohen & Goldhaber, 2016; Sykes & Wilson, 2016), there has been widespread and enduring agreement that observing teachers' in the classroom provides valuable information.

Though there is singular agreement on the value of classroom observation, there is considerable variation in the goals that schools seek to accomplish in collecting observational information. These uses range from individual personnel decisions to collective professional development to school improvement (Wise et al., 1984). Systems often initiated to fulfill one of these goals eventually get tasked with fulfilling additional goals, as adapting existing routines is seen as easier than building new, possibly redundant ones (Hill & Grossman, 2013). This becomes problematic because though observation may be central to each of these aims, the same information—and the same routines and systems for collecting it—is not equally suited to achieve all them. The information and tools required to support these various goals differ dramatically in terms of the required aperture, character, and quality. For instance, the detailed descriptive accounts necessary to provide formative information to support instructional quality are likely quite different than the standardized measures necessary for summative evaluations that reward and sanction teachers on the ends of the performance distribution (Hiebert & Stigler, 2017; Hill & Grossman, 2013; Papay, 2012).

To be clear, we are not making a normative case for the value of one kind of information over another. To the contrary, different tools vary in their utility depending on the goal. We may well want teachers to be able to demonstrate certain kinds of observable practices for licensure, tenure, and other

consequential decisions—and scoring them on observable practices may be the most practical and reliable way to capture this information. But the information necessary to determine whether a prospective teacher is minimally qualified to enter a classroom is not necessarily the same kind of information we need to support teachers in demonstrating high levels of instructional quality in the classroom, nor is it likely sufficient in communicating with teachers the specific ways they might improve their instruction, particularly when working within and across grade levels (Hill & Grossman, 2013).

This tension points to a general truism about observational measures regardless of their purpose: to construct an observation protocol is necessarily to emphasize certain observable portions of instruction and, in so doing, to bring some things into focus and to crop out, or blur, others. The choice of what to bring into focus and what to exclude from the frame has consequences not only for the suitability of the protocol's use for different purposes (e.g., development or evaluation) but also for its specific portrayal of the construct of "good teaching." Crucially, every observational instrument, regardless of its components, rests on a theory of instruction that is contested and supported by competing, and often contradictory, empirical claims of effectiveness. For instance, some tools, like Hill et al. (2008) Mathematical Quality of Instruction (MQI) tool, conceptualize "good teaching" by foregrounding the clarity and accuracy of teachers' representation of academic content and their instructional explanations (Hill et al., 2008). Others, like Pianta and Hamre's (2009) CLASS tool, focus on the extent to which teachers foster a positive classroom climate and engage in affirming interactions with students (Pianta & Hamre, 2009). Districts seeking to help teachers develop instructional quality would need to collect different information and provide different supports, depending on which conception of quality they sought to support. Using the MQI or the CLASS in the same classroom would provide distinct information about instruction and what the teacher should do to better support students (Berlin & Cohen, 2018; Praetorius & Charalambous, 2018).

If this were not complicated enough, there are also questions about the alignment between the theory of instruction embedded in the protocol and the theory of instruction implied by the operative standards for students. One upshot of standards-based reform becoming increasingly intertwined with—some would argue inseparable from—accountability systems is that the importance of alignment among standards, curricular materials, and instruction has never been greater (Polikoff, 2017; Polikoff & Porter, 2014). Thus, each new shift in standards and/or corresponding curricula raises the question of whether the observational tools schools use to support instructional quality are still appropriate and conceptually aligned.

Using Observations to Support Instructional Quality

Districts spend considerable time, effort, and money investing in structures and tools *both* to monitor and support teaching (Gitomer, 2008; Papay, 2012). These investments take multiple forms such as organizational routines, communal sense making, and developing common languages and coaching protocols around these practices and the associated goals (Cohen & Goldhaber, 2016; Hill & Grossman, 2013). These outcomes are hard-earned, making the cost of disrupting or abandoning them in the wake of a new reforms very high. Thus, there is a strong incentive to avoid reform churn and, instead, to try and adapt existing tools to new standards and tasks. Indeed, in the wake of new standards, authors of observation rubrics have sought to assist with this adaptation by highlighting existing alignment between the conceptualization of instruction in their tools and college and career ready standards (Grossman, et al., 2014; Danielson, 2012).

Leveraging existing observational measures to promote teaching aligned with new standards makes conceptual sense. If districts want teachers to shift their instruction to support students in meeting new goals, they need tools to clarify to teachers what such instruction might look like. Observation rubrics provide users a common lens for making sense of classroom instruction in the midst of reform (Hill & Grossman, 2013) and when designed carefully, can spotlight important facets of desired practice to teachers (Adnot, 2016; Grissom & Youngs, 2015) and the coaches and instructional leaders who support them (Kraft et al., 2018). In fact, there is empirical evidence that teachers work to align their instruction to the practices delineated in an observation rubric, particularly during classroom observation windows (Phipps & Wiseman, 2020). Coaches and principals, too, shift their foci based on evaluation tools selected by a district (Cohen, et al., 2020), and are differentially helpful to teachers depending on what they choose to highlight in coaching sessions (Allen et al., 2011; Kraft et al., 2018). Thus, given the centrality of observational measures in current evaluation and support systems (Cohen & Goldhaber, 2016; Goldring et al., 2015), the shortcomings of these tools can adversely affect the overall success of reform efforts, especially if they encourage the adoption of practices that emphasize the most visible elements of new pedagogical practices, rather than those most instrumental to supporting students' meeting new standards.

Visible and Invisible Aspects of Instructional Quality

The possibility that observation protocols might be poorly equipped to capture the kind of instruction envisioned by the standards remains relatively

uncharted territory both in the research literature and in practice. Despite the long tradition of classroom observation and the many strengths of current observational protocols, particularly for summative evaluation, traditional observational methods may well be insufficient for formative information and/or coaching and other supports for the broader and less visible aspects of instructional quality. In the context of newly adopted standards, we know very little about what constitutes “Common Core aligned” teaching and/or whether we can consistently observe it using existing rubrics. This, however, has not stopped districts and states from continuing to require classroom observations as part of annual teacher evaluations (Polikoff, 2017).

Most observation protocols developed in the last decade use the visible elements of classroom practice as the sole source of evidence of instructional quality (Kane & Staiger, 2012). The Danielson Framework for Effective Teaching (FFT) represents an exception, with domain one (planning and preparation) and domain four (professional responsibilities) incorporating some of the “behind the scenes’ work of teaching” (The Danielson Group, 2017). Though the framework itself recognizes these additional dimensions of instructional quality, very few states actually incorporate scores from these domains in evaluation ratings (Steinberg & Donaldson, 2016).

For the vast majority of states, what a teacher does in front of the class during a lesson is synonymous with instructional quality. Yet, there are clearly aspects of instructional quality called for in College and Career Ready standards that are not visible within the context of a single lesson. We attempt to capture the features of this problem in the visual model below (see Figure 1). The figure illustrates that instructional quality involves a complex interaction between Teachers (T), Students (S), and Materials (M), which is often referred to as the “instructional triangle” (Cohen et al., 2003). This interaction plays out from class to class and lesson to lesson, in a particular sequence of learning experiences. The nature and tenor of those classroom interactions is a function of what teachers actually do in a particular lesson—asking pointed questions, providing a specific and clear instructional explanation, capitalizing on students’ cultural backgrounds to engage them in a lesson. This is the part of instructional quality that is observable through traditional means (as indicated by the solid lines in the figure).

But showing up in a classroom and assuming that this is the whole of instructional quality is a little like watching a single play of a football game and assuming that the well-choreographed, synchronized motion on the field was all the result of spontaneous (and remarkable) improvisation. What is hidden from view is all of the careful, thoughtful, and strategic pre-planning that preceded the executed events observed on the field, or in the classroom. Important features of teachers’ pre-lesson planning that bear directly on

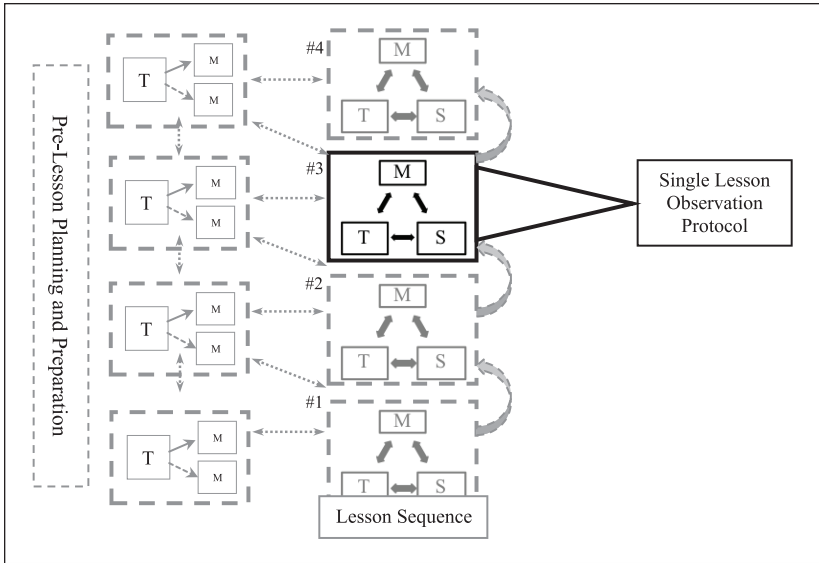


Figure 1. Visible and invisible aspects of instructional quality.

matters of instructional equality include the selection of appropriate materials from multiple potential options, as well as a consideration of how the lesson and materials fit other prior lessons and future planned lessons (represented in Figure 1 by dashed lines). Single lesson observations only capture a small fraction of this dynamic (represented by the solid lines in Figure 1)—a single day, in-classroom interaction between student (S), teacher (T), and materials (M) are visible. Adding more observation days or altering the length of these observations could capture more of this dynamic (turning more of the Figure’s dashed lines solid) but would not change which elements of instructional quality remain visible.

Equally invisible are the ways in which teachers work together to ensure content and instructional experiences are appropriately sequenced across grade levels, so material is neither redundant nor inaccessible (Engel et al., 2013). This is hard to depict visually, but the reader can imagine duplicating Figure 1 for sequences of teachers spanning grade levels—the choices made by one teacher providing the opportunity for future instructional opportunities. In this way, sequences of instructional interactions within and across grades and teachers work in concert to support students in achieving learning goals.

A foundational premise of all these observational measures, as with more contentious value-added measures, is that the individual teacher is a quality

purveyor. That is, instructional quality is operationalized as the characteristic of an individual teacher's classroom which has the effect of reinforcing—conceptually, and in practice—what scholars have described as the “egg crate” model of teaching (Johnson, 2015, 2019; Lortie, 1975). It stems from this view of instructional quality—as an observable characteristic of a classroom—that a single lesson from a single teacher makes sense as the primary unit of analysis, either for evaluation or coaching support. Regardless of the particular content or topic of the lesson, the theory goes, the instructional quality of the teacher will be on display and apparent to the observer.

Even when districts and researchers opt to aggregate lesson level scores over time to create a broader picture of instructional quality and to make inferences about the general quality of teachers' practice (Hill et al., 2012), the arithmetic does not change the assumption that each teacher provides a discrete contribution to the measured educational ends, and that each lesson provides a valuable window into a teacher's instruction. This focus on individual teachers may prove helpful for summative evaluation purposes, as evaluation decisions are indeed made at the individual level. However, this focus on an individual teachers' performance on single lessons may disincentivize collaborative activities, especially in evaluation systems that require a certain percentage of teachers be assigned to each evaluation tier. Others have noted that while collaboration has tremendous potential upside for teachers and students, effective collaboration is difficult to measure, in either formative or summative ways (Jones et al., 2019; Saunders et al., 2009). The challenges in developing formative systems that also incentivize and support breaking down the walls of the egg crate school may be worth the effort, given multiple studies that indicate an association between high-quality teacher collaboration and changes in teacher practice and student achievement (Goddard et al., 2007; Jones et al., 2019; Ronfeldt et al., 2015; Vescio et al., 2008; Wiley, 2001).

As Figure 1 highlights, we need to widen the aperture of our formative systems of observation and corresponding coaching supports in order to accommodate both capacious notions of student learning (as codified in state standards) and a complete view of instructional quality. If not, we risk inscribing an overly narrow conception of instructional quality that ignores the involved instructional work that precedes a teacher's classroom perform, as well as the potentially complex relationships between lessons and among teachers. These aspects of quality pushed out of the aperture may well be foundational to the goal of students achieving the most ambitious learning standards.

In the next section, we show that these challenges are not just hypothetical, but they were practical obstacles for work we were doing with DCPS to

use observational protocols to understand how teachers adapted their practice to new learning standards. Extrapolating from these challenges, we identify the specific tensions between newer standards, existing observation tools, and current understandings of instruction and instructional change. Identifying these friction points is the first step to developing the new tools and conceptions to support instructional change.

Supporting Instructional Quality in Common Core Aligned Teaching

Beginning in the 2016 to 2017 academic year, DCPS implemented a district-wide professional development program called Learning Together to Advance our Practice (LEAP), designed to support teachers in helping students meet CCSS in mathematics and language arts. DCPS had done important work on teacher evaluation with IMPACT (Dee & Wyckoff, 2015) and district leadership were invested in creating systems of teaching development, too. The district was, therefore, an ideal site for trying to understand the implementation and effects of a content specific program designed to support teachers through the transition to new standards. DCPS trained 600 “LEAP leaders” to support content-based teams at their respective schools, which met weekly for a ninety-minutes to support teacher analysis of student work and instructional materials, as well as to build content-based pedagogical knowledge and skills. LEAP leaders also engage in classroom observations and structured debriefs to provide teachers with formative feedback.

To measure Common Core aligned instructional quality in this context, we collaborated with Student Achievement Partners (SAP),¹ a non-profit organization dedicated to helping teachers work toward new learning standards. In partnership with SAP, we spent a year designing observation protocols that could reliably measure instructional practices aligned to the expectations of the CCSS in ELA and Mathematics. The focal point of these efforts was adapting SAP’s Instructional Practice Guides (IPGs), a set of coaching rubrics designed to support “Common Core aligned instruction,” into standardized observation tools that could be used reliably at scale.

The IPGs focus on three subject-specific shifts in both mathematics and ELA instruction that are necessary to meet the learning demands of CCSS.² According to CCSS, in mathematics, teachers should: (1) focus narrowly on grade-specific foundational topics with adequate depth, (2) build content connections both across and within grades, and (3) attend to conceptual understanding, procedural skill and fluency, and application of knowledge with equal intensity; and in ELA, teachers should: (1) regularly focus on

grade-level complex texts and vocabulary with their students, (2) ground reading, writing, and speaking skills in text-based evidence, and (3) build content knowledge through the inclusion of content-rich, non-fiction texts. These practices were not entirely absent from prior generations of standards but were emphasized to a lesser degree than in the CCSS.

In consultation with SAP, we developed the Instructional Practice Research Tool (IPRT) for mathematics and ELA as a complement to the IPGs to understand the degree to which teaching in DCPS reflected the shifts described above. Over 2 years, these observation tools were used to analyze teaching practices in over 1,000 lessons. Below, we focus on our experience in using these tools and the concrete issues raised about visible and invisible aspects of instructional quality.

Visible Elements of Instructional Quality

We were able to identify and reliably score multiple facets of teaching that are conceptually aligned with and in service of CCSS goals for students. In mathematics, for example, mathematical practice 3 (MP3) requires students “construct viable arguments and critique the reasoning of others.” We created two distinct scales to measure this. The first focused on how teachers provided students opportunities for justification and critique. The second measured students’ depth of argumentation and reasoning, privileging student responses that included a logical sequence of mathematical reasoning.

ELA standards expect students to carefully read text. To capture whether teachers pose general questions about text (“What is the author’s purpose?”) versus questions that direct students to a particular aspect of text (“Why does the author include the images on page 21?”), we created two distinct scales. One distinguishes between non-text questions that ask students to draw on their own experience (e.g., “In the text, Casey strikes out. Describe a time when you failed at something.”) rather than support textual claims. The other distinguishes between text-based questions that ask students to use the text more generally (e.g., “What is the theme of the story?”) and those that require students to understand specific elements of the text (e.g., “How does the main character’s conflict help the reader understand the theme of this passage?”).

Including two scales targeting separate aspects of text-based questioning allowed us to better identify the range of practice in ELA classrooms in DCPS. The standards foreground the notion that students should critically engage with text to understand it more deeply. Rather than posing questions that might be asked of any text, the standards call for questions and tasks that require students to dig into the content and language of a particular text. Our separation of text-dependent versus text-specific questioning allowed us to

parse whether teachers in DCPS were asking students to respond to text-based questions and tasks broadly, or in the more specific way the standards envision.

Challenges in Measuring Common Core Aligned Teaching

Although we were able to capture some important elements of teaching aligned with CCSS, the observation protocols we developed were ultimately insufficient for measuring the full range of the shifts delineated by SAP and others who support teachers in shifting instruction to support new learning goals for students. These challenges coalesced around the visibility of curricular resources and instructional materials used, coupled with temporal challenges and determining the appropriate unit of analysis.

Temporal issues and appropriate units of analysis. Determining the unit of analysis is an integral part of every measurement decision because it indicates the level at which conclusions will be drawn (Cronbach et al., 1972). Protocols rely on different sampling parameters to generalize about a teacher's "practice" (Bell et al., 2012; Herlihy et al., 2014; Hill et al., 2012). Despite considerable work over the last decade, questions remain about the ability to make accurate inferences about instructional quality from classroom observation (Cohen & Goldhaber, 2016).

Common Core aligned teaching raises additional questions about the unit of analysis because of the standards' emphasis on sequencing knowledge both within a lesson and across broader units of study. Many of the shifts that teachers might make in their instruction to fulfill this requirement are not visible at the lesson level. For example, one shift in mathematics entails making connections between content both within and across grade levels. In the context of a single lesson in DCPS, outside raters were unable to determine whether content meaningfully built on previous grade-level topics or material learned in earlier grades. While we were able to measure explicit connections to prior content (e.g., "We've been working on . . . ; today we'll use what we've learned to. . ."), we were unable to measure the full intent of the shift in a fulsome way.

Analyzing coherence across broader units of study using a single lesson as the unit of analysis also proved challenging. The new mathematics standards require teachers to "pursue conceptual understanding, procedural skill and fluency, and application with equal intensity" (Council of Chief State School Officers, 2010), a task SAP terms "rigor." This balance between conceptual understanding and procedural fluency is expected at the topic level (e.g., within a series of lessons on fractions), rather than at the individual lesson

level (e.g., within one lesson on adding fractions with common denominators). In DCPS classrooms, we observed some lessons that included all aspects of rigor, but most focused on just one or two. This makes good sense given the way lessons are often sequenced over time, but it complicates inferences about “quality” based on individual lessons, even in instances in which individual lessons are aggregated over time. Observing individual lessons that were not necessarily in a sequence did not allow our raters to know if teachers were achieving a balance of rigor across the topic as a whole, which was necessary to score the rigor of the instruction in a way that was consistent with the standards. Most often we found that, within a given topic, teachers seemed to teach separate, complementary lessons—some focused on developing conceptual understanding, others on procedural skill or application of knowledge. Our attempts to observe and score single lessons meant that the rubric score could not accurately capture the instructional quality of teachers who achieved this shift at the topic, rather than lesson, level, despite the fact that such instructional decisions might make good pedagogical sense.

We encountered a comparable problem in observing elements of Common Core aligned ELA practice. One of the primary shifts in ELA calls for a very specific balance between fiction and non-fiction texts. For example, in elementary schools, the recommended ratio is 50/50 across the year. How this final balance is obtained is a question of instructional planning over the course of a school year, rather than a single lesson, making that balance unobservable using our observational tool, the IPRT-ELA, and all other widely used observation protocol procedures (Marzano, 2007; Pianta & Hamre, 2009), including those that are purportedly “Common Core aligned” (Danielson, 2013).

To be clear, these aspects of instructional quality are not literally invisible. A coach or principal could observe the balance of texts or mathematical rigor if they were in a teacher’s classroom every day over the course of a unit in mathematics, or year in ELA. However, these shifts are *functionally invisible* in the context of observational systems that focus on sampling individual lessons.

Beyond the issue of whether and how this instructional balance is achieved, there is an additional temporal component that is not addressed by existing observational frameworks. The extended period of instructional interest within the standards themselves means the number of possible ways and means of securing the desired instructional result are far more numerous than the comparatively small set of optimal instructional practices highlighted by observational frameworks that focus on individual lessons. How to track and evaluate those decisions—even putting aside questions of how to assess the relative strength of alternative approaches or alternative sequences—is far

beyond the capacity of any available observational framework and, in truth, beyond anything explored in the existing research literature. While many scholars have studied the number of scored lessons needed to make inferences about teacher quality (Hill et al., 2012; Mashburn et al., 2014), we know next to nothing about the optimal spacing or intervals between such lessons as they pertain to inferences about instructional quality, more broadly defined. We also know little about how coaches make sense of and support instructional quality based on what and when they observe the teachers with whom they work (Kraft et al., 2018). These are both important areas for future work.

Materials and curricular resources. In addition to the temporal issues detailed above, other facets of standards-aligned instruction also proved difficult to observe. A major focus of the CCSS standards in ELA is using complex, grade-level text, which necessitates a teacher making a thoughtful text selection based on contextual knowledge of students. Measuring this construct through classroom observation was difficult.

The standards delineate a three-pronged approach to judging text complexity for students: (1) A teacher may consider a range of texts that meet the appropriate quantitative criteria (e.g., using Lexile levels); (2) The teacher reads each text carefully, paying attention to the meaning, structure, language, and knowledge demands (the qualitative criteria for literature); (3) The teacher considers the match between text, reader, and task (e.g., what background knowledge might make this text less challenging?) and balances the goals of the unit and lesson for knowledge and skill development, in light of the learning goals called for by the standards.

When it comes to evaluating a teacher's performance on text selection, only the final selection itself is visible to observers. The process by which a teacher arrived at that selection, by contrast, is totally invisible to a classroom observer and it is difficult to imagine how it could be made visible: teachers may engage in this process during planning sessions, perhaps using district-provided lists of complex texts by grade level (Valencia et al., 2014); or they may do so through more informal consultation with colleagues or trusted mentors; or through reference to their past experience with similar students. And yet, these different sources of information and distinct decision-making processes matter for coaching or supporting the teacher in this skill.

In working with SAP, we considered varied approaches to examining text selection, given its unobservable qualities. Instead of directly observing a teacher's process, raters identified the text used in a lesson and rated it quantitatively. While this approach is useful in determining the overall complexity of the text, it does not capture any dimension of its contextual

appropriateness. This approach fails to consider both qualitative complexity of the text, which would require the full text and human readers, rather than software programs, as well as an assessment of the text-reader match, which would require knowledge of students that outside observers cannot possess. Thus, while we were able to capture some information about text, we were not able to measure the full intent of the CCR standards' focus on text complexity.

Similarly, the standards in mathematics call for attention to teaching that fully captures the depth of grade level cluster and content standards (Council of Chief State School Officers, 2010). To engage in this practice, teachers must be able to classify different approaches to mathematical tasks by grade level. For instance, while multiplication with a product within 100 is aligned to the standards in third grade, an identical problem that asks students to make mathematical comparisons is aligned to the standards in fourth grade. So, for example, "Gabriel had three pots. He planted seven seeds in each pot. How many seeds did he plant?" is aligned to third grade mathematical standards, while, "Gabriel planted seven seeds. His mom planted three times as many seeds as he did. How many seeds did his mom plant?" is aligned to fourth grade mathematical standards. This measure of *what* is being taught, rather than simply *how* it is being taught, represents a broader shift to the elements of ambitious teaching involved in the planning and preparation phase of a lesson across both content areas. This aspect of instruction is enacted outside live classroom practice, in planning sessions, and is also contingent on which curricular materials teachers have access to and which ones are promoted in a given district.

As we detailed in this section, the instructional shifts to using more complex texts in the classroom and to ensuring that lessons and units reflects the full rigor of the standard targeted—both key features of CCSS standards in ELA and mathematics—represents a shift in how we conceptualize the scope of available evidence related to instructional quality.

To oversimplify these findings only slightly: our current protocols can capture some of the *how* of teaching but cannot see the *what or when* of instructional quality. That is, current protocols—developed in an era rightly concerned with identifying pedagogical practices that can be reliably measured for summative evaluation decisions—draw our attention to the visible instructional moves undertaken by a teacher. While discrete lesson observations were sufficient for determining whether teachers had these practices in their repertoires, they are insufficient for responding to our new standards that ask teachers to engage more deeply with questions of material selection (i.e., *what*) and sequencing within and across years (i.e., *when*).

Most observation rubrics reflect the position that instructional quality should be attributed to individual teachers. Our work in DCPS suggests some teaching practices aligned to CCR standards are invisible, given current observation protocols. In mathematics, a key instructional shift requires a balance of procedural, conceptual, and application-oriented instruction across a unit of study, which cannot be observed without observing multiple lessons from the same unit. In ELA, a key instructional shift requires thoughtful text selection based on knowledge of students' experiences and readiness, which also cannot be observed without access to the planning process.

Next Steps for Conceptualizing and Measuring Implementation of Standards-Based Reforms

New, more ambitious learning standards for students require new and more ambitious forms of teaching. If districts are to develop teachers as they pursue these more ambitious forms of instruction, then they would likely benefit from tools capable of accurately analyzing and ultimately supporting instruction in light of these new goals. If districts only employ observation protocols and other supports that either misidentify, fail to identify, or completely ignore aspects of ambitious instruction, districts run the risk of circumscribing ideas about what ambitious teaching looks like or, worse, disincentivizing teachers from pursuing the full-range of instructional practices necessary to support students in meeting next-generation standards.

We are not suggesting that classroom observation has no place measuring instructional quality aligned with new standards, especially for summative evaluation and assessment purposes. Indeed, many standards-aligned teaching practices are visible in a single lesson. In mathematics, we captured the opportunities teachers provided for students to justify and critique, as well as the ways in which students took advantage of those opportunities. In ELA, we reliably measured text-based questioning with respect to the demands of the standards, classifying classroom questions as non-text, text-dependent, or text-specific. In the wake of a new policy reform, however, researchers and policy makers should thoughtfully consider the extent to which previous measures are aligned to new policy goals and adapt accordingly.

Our central goal in this essay is to make clear that many distinctive features of Common Core aligned teaching cannot be supported with traditional observational systems. And this is for good reason: new standards are more ambitious precisely because they place demands on instruction that ask teachers to move beyond atomized thinking about the presentation of individual lessons. However, there is nothing that prevents a state or district from

requiring non-evaluative, formative observations and coaching supports focused on some of the issues we raise here, including the texts or other instructional materials used, or oriented around the interconnectedness and sequencing of lessons, over time. In some ways, the policies have been too focused on summative evaluation with too little attention paid to supporting teachers as they undertake the difficult task of transitioning their instruction. We argue that now is an important time to take a broader view of the issue and develop correspondingly robust systems of formative support for teacher development aligned with new learning standards. These two systems can and should work in tandem and to promote instructional quality in its broadest sense.

Addressing all the shortcomings of our current tools will be an unquestionable challenge and require the development of new tools and, likely, the adoption of new policies to implement their use. Indeed, we believe there are elements of ambitious instruction that are fundamentally incommensurable with our current systems of instructional support grounded in organizational practices of teacher observation. That said, we believe it is possible to widen the aperture of our current system of observation system to bring into view more aspects of teacher practice we include in Figure 1, with district resources and support and policies that foreground the importance of invisible aspects of instructional quality.

These policies and supports could take a variety of forms ranging from the commitment of financial resources to develop new tools or reporting systems to capture additional information; the commitment of personnel resources necessary to train people on how to use and implement these tools; and the adoption of explicit state, district, or school policies (or the promulgation of specific guidance) that specify the minimum expectation around the use of these tools and the incorporation of these practices. We would note that these are policies that accompanied the adoption of CCSS in the first place—that is, states adopted the standards, committed resources to new textbooks, and facilitated various forms of professional development. We are suggesting that another round of similar investments is necessary to create the corresponding practices to support teachers' instructional development, as well as an empirical base about the degree to which and ways in which coaches and other support providers could widen the aperture of their view of instructional quality to include more about the what and when of teaching practices.

With the two specific challenges raised by our work in DCPS in mind, we offer suggestions on how these issues might be mitigated in the short-term through the incorporation of explicit practices aimed at developing multiple vantage points of teacher instructional practice. These short-term fixes, however, must be succeeded by policies that reframe the whole system of

instructional support in a way that reduces our reliance on what can be *seen* of instructional practice and toward a more holistic and collaborative view of teaching.

Temporal Issues: Capturing the “When” of Instruction

Just as we are not arguing that observation has no role to play in supporting teachers’ instructional practice, we are not arguing that lesson level observations should cease playing an important role in systems of instructional evaluation. We may well want teachers to be able to demonstrate particular practices in a given lesson, and districts do need to be able to observe and measure those practices reliably, a hallmark of evaluation policies in DCPS and around the country, in order to make fair personnel decisions (Cohen & Goldhaber, 2016). However, our experience in DCPS made clear that observations and current observation systems cannot get us all the way to supporting some of the broader and less readily observable aspects of instructional quality described here. Reducing reliance on lesson level snapshots likely necessitates a multi-pronged approach that involves shifting to larger units of observations, as well as developing ways for teachers to make visible key aspects of their instructional planning and decision-making. These changes would allow for more (and more kinds of) information to be captured. If we want formative assessment systems for teachers, then incomplete information can at best be unhelpful in a teacher’s efforts to improve, and at worst, distort the teacher’s focus toward visible aspects of instruction.

For instance, new sampling approaches for observational data might better capture the interrelated character of individual lessons. An ideal policy might structure district policies around observations at the curricular topic level. Collecting data for an entire sequence of lessons focused on a particular topic would help answer questions about the ways in which lessons were sequenced to scaffold student understanding over time and could also provide information about the weight teachers give particular topics or texts. This approach, in which several lessons are observed back-to-back, would require much higher levels of coordination between administrators and teachers than current one-off observation systems, but it may ultimately yield more useful information about instructional quality than would the same number of lessons observed in spaced-out intervals over the course of the year. This is an empirical question that we have not tested, but at the very least its structure is theoretically aligned to the goals articulated in the Common Core Standards.

In addition to modifying existing observation systems to shift from lesson level snapshots, districts may also benefit from developing formative

assessment and instructional support systems that require teachers, coaches, and other school-based leaders collect a wider range of evidence that speak directly to the components of standards-aligned teaching we have highlighted here. These systems do not require, but would be helpfully reinforced by, the development of explicit policies. These policies might detail the kinds of information that could be used or collected for these purposes. They will also need to grapple with how best to analyze the correspondence between the intended plan and the actual lesson—taking into account that improvised deviation is sometimes necessary and often a mark of good pedagogy. When we begin to consider the amount and kinds of information necessary to capture and support more ambitious forms of instruction, it becomes clear how limited our current systems are for these purposes. A great deal of research on all these elements is required before we will have a fully robust system, but we will only gain this knowledge if we raise our expectations for these systems.

Finally, and perhaps most radically, districts may benefit from models that zoom out from the traditional focus on individual teachers and instead bring into view the instructional quality provided by teacher teams that work together over time. We need to create school-level systems and corresponding policies for promoting instructional quality, which necessitates moving beyond thinking of individual teachers as quality purveyors (Johnson, 2019). Though teachers are instrumental in supporting students' learning, an individual teacher's work will always be amplified when we support coordination across classrooms and grade levels, and when we create systems that prioritize and reward such efforts. Coaches could work with groups of teachers to ensure coherent and rigorous sequences of learning over time, pushing teachers to explain the ways in which their instruction builds on prior learning experiences and readies students for the expectations in future grades.

Indeed, DCPS' LEAP policies were in many ways constructed in service of such goals. DCPS built LEAP teams to be content-focused but span grade levels; a LEAP leader might work with an ELA team composed of teachers from the 3rd, 4th, and 5th grades. The structure makes conceptual sense, but if we want to generate coherent vertical trajectories of teaching and learning, then we likely need corresponding tools to help guide and support that work. This, too, will necessitate instrument development and future research on how such tools promote instructional quality across classrooms and over time. Such models also present logistical challenges, including a stable teacher workforce, a rare commodity in many schools, even in the best circumstances. In addition, such systems also necessitate early feedback on the results of implementation processes as these results

might counsel a substantial change in direction or approach. This kind of dynamism should be welcomed and rewarded, but it poses challenges for those looking to measure and support the developmental capacities of instructional teams (Jones et al., 2019).

Material and Resources Selection: Capturing the “What” of Instruction

Our work in DCPS surfaced the ways in which classroom observations can treat the materials used and the content of instruction as either static and decontextualized (e.g., rating a text’s lexile level) or as immaterial (i.e., rating the quality of questions a teacher asks about a text without considering the appropriateness of the text for a particular group of students). Both approaches run counter to the central role that materials play in the “instructional triangle” and many other conceptual models of instructional quality (Cohen et al., 2003; Hiebert & Stigler, 2017).

Artifact analysis may prove a powerful complement to classroom observation in measuring instructional quality, to capture some of the materials and content so central to the instructional triangle we feature in Figure 1 (Cohen et al., 2003). One approach, used primarily to assess science teaching, asks teachers to respond to written reflection prompts and submit evidence of student work and classroom materials (Kloser et al., 2017). Raters then assess these portfolios or “notebooks” against instructional criteria (Martínez et al., 2012). Artifact analysis would have allowed us to gather richer information about teacher decision-making, particularly with respect to text selection. In mathematics, it might have made accessible a new dimension of student reasoning, as well as some mathematical practices obscured from observers, such as whether students used mathematical tools strategically. Districts, for instance, could require teachers to submit and review notebooks with coaches or with other instructional support providers, which would allow districts to signal the importance of elements of practice that go beyond what is observable in a classroom.

We believe that shifting policies to include formative assessments that extend beyond observable teaching practices, if implemented well, would appreciably improve the alignment between our new learning goals and our systems for understanding and promoting high quality instruction. As a path forward, the research community must think creatively about multi-measure systems, including more formative assessment systems that promote and support instructional quality at scale. Modifying existing systems and tools may well make instructional assessment and support systems more robust and better aligned with ambitious standards, but such modifications would

also require considerable work and resources. The trade-off might be worth it, though, if it meant that our assessment systems provided the kind of information necessary to support the continual development of instruction that moved students toward the ambitious learning goals outlined in new standards (Hiebert & Stigler, 2017).

Conclusion

In this paper, we have tried to accomplish two goals. First, we highlight the many ways in which current observation protocols that are in wide use nationwide give short shrift to the ambitious instruction imagined by newly adopted state standards. Because teachers take important cues from these protocols, and because districts use them to provide instructional support (and at times evaluation and assessment) for teachers, this is a budding policy issue that needs to be addressed. We have also tried to suggest some of the directions that policy might go in meeting these challenges, in large part by illuminating the gaps in our current systems. Though we have refrained from particular prescriptions, we have tried to highlight the extent to which a variety of approaches already in use in to varying degrees in other contexts (e.g., collecting artifacts, working with coaches, reflections on teaching) could be adapted to fill these gaps. But to do this work we need to develop new tools and protocols that communicate and capture a more fulsome picture of ambitious instruction. We will also need researchers to evaluate the effectiveness of these new tools and to identify best practices in their development and application.

We recognize that in proposing these changes we are making a big ask not only of researchers to develop these tools, but also of the districts that would seek to find ways to implement and support this work. But it would seem that this is the steady work necessary to have our supports for instructional improvement match our aspirations for student learning.

Acknowledgments

The authors want to thank everyone at Student Achievement Partners for their partnership in this work. In particular, Jessica Eadie, Lisa Goldschmidt, Susan Pimentel, Jason Zimba, and Amanda Vitello provided invaluable feedback on earlier versions of the manuscript. Barbara Beske and Rachel Etienne spent countless hours discussing videos with us, as we worked together to develop the measures described here. Chong-Hao Fu and his colleagues at Leading Educators also provided insight and feedback about CCR-aligned teaching in mathematics. Our many partners in DCPS were instrumental in facilitating this work, including Jason Kamras, Scott Thompson, Liz McCarthy, and Alison Williams.


Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was funded through the generous support of the Charles and Lynn Schusterman Family Foundation and the Overdeck Family Foundation. In addition, our work was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant #R305B140026 to the Rectors and Visitors of the University of Virginia. The opinions expressed are those of the authors and do not represent views of our funders.

ORCID iD

Julie Cohen  <https://orcid.org/0000-0001-6794-7993>

Notes

1. The first, third, and fourth author worked closely with SAP throughout the LEAP study.
2. We focus on CCSS in mathematics and ELA because our work in DCPS focused on these subjects. We recognize there are other “next generation” or “college and career ready” standards in other subjects, including science and social studies.

References

- Adnot, M. (2016). *Effects of evaluation and incentives on instructional practice: Evidence from the District of Columbia Public Schools' IMPACT teacher evaluation system* (Unpublished Doctoral dissertation). University of Virginia.
- 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.
- Bell, C. A., Gitomer, D. H., McCaffrey, D. F., Hamre, B. K., Pianta, R. C., & Qi, Y. (2012). An argument approach to observation protocol validity. *Educational Assessment, 17*(2–3), 62–87. <https://doi.org/10.1080/10627197.2012.715014>
- Berlin, R. & Cohen, J. (2018). Understanding instructional quality through a relational lens. *ZDM: The International Journal on Mathematics Education, 50*(3), 367–379.
- Brophy, J., & Good, T. (1986). Teacher behavior and student achievement. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 376–391). Macmillan.
- Cohen, D. K., & Hill, H. C. (2008). *Learning policy: When state education reform works*. Yale University Press.

- Cohen, D. K., Raudenbush, S. W., & Ball, D. L. (2003). Resources, instruction, and research. *Educational evaluation and policy analysis*, 25(2), 119–142.
- Cohen, J. & Goldhaber, D. (2016). Building a more complete understanding of teacher evaluation using classroom observations, *Educational Researcher*, 45(6).
- Cohen, J., Loeb, S., Miller, L., & Wyckoff, J. (2020). Policy implementation, principal agency and strategic action: Improving teacher quality in New York City middle schools. *Educational Evaluation and Policy Analysis*, 43(1), 134–160.
- Council of Chief State School Officers. (2010). *Common core state standards*. National Governors Association Center for Best Practices.
- Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. Wiley.
- Danielson, C. (2012). Observing classroom practice. *Educational Leadership*, 70(3), 32–37.
- Danielson, C. (2013). *The framework for teaching: Evaluation instrument*. <https://www.danielsongroup.org>
- 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>.
- Engel, M., Claessens, A., & Finch, M. A. (2013). Teaching students what they already know? The (mis) alignment between mathematics instructional content and student knowledge in kindergarten. *Educational Evaluation and Policy Analysis*, 35(2), 157–178.
- Gitomer, D. H. (2008). *Measurement issues and assessment for teaching quality*. Sage.
- Goddard, Y. L., Goddard, R. D., & Tschannen-Moran, M. (2007). A theoretical and empirical investigation of teacher collaboration for school improvement and student achievement in public elementary schools. *Teachers College Record*, 109(4), 877–896.
- Goldring, E., Grissom, J. A., Rubin, N., Neumerski, C. M., Cannata, M., Drake, T., & Schuermann, P. (2015). Make room value added: Principals' human capital decisions and the emergence of teacher observation data. *Educational Researcher*, 76(2), 173–208.
- Grissom, J. A., & Youngs, P. (Eds.). (2015). *Improving teacher evaluation systems: Making the most of multiple measures*. Teachers College Press.
- Grossman, P., Cohen, J., Ronfeldt, M., & Brown, L. (2014). The test matters: The relationship between classroom observation scores and teacher value-added on multiple types of assessment. *Educational Researcher*, 43(6), 293–303.
- Herlihy, C., Karger, E., Pollard, C., Hill, H. C., Kraft, M. A., Williams, M., & Howard, S. (2014). State and local efforts to investigate the validity and reliability of scores from teacher evaluation systems. *Teachers College Record*, 116(1), 1–28.
- Hiebert, J., & Stigler, J. W. (2017). Teaching versus teachers as a lever for change: Comparing a Japanese and a US perspective on improving instruction. *Educational Researcher*, 46(4), 169–176.

- Hill, H. C., Ball, D. L., & Schilling, S. G. (2008). Unpacking pedagogical content knowledge: Conceptualizing and measuring teachers' topic-specific knowledge of students. *Journal for Research in Mathematics Education*, 39(4), 372–400.
- Hill, H. C., Blunk, M. L., Charalambous, C. Y., Lewis, J. M., Phelps, G. C., Sleep, L., & Ball, D. L. (2008). Mathematical knowledge for teaching and the Mathematical Quality of Instruction: An exploratory study. *Cognition and Instruction*, 26(4), 430–511. <https://doi.org/10.1080/07370000802177235>
- Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the G-study. *Educational Researcher*, 41(2), 56–64.
- Hill, H. C., & Grossman, P. (2013). Learning from teacher observations: Challenges and opportunities posed by new teacher evaluation systems. *Harvard Educational Review*, 83(2), 371–384. <https://doi.org/10.17763/haer.83.2.d11511403715u376>
- Johnson, S. M. (2015). Will VAMS reinforce the walls of the egg-crate school?. *Educational Researcher*, 44(2), 117–126.
- Johnson, S.M. (2019). *Where teachers thrive: Organizing schools for success*. Harvard University Press.
- Jones, N. D., Bettini, E., & Brownell, M. (2019). Competing strands of educational reform policy: Can collaborative school reform and teacher evaluation reform be reconciled? *The Elementary School Journal*, 119(3), 468–486.
- 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.
- Kloser, M., Borko, H., Martínez, J. F., Stecher, B., & Luskin, R. (2017). Evidence of middle school science assessment practice from classroom-based portfolios: Dimensions of quality science assessment. *Science Education*, 101(2), 209–231. <https://doi.org/10.1002/sce.21256>
- 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.
- Lortie, D. C., & Clement, D. (1975). *Schoolteacher: A sociological study*. University of Chicago Press.
- Martínez, J. F., Borko, H., Stecher, B., Luskin, R., & Kloser, M. (2012). Measuring classroom assessment practice using instructional artifacts: A validation study of the QAS notebook. *Educational Assessment*, 17(2–3), 107–131. <https://doi.org/10.1080/10627197.2012.715513>
- Marzano, R. J. (2007). *The art and science of teaching: A comprehensive framework for effective instruction*. ASCD.
- Mashburn, A. J., Downer, J. T., Rivers, S. E., Brackett, M. A., & Martinez, A. (2014). Improving the power of an efficacy study of a social and emotional learning program: Application of generalizability theory to the measurement of classroom-level outcomes. *Prevention Science*, 15(2), 146–155.

- Papay, J. (2012). Refocusing the debate: Assessing the purposes and tools of teacher evaluation. *Harvard Educational Review*, 82(1), 123–141.
- Pepin, B., & Haggarty, L. (2001). Mathematics textbooks and their use in English, French and German classrooms: A way to understand teaching and learning cultures. *Zentralblatt Für Didaktik Der Mathematik*, 33(5), 158–175. <https://doi.org/10.1007/BF02656616>
- Phipps, A. R., & Wiseman, E. A. (2020). Enacting the rubric: Teacher improvements in windows of high-stakes observation. *Education Finance and Policy*. Advance online publication. https://doi.org/10.1162/edfp_a_00295
- Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher*, 38(2), 109–119. <https://doi.org/10.3102/0013189X09332374>
- Polikoff, M. S., & Porter, A. C. (2014). Instructional alignment as a measure of teaching quality. *Educational Evaluation and Policy Analysis*, 36(4), 399–416.
- Polikoff, M. S. (2017). Is Common Core “working”? And where does Common Core research go from here? *AERA Open*, 3(1), 2332858417691749. <https://doi.org/10.1177/2332858417691749>
- Praetorius, A. K., & Charalambous, C. Y. (2018). Classroom observation frameworks for studying instructional quality: Looking back and looking forward. *ZDM*, 50(3), 535–553.
- Reese, W. J. (2013). *Testing wars in the public schools*. Cambridge Harvard University Press.
- Remillard, J. T., & Heck, D. J. (2014). Conceptualizing the curriculum enactment process in mathematics education. *ZDM*, 46(5), 705–718.
- Ronfeldt, M., Farmer, S. O., McQueen, K., & Grissom, J. A. (2015). Teacher collaboration in instructional teams and student achievement. *American Educational Research Journal*, 52(3), 475–514.
- Saunders, W. M., Goldenberg, C. N., & Gallimore, R. (2009). Increasing achievement by focusing grade-level teams on improving classroom learning: A prospective, quasi-experimental study of Title I schools. *American Educational Research*, 46(4), 1006–1033.
- Steinberg, M. P., & Donaldson, M. L. (2016). The new educational accountability: Understanding the landscape of teacher evaluation in the post-NCLB era. *Education Finance and Policy*, 11(3), 340–359. https://doi.org/10.1162/EDFP_a_00186
- Stein, M. K., & Kim, G. (2011). The role of mathematics curriculum materials in large-scale urban reform: An analysis of demands and opportunities for teacher learning. In J. T. Remillard, G. Lloyd, & B. Herbel-Eisenmann (Eds.), *Teachers' use of mathematics curriculum materials: Research perspectives on relationships between teachers and curriculum* (pp. 57–75). Routledge.
- Sykes, G., & Wilson, S. M. (2016). Can policy (re) form instruction? In C. Bell & D. Gitomer (Eds.), *Handbook of research on teaching* (pp. 851–916). American Educational Research Association.

- The Danielson Group. (2017). *Questions about the evaluation of Domains 1 and 4*. <https://www.danielsongroup.org/questions-about-the-evaluation-of-domains-1-and-4/>
- Valencia, S. W., Wixson, K. K., & Pearson, P. D. (2014). Putting text complexity in context. *The Elementary School Journal*, 115(2), 270–289. <https://doi.org/10.1086/678296>
- Vescio, V., Ross, D., & Adams, A. (2008). A review of research on the impact of professional learning communities on teaching practice and student learning. *Teaching and Teacher Education*, 24(1), 80–91.
- Walkowiak, T. A., Berry, R. Q., Meyer, J. P., Rimm-Kaufman, S. E., & Ottmar, E. R. (2014). Introducing an observational measure of standards-based mathematics teaching practices: Evidence of validity and score reliability. *Educational Studies in Mathematics*, 85(1), 109–128. <https://doi.org/10.1007/s10649-013-9499-x>
- Wiley, S. D. (2001). Contextual effects on student achievement: School leadership and professional community. *Journal of Educational Change*, 2(1), 1–33.
- Wise, A., Darling-Hammond, McLaughlin, M., & Bernstein, H. (1984). *Case Studies for teacher evaluation: A study of effective practices* (N-2133-NIE). Rand.

Author Biographies

Julie Cohen is an Associate Professor of curriculum, instruction, and special education at the University of Virginia's School of Education and Human Development. She studies teachers and teaching, with a focus on how policies and practices support the development of effective instruction.

Ethan Hutt is an Assistant Professor of culture, curriculum, and teacher education at the University of North Carolina at Chapel Hill School of Education. His research focuses on the numbers and metrics that we use to describe, define, and evaluate American schools.

Rebekah Berlin is a Program Manager at Deans for Impact. Her work focuses on measurement and improvement efforts in teacher education.

Emily Wiseman is a doctoral student in education policy at the University of Virginia's School of Education and Human Development. She studies policies that support teacher development, especially coaching initiatives.