

Using Classroom Observations in the Evaluation of Special Education Teachers

Nathan D. Jones

Boston University

Courtney A. Bell 

University of Wisconsin-Madison

Mary Brownell 

University of Florida

Yi Qi

Educational Testing Service

David Peyton 

Appalachian State University

Daisy Pua 

University of Florida

Melissa Fowler

Steven Holtzman

Educational Testing Service

We examine whether one of the most popular observation systems in teacher evaluation—the Framework for Teaching (FFT)—captures the range of instructional skills teachers need to be effective. We focus on the case of special educators, who are likely to use instructional approaches that, although supported by research, are de-emphasized in common observation systems. Drawing on 206 lessons from 51 teachers, we compare FFT scores to an observation system from special education. We find that FFT’s psychometric properties are consistent with previous studies, but the system is limited in assessing the quality of instructional practices used in special education. We discuss implications of these findings for two practical uses of observations—supporting teacher development and informing human capital decisions.

Keywords: *assessment, educational policy, evaluation, instructional practices, special education, validity/reliability, descriptive analysis, factor analysis, observational research, psychometrics*

Introduction

VALIDATED observation systems have figured prominently in the teacher evaluation and development landscape over the past decade. At present, all 50 states have policies requiring that formal observations be included in teacher evaluation systems. Incentivized by federal efforts to improve teaching (e.g., Race to the Top, Every Student Succeeds Act), states and districts have adopted observation systems that can be used in standardized ways with all teachers (Steinberg & Donaldson, 2016).

Districts commonly use observation systems for two purposes—to inform human capital decisions and to guide teacher improvement efforts. When included in formal teacher evaluation, observation scores provide data on the quality of a teacher’s professional practice and complement other information, such as student academic progress. When used for teacher improvement, observation scores guide administrators and teachers in identifying instructional practices where teachers need further professional development. More generally, observation systems often become the organizational framework through which districts focus their instructional program (e.g., Donaldson & Woulfin, 2018), signaling to teachers and administrators what *good teaching* looks like.

Although there is a large, growing body of research on classroom observation systems in teacher evaluation, an important question remains largely unanswered: How well do common observation systems¹ capture the full range of teaching skills and behaviors needed to provide effective instruction? Researchers have demonstrated that effective teaching depends on a variety of factors, including the specific content taught and the needs of individual learners (e.g., Connor & Morrison, 2016; Morgan et al., 2015; Sonnenschein et al., 2010; Xue & Meisels, 2004). Notably, when students are first mastering a skill or new content, it is often beneficial for them to have teachers take on an active role, providing what the literature refers to as explicit and systematic instruction. For students with disabilities (SWDs), many continue to benefit from teachers playing an active role, even when solving higher-order problems (e.g., summarizing text, discussing essential questions in texts). Conversely,

students with high levels of prior knowledge or skill may need less initial scaffolding from teachers before engaging in instructional approaches that prioritize student exploration. From this perspective, neither *teacher-directed* or *student-directed* instruction is inherently more effective (nor are they always the most appropriate instructional approach for specific students). Instead, successful teachers are those who can deploy a full range of instructional methods to meet their students’ individualized needs.

But the observation systems required by teacher evaluation policies rarely reflect this reality. Take the case of Charlotte Danielson’s Framework for Teaching (FFT), one of the most widely used observation systems in the United States. The FFT is meant to be comprehensive in nature, capturing a complete picture of a teacher’s instructional quality. It is also meant to be content agnostic—that is, appropriate for use in a variety of subjects, grade levels, and instructional contexts. Importantly, the FFT also adheres to one specific vision of teaching quality. As Danielson (2013) describes, the FFT is based on an inquiry-oriented conception of teaching and posits that education is most powerful when students guide their own learning and teachers take on the role of facilitator. For teachers who work with students who may benefit more from more time engaged in teacher-directed instruction, such as special educators, their instruction may not be adequately captured by tools that prioritize student-centered teaching (Jones & Brownell, 2014; Morris-Mathews et al., 2020).

Hypothetically, the lack of alignment between teaching effectiveness and observation systems could have direct negative consequences for teachers, particularly in more high-stakes uses. But all evidence we have suggests that, in practice, teachers receive nearly uniformly high observation scores in evaluation (Kraft & Gilmour, 2017). If this is the case, some might question the extent to which we are concerned about misalignment? We argue there are a number of reasons that still warrant out attention. First, when a district adopts an observation system, it may signal the importance of some practices relative to others. High observation scores on a tool clearly aligned with student-centered instruction provide a signal to teachers about their performance and where

they might need to invest efforts to improve (or not). Such a tool would fail to provide teachers with incentives or appropriately targeted information for improving practices in ways aligned with the needs of students who benefit from more teacher-centered practices. Furthermore, the use of an observation system only aligned with a single perspective on teaching could potentially weaken the alignment of a district's professional learning system. As one example, we have increasingly seen districts adopt new instructional programs aligned with the *science of reading*, which often emphasize teacher-led instruction. The success of a district's curricular program or its evaluation policies are likely to be undermined when the two don't align. Although we should not expect any single observation tool to capture the full range of instructional approaches used in schools, research has yet to adequately address the trade-offs of using one tool versus another in evaluation contexts.

Purpose of Our Study

Our study investigates the consequences of using observation systems like Danielson's FFT in contexts where student-led instruction is less common. We explore the validity evidence that supports using the FFT for human capital and improvement purposes with special educators, specifically those providing instruction to high-incidence student populations in Grades 3 to 8.² Special educators make up approximately 12% of teachers in schools, and although this is not a majority of teachers, it is a robust minority and therefore worthy of consideration through a validity lens. It is worth noting that Danielson provides scoring supports for observing special education instruction; however, these supports are not designed for special education teachers serving students with high-incidence disabilities (e.g., specific learning disabilities), which is the largest group of SWDs in schools. Furthermore, most states have taken steps to refine their observation systems for special educators or provide supplementation guidance to support the evaluation of these teachers (Gilmour & Jones, 2020).

Drawing on 206 digitally recorded lessons from 51 special education teachers in Rhode

Island, we investigate FFT scores created by well-trained raters using standardized scoring procedures. We ask two research questions. First, to what degree does the FFT provide accurate and reliable estimates of teaching quality among special education teachers? Then, leveraging the same sample of lessons, we ask whether FFT scores align with scores on an observation system that explicitly captures teacher-directed instruction: the Quality of Classroom Instruction (QCI) instrument (Doabler et al., 2015). Our findings provide the first comprehensive empirical evidence of how a popular observation system functions for special education teaching.

Aligning Instruction With Students' Needs

Research from the past three decades provides evidence that instructional approaches are most effective when they are aligned to students' prior learning; when teachers adhere strictly to a single view of teaching, it is likely to have a detrimental impact on student learning and long-term outcomes. From this perspective, all students likely benefit from some level of teacher-directed instruction (Clements et al., 2013; Cohen, 2018; Stockard et al., 2018; Taylor, 2018), and simultaneously, no students should be denied opportunities to develop the higher-level comprehension skills that are often best supported through inquiry instruction. The issue is the balance between the two, and the positive contributions of teacher-directed instruction are most pronounced among students with lower levels of initial skills (Al Otaiba et al., 2011; Connor et al., 2018; Fuchs et al., 2009; Mercer et al., 1996). What explains these findings? In research on cognition, it is well established that the cognitive demands of learning are higher when students have limited prior knowledge; when novice learners encounter new problems, they have to dedicate more working memory to coding new information, while experts have the schema necessary to flexibly—and independently—adapt new information (Kirschner et al., 2006; Martin, 2016; Mayer, 2003; Sweller, 1988). In the context of teaching, this means that students with limited knowledge are more likely to benefit when teachers take a more active role in structuring learning opportunities, making instruction

explicit to aid in both recall and retention of key skills, routines, and strategies. In contrast, inquiry (or student-directed) instruction is more likely to lead to deep conceptual understanding when students already have the requisite prior knowledge. This idea is referred to in the literature as the *expertise reversal effect* (Kalyuga, 2007), or alternatively *aptitude–treatment interaction* (Snow, 1989), or more recently, *child–instruction interaction* (e.g., Al Otaiba et al., 2011; Connor et al., 2018). According to these theories, depending on specific skills or strategies, SWDs may need to spend more time engaged in explicit, systematic instruction to support their independent participation in inquiry-based instruction. This is not to suggest that student-centered approaches are not appropriate for SWDs, only that students typically need teachers to initially play a scaffolding role when they are engaging with new skills, strategies, or concepts.

Several studies have documented the relationship between student prior performance and the effectiveness of instructional approaches. In a series of studies, Connor and colleagues (2004, 2007, 2009, 2018) have demonstrated that elementary-grade students make greater growth in reading and mathematics when teachers individualize their instructional approaches based on students' needs. That is, when students exhibit low levels of skills in reading or math, they benefit most from more time in teacher-directed instruction, and students with higher levels of skill benefit from more student-centered instruction. Importantly, as students' skills improve over the course of the year, the recommended amount of student-centered instruction increases. Morgan et al. (2015) found similar patterns for first-grade students in mathematics, drawing on teachers' reports of their instructional practices in the *Early Childhood Longitudinal Study, Kindergarten Class of 2010–11*. Their findings suggested that students with mathematics difficulties benefited more from teacher-directed instruction, while students without such difficulties seemed to benefit equally from either teacher-directed or student-directed approaches. These patterns hold across several additional studies, with particularly strong differences in the early grades (Le et al., 2006; Sonnenschein et al., 2010; Xue & Meisels, 2004). Given these findings, we now turn to how instruction is conceptualized in the FFT and in the special education literature, respectively.

FFT's Conceptualization of Student-Centered Teaching. The FFT is a general observation system, developed to be used across grades and subject areas (Danielson, 1996, 2013). Danielson's FFT divides teaching into four domains: Planning and Preparation, Classroom Environment, Instruction, and Professional Responsibilities. This study focuses only on the Classroom Environment and Instruction domains (see the appendix).

Although the FFT is intended to be content and curriculum agnostic, the FFT is founded in an inquiry approach to teaching (Danielson, 2013). It posits that students guide their own learning, and that effective instruction requires teachers to play the role of facilitator. Teachers can promote students' construction of understanding by enacting strategies that deeply engage students in content, promote critical thinking and reasoning skills, and encourage students in intellectual argumentation (Danielson, 2013). In this view, teachers encourage student autonomy and allow students to direct their own inquiry and learning, as well as participate in the learning of peers. The FFT also assumes that advanced teaching is based upon the extent to which student learning is anchored to conceptual knowledge (Morris-Mathews et al., 2020). By that, we mean students engage with increasingly complex content and apply analysis to discern themes, make predictions, and engage in debate and open-ended questioning.

The student-led nature of effective instruction described in the FFT benefits students with stronger prior academic skills (e.g., Connor et al., 2009, 2018). The degree, however, to which such instruction, is effective for promoting conceptual understanding among students who demonstrate delays in their acquisition of foundational skills and cognitive and metacognitive strategies (i.e., SWDs and other students experiencing low achievement) is less established. In the next section, we compare this approach to current research in special education.

Conceptions of Effective Teaching in Special Education and Supporting Research. Scholars from special education assert that SWDs often need explicit, systematic instruction to promote their acquisition of new concepts, skills, or strategies (Jones & Brownell, 2014; Klingner et al., 2016; Morris-Mathews et al., 2020). This is not to suggest that SWDs cannot or should not be

engaged in student-directed instruction, only that they likely benefit from additional teacher actions to support their development of conceptual understanding (Fuchs et al., 2021). Results from intervention and observational studies of teaching in special education have demonstrated that SWDs learn lower and higher-order concepts and skills best when instruction is intensive, teacher-directed, explicit, and systematic—findings that also hold for students experiencing low achievement in reading, writing, and mathematics.

Explicit, systematic instruction (a) involves teacher modeling and controlled practice; (b) is highly interactive; (c) uses multiple, repeated practice opportunities and consistent feedback; and (d) incorporates ongoing progress monitoring to see if changes should be made to instruction (Doabler et al., 2015; Fuchs et al., 2021; Klingner et al., 2016; Vadasy et al., 2005; Vaughn et al., 2000). Furthermore, teachers apply these instructional processes systematically—progressing from less to more complex skills. Teachers also apply skills and strategies with controlled instructional materials that involve less cognitively demanding examples prior to introducing more demanding ones. Even when cognitive and meta-cognitive strategies and conceptual understanding are taught using peer-mediated instruction, they are modeled first by the teachers and well structured in terms of specific strategies or skills that are applied (e.g., Vaughn et al., 2011).

Although special education teachers use practices representative of “distinguished” practice on the FFT (e.g., questions and assessments are used regularly to diagnose evidence of learning by individual students), they are also likely to rely heavily on practices (e.g., explicit and systematic instruction) that are not well represented in the Instruction domain of the FFT (Morris-Mathews et al., 2020). In a systematic content analysis of the FFT, in which the authors analyzed all sentence units within the FFT’s Instruction domain, Morris-Mathews and colleagues found that the FFT largely excluded instructional practices documented as effective in special education, math, and reading (e.g., Doabler et al., 2015; Wanzek et al., 2013), and in observational studies of effective reading and special education teachers (Brownell et al., 2009; Connor et al., 2009; Johnson et al., 2021). They found that practices like explicit and systematic

instruction were not present in 93% of the sentence units. Furthermore, some aspects of effective special education instruction—such as providing multiple practice opportunities—were never mentioned in the tool. Thus, the FFT captures some practices used to support the learning of SWDs; however, it overlooks practices that are considered essential to intervention and remedial instruction.

Findings from one study that examined the use of the FFT for assessing special education teachers compared with a tool designed to assess use of explicit instruction supported the conclusions of Morris-Mathews and colleagues (2020). Johnson and colleagues (2020) found that special education teachers’ ratings on the FFT systematically differed from their ratings on Recognizing Effective Special Education (RESET), a tool designed specifically to assess explicit instruction. Special education teachers’ performance on RESET was systematically higher than their performance on the FFT; of 125 correlations between performance on RESET and the FFT, only 73 were significant and could be considered moderate to low. Furthermore, in a second study using RESET, Johnson and colleagues (2021) found, using many-facet Rasch measurement (MFRM), that special education teachers’ performance on the most difficult items defining explicit instruction accounted for 4.5% of the variance in the achievement of SWDs.

Collectively, existing research from intervention and observations studies raises questions about the FFT’s potential for validly capturing the type of instruction shown to improve the performance of SWDs and other students with low achievement who are typically served exclusively by general education teachers. Furthermore, if the FFT overlooks some of these critical practices, it may be of limited use for providing feedback to special education teachers. With these concerns in mind, we introduce a conceptual framework for how we might systematically examine the validity of using the FFT in the evaluation of special educators.

Conceptual Framework: Examining Validity Evidence From Measurement and Functional Perspectives

Given questions about the appropriateness of using the FFT to capture special education teaching practices, we follow M. T. Kane and Wools’ (2019)

framework for assessing the validity of classroom assessments.³ Specifically, we investigate the observation system's validity evidence from both a measurement perspective (i.e., the extent to which the assessment is accurate, reliable, and generalizable) and a functional one (i.e., the extent to which the assessment supports specific goals). M. T. Kane and Wools (2019) argue that effective assessments attend to both of these perspectives. In the case of observations, we would want to ensure that the scores they produce are defensible (for reasons of fairness) and they are able to achieve their purpose. Cronbach (1988) describes this as the distinction between *truthfulness* and *worth*. He emphasizes that although assessments may be "truthful" (i.e., they have adequate measurement characteristics), they may or may not be worthwhile. For example, if the FFT rank orders special education teachers accurately but cannot provide insight into how to improve their teaching, scores would be considered truthful but have little worth in improvement efforts. Furthermore, M. T. Kane and Wools (2019) argue that the measurement and functional perspectives are complementary, in that "the relative importance of the two perspectives in evaluating an assessment will depend on the goals and contexts of the assessment" (p. 12). Below, we describe how these two perspectives frame our investigation of the FFT's validity in the evaluation of special educators.

The Measurement Perspective

A measurement perspective has been used in other validation arguments of observation systems such as the *Classroom Assessment Scoring System* (CLASS), the *Mathematical Quality of Instruction*, and the *English Language Learner Classroom Observation Instrument* (Baker et al., 2006; Bell et al., 2012; Mantzicopoulos et al., 2018). M. T. Kane and Wools (2019) write,

The measurement perspective views assessments primarily as measurement instruments, and as a result, it focuses on certain technical criteria, particularly the generalizability (or reliability) of scores and their accuracy as estimates of the attribute of interest. It emphasizes standardization and objectivity. (p. 15)

We consider four aspects of validity generally associated with this measurement perspective:

the accuracy, consistency, bias, and generalizability of scores (Bell et al., 2012).

Accuracy. Scores should be accurate. Every observation system relies on experts who establish what each score point means. These experts, often called "master raters," decide what basic or proficient or exemplary teaching and learning looks like in behavioral terms. If raters do not apply ratings in the same ways as master raters, the meaning of the observation scores are undermined. This has historically been a serious problem in teacher evaluation, with most teachers receiving high scores on observation scales despite moderate levels of teaching quality (Kraft & Gilmour, 2017; Weisberg et al., 2009).

Consistency. Scores from the observation system should also be consistent. Raters should consistently agree on the quality of the interactions they observe. This is analyzed by considering the degree to which raters assign the same ratings when rating the same lesson, also known as inter-rater agreement. Another form of reliability concerns the degree to which the observation system's scores reflect variation in teaching, rather than variation in facets of the assessment system that are not relevant to the construct (Hill et al., 2012). These facets could be raters, lessons, or different classroom compositions. Generalizability studies frequently provide relevant evidence on this point (cf. Praetorius et al., 2014).

Bias. Observation scores can be compromised if they are biased. Two sources of potential bias that are prevalent in the teacher evaluation context are the application of the rating scales and the assignment of raters to lessons. Raters often have expertise with certain populations of students, content areas, and general education or special education classrooms. This expertise can bias scores in unpredictable but systematic ways. For example, administrators with special education backgrounds could potentially rate special educators more harshly because they have strong views of "good" special education teaching, resulting in the assignment of lower scores than would be assigned by a rater without special education expertise. If the teacher is only ever observed by an administrator with a special education background, this type of bias could potentially result

in a teacher receiving systematically lower scores than other teachers who are observed by administrators without a special education background, who may tend to assign higher scores.

Generalization. All observation scores are created from only a sample of observed lessons, not the universe of lessons we want to characterize. For teacher evaluation, we often generalize to a year of teaching in all of the classes the teacher teaches, but only have observations of two or three lessons in a single classroom of students. Generalization concerns the degree to which generalizing from the sample to the universe is appropriate.

The Functional Perspective

A second validity perspective—the functional perspective (Cronbach, 1988)—focuses on the degree to which an assessment tool supports achieving a specific purpose(s). In the case of the FFT, we are concerned with the degree to which scores support the evaluation and improvement of teaching. We are also interested in the degree to which scores are likely to minimize undesirable outcomes. In the current study, the FFT and QCI scores were not used to assign consequences to teachers; however, in the tradition of similar researcher-created scores (e.g., Hill & Grossman, 2013; T. J. Kane & Staiger, 2012; Semmelroth & Johnson, 2014), we analyze potential consequences in our discussion.

Aligned to the Community's Views. The worth of the FFT to the two stated purposes depends in part on how well it aligns with the special education community's views of teaching quality, particularly with regard to high-quality special education teaching and learning. Their views are especially important because they may not be well represented among administrators. A recent nationally representative survey of principals found that only 10% had special education degrees (Bettini et al., 2019), and many administrators have been unable to identify effective instructional practices for SWDs in studies based on surveys and interviews (e.g., Bays & Crockett, 2007; Steinbrecher et al., 2015; Stelitano et al., 2020). We account for the special education community's perspectives by examining

FFT scores vis-à-vis an observation system reflecting what special educators consider to be high-quality teaching.

Support Appropriate Consequences. To improve teaching quality, observation scores can be used to inform leadership, certification, and probationary decisions (the human capital management goal) and to guide professional conversations and professional learning opportunities (the improvement goal). In Rhode Island, where our research study is conducted, observation scores are combined with other measures of teaching quality (e.g., student growth). These combined measures are then used to assign overall teacher evaluation scores that are associated with specific consequences. Those consequences include being eligible for specific teacher leadership opportunities or being put on probationary status, which requires more observations and a professional learning plan (e.g., Rhode Island Department of Education, 2017). The observation scores are also used in formalized reflection and improvement conversations between administrators and teachers, as dictated by state policy.

The appropriateness and fairness of such consequences should be evaluated empirically and logically. If, for example, research-based practices used by special education teachers cause them to score lower on an observation tool, those teachers may be encouraged to use other practices that score higher on the FFT. Such an outcome would undermine the worth of the scores because they would encourage teachers to use practices not viewed as effective practice in the special education community.

Our study did not involve consequences for teachers; therefore, we focus our analysis (and discussion) on the degree to which FFT scores may support teaching improvement conversations.

To summarize, validating the use of the FFT to evaluate special educators in the current policy environment will require measurement and functional evidence, or, information supporting the system's truthfulness and worth. Our study seeks to investigate such evidence. Our first research aim—examining whether the scores produced on FFT are accurate and reliable—draws on the measurement perspective of validity. The second—examining whether FFT provides similar

information about teaching as the QCI—draws on the functional perspective.

Method

The current study was designed to validate the FFT for identifying effective special educators, as required in many states' teacher evaluation policies.

Teacher Sample

We recruited a sample of elementary and middle school special educators ($N = 51$) from Rhode Island.⁴ We secured data from the state for the purpose of comparing the sample teachers to the broader population of special education teachers in Grades 3 to 8 in Rhode Island schools. Approximately 51% of the teachers in our sample were elementary school teachers ($n = 26$); the remaining 49% taught middle school students ($n = 25$). With volunteer samples such as ours, there are always concerns that the sample varies in meaningful ways from the target population. Using the data available to us, the teachers appeared to have similar levels of experience as those across the state; our 51 teachers had 11.05 years of experience on average, compared with the statewide average of 11.53 years among special educators ($n = 11,960$). The teachers in our sample did have higher evaluation scores on the statewide evaluation system than the broader population (3.63 vs. 3.50 on a 4-point scale). The teachers in our sample did not appear to disproportionately represent schools with regard to urbanicity or student family poverty levels; similar patterns across these teacher characteristics were found when comparing educators in our sample to other teachers in the same district.

Four video lessons were collected from each teacher during the 2016–2017 school year. Participating teachers were offered the option of self-recording lessons or having a research assistant attend scheduled class periods to collect the video recordings.

Lesson Sampling

Lessons were sampled to mirror teachers' instructional responsibilities. As many scholars have suggested (Casabianca et al., 2015; Cash &

Pianta, 2014; Mashburn, Downer, et al., 2014; Meyer et al., 2011), care should be taken to map a sample of lessons onto the population of a teachers' lessons over the course of the year, being mindful, for example, to not oversample lessons from any one subject area or from two time points too close to one another. Therefore, lesson sampling followed teachers' workloads. For example, if a teacher spent close to 75% of their time in a coteaching setting and 25% of their time in a resource room setting, members of the research team conducted three coteaching lesson observations and one resource room observation. In cotaught lessons, raters were directed to only assess the instruction provided by the special education teacher. As an example, if the general education teacher was providing whole-class instruction and the special educator was working with a small number of students, the rater would only assess the special educator. On occasion, special educators in cotaught classes would not actively teach (while the general educator taught the whole class). However, these instances were infrequent and short in duration. Raters typically had sufficient information even with these lapses to still assign scores in 15-minute segments. Lessons were recorded across the school year, with no recordings scheduled during the early part of the school year (September–October), the end of the school year (late May–June), around holidays, or near testing windows to limit bias associated with those times of year. Table 1 summarizes the characteristics of lessons observed. The majority of teachers (84%) were observed in both English language arts (ELA)/reading and mathematics lessons. More teachers were observed in resource room or self-contained settings (92%) than in cotaught settings (37%), and only 18% of teachers were observed in both settings.

Measures

Once all recordings were complete, individual lessons were scored using two observation systems: the FFT and the QCI (Doabler et al., 2015).⁵ Data collected from these measures provided the basis for our validation. Of note, we paid close attention to scoring patterns between the FFT and the QCI, which closely reflects our definition of effective special education teaching.

TABLE 1
Characteristics of Lessons in the Sample

Total	%
Grade level	
Elementary school	52.4
Middle school	47.6
Subject area	
Reading/ELA	54.4
Math	44.7
Other	1.0
Setting type	
Resource room/self-contained	73.8
Cotaught	26.2

Note. $n = 206$.

FFT. The FFT is intended to reflect the range of instructional and other teaching responsibilities. Altogether, the FFT includes four domains: Planning and Preparation (Domain 1), Classroom Environment (Domain 2), Instruction (Domain 3), and Professional Responsibilities (Domain 4). Many states, including Rhode Island, use a variation of the FFT that includes only Domain 2: Classroom Environment and Domain 3: Instruction. Since our teacher sample was based in Rhode Island, they only received ratings across these two domains. Within each domain, raters provide scores ranging from unsatisfactory (1) to distinguished (4) on four *components*, which each represent instructional practices or characteristics. There are eight components overall across Domains 2 and 3.

In line with existing conventions for creating observation scores (Bell et al., 2018), FFT ratings were applied to 15-minute segments of a lesson. For example, a 60-minute lesson would consist of four scored segments. If there were 7.5 minutes or more remaining to be rated in the final interval, a new interval rating was created. If there were less than 7.5 minutes, those minutes were rated with the preceding segment. Throughout our findings, we report out lesson-level scores that were created by aggregating across segments. With the exception of analyses comparing lesson-level scores across raters, all analyses report lesson-level averages across the two raters who scored each video.

FFT Scoring Procedures. All 206 videos were double scored on the FFT, with observa-

tions spread across 12 raters. An Educational Testing Service (ETS) statistician generated rater assignments with the goal of maximizing, to the extent possible, unique combinations of teachers and raters such that raters did not observe the same teacher more than once. Random assignments were made to achieve a balanced distribution in terms of grade level, subject matter, and special education classroom settings (coteaching, resource room, and substantially separate). One rater removed herself from the project during Week 5 of scoring; her remaining videos were reassigned among the other 11 raters.

Among the 11 raters who participated through the entire scoring period, each rater scored 8% to 10% of the digitally recorded lessons. To ensure score quality following initial training, we used two scoring strategies that have been recommended in existing literature on observations (Bell et al., 2019). To improve rater reliability and to reduce the chances of rater drift, raters scored a weekly *calibration* exercise. This gave trainers an opportunity to assess rater scoring and intervene if necessary. Scoring also included a *validation* exercise, which was used to enable rater-trainers to monitor raters' scoring accuracy. Five prescored videos were embedded into each rater's assigned video queues, without being identified as validation videos. These data allowed the research team to assess how accurately raters were aligning with scores assigned by master raters.

QCI. In line with our validity framework, we sought a comparison observation system that would allow us to assess the FFT's function, or the extent to which it could lead to improvement in special educators' instruction. We used the QCI system. The QCI asks raters to assess instruction along eight key instructional principles, including modeling, transitions, pacing, timely checks for student understanding, student engagement, encouragement, and ensuring high rates of success for all students. Collectively, these constructs reflect the kind of explicit, systematic instruction that is widely endorsed in the special education literature. Scores are assigned at the end of each lesson on a scale of 1 (low quality) to 3 (high quality).

QCI Scoring Procedures. QCI scoring procedures mirrored FFT scoring procedures, with

a few exceptions. Only 20% of lessons were double-coded on QCI. We devoted additional resources to double scoring FFT lessons to support our generalizability studies, but double scoring of 20% is still in line with common scoring approaches in the field. In all, 206 videos were completed by six raters.⁶ Each rater was allotted an equal number of videos at the outset of coding assignments, and like with FFT, we took steps to balance the distribution of lesson types across raters.⁷ As in our FFT training, raters completed routine calibration exercises for the purposes of quality control and to guard against rater drift. As a secondary check, we also used validation videos, where we checked rater agreement with master raters.

Data Analysis Plan

To investigate the appropriateness of using the FFT in special education, a natural starting point was to examine how special educators scored on the FFT, on average, and the distribution of their scores within and across FFT components. To ensure that raters were able to score lessons (i.e., operationalize the scales on the instrument) in a way that was psychometrically sound, we examined how raters used the scoring scales. We examined the distribution of lesson ratings for each individual FFT component, as well as for the two domains, Classroom Environment and Instruction. In addition, we anticipated that the FFT might function differently across varied aspects of special educators' classroom context, including subject (reading vs. math) and setting (resource room/self-contained vs. coteaching). We hypothesized that special educators would be more likely to use explicit, systematic instruction in resource and self-contained settings, where they have more control over the instructional design. It was not as readily obvious whether we would be more likely to see instruction that is aligned with the FFT in either reading or math lessons. By upper elementary grades and middle school (the grade levels in our sample), reading instruction is likely to shift away from foundational skills—such as phonics and phonemic awareness—to comprehension, but at the same time, SWDs are more likely than their peers to continue receiving targeted instruction in these areas. In

mathematics, instruction for SWDs commonly prioritizes explicit, systematic instruction.

One important limitation in analyses comparing FFT scores across instructional contexts is the uneven representation of instructional contexts within teachers. We would like to be able to assess whether the same teacher would score differently in a cotaught or self-contained setting, but only nine of 51 teachers were observed in both settings. We could alternatively run comparisons of cotaught and self-contained lessons across the sample as a whole, but it would be difficult to make claims about whether differences in scores across settings were attributable to the settings themselves or underlying differences across teachers. Ultimately, given these concerns, we elected to exclude these comparisons from our results; however, it is worth noting that across the restricted and full samples, differences across settings were minimal and nonsignificant. The same concerns were not present with the ELA and math comparisons, with 43 of 51 teachers having lessons across both subjects. In the Findings section, we present subject area comparisons for both the FFT and the QCI for the restricted sample.

Next, in line with the measurement perspective of validity, we examined whether the FFT factor structure conformed to theory—in this case, a two-factor structure in line with the FFT's two domains. We assessed the factor structure of FFT scores by using multilevel multivariate factor analysis. Recent studies (McCaffrey et al., 2015; Oliveri et al., 2017) demonstrate that traditional factor analyses may not be appropriate for observation data. This is because raters' scores and related errors are hierarchical (scores are recorded in 15-minute segments, aggregated to the lesson, then aggregated across raters) and because any rater-specific errors are likely to be correlated across dimensions. Therefore, we employed a two-level factor analysis. In line with McCaffrey et al. (2015), our first step was to estimate and remove rater-specific effects from segment-level scores. Then, aggregating to the lesson, we ran confirmatory factor analyses (CFAs), keeping the within-teacher factor structure constant at two factors and modeling either one or two between-teacher factors to assess the FFT's theorized two-domain structure.

From the measurement perspective of validity, it was also important to collect data on FFT scores' accuracy, consistency, bias, and generalizability. To assess *accuracy* of rater scores, we examined the extent to which FFT raters aligned with master raters. To assess *consistency* of rater scores, we examined the degree to which FFT raters agreed on the 4-point FFT scale. Interrater reliability for QCI scores was also calculated using regular kappa and weighted kappa statistics with both linear and quadratic off-diagonal weighting. We then conducted a series of *generalizability* studies (G study) and *decision* studies (D study) (Brennan, 2001; Cronbach et al., 1972; Shavelson & Webb, 1991) for the FFT. Generalizability studies provide insight into the proportion of variation in scores that can be attributed to aspects of the observation system—for example, raters, the lessons observed, measurement error. For the G studies, we estimated the variance of total FFT scores that were attributable to variability across teachers, lessons, segments, raters, and residual error, and interactions between each of these sources of variation with the lesson. We did not include interactions between teachers and raters because our rater assignment intentionally avoided raters observing the same teacher more than once. With a sample size of 51 teachers, we could not analyze the variation that was attributable to specific aspects of teacher assignments (e.g., instructional level) or lesson type (e.g., coteaching vs. resource).

D studies were then carried out to determine the practical implications of the sources of variation in scores. The D studies were carried out by using the variance estimates from the G study to develop insights regarding the optimal approach to observations, given a specific desirable level of reliability of the observation scores. School administrators often need to make choices about how many lessons are observed and how many observers rate a lesson. Therefore, the D studies estimated how reliability changed under three different scoring designs: (a) one rater, one lesson; (b) two raters, one lesson; and (c) two raters, two lessons. Reliability is calculated as a dependability coefficient. For examples of this application of generalizability theory to different observation tools, see Hill and colleagues (2012) work on a mathematics specific tool or Mashburn

and colleagues work on a general observation system (Mashburn, Downer, et al., 2014).

Then, to empirically assess the degree to which FFT scores are likely to be able to support their intended uses (in other words, gathering validity evidence from the functional perspective), we included two additional analyses based on QCI data. To understand whether FFT scores might systematically sort teachers in ways that do not reflect effective special education teaching practice, we compared teachers' rank ordering on the FFT with their ordering on the QCI. To investigate how well the FFT might support the improvement goal, we considered how the FFT score distribution compared with the QCI score distributions, noting the teaching practices that do not score similarly. By expressly specifying our inferences and presenting our evidence along these claims, we can inform the validity argument (M. T. Kane, 2006) for FFT scores being used for their intended purposes.

Findings

Research Question 1: To What Degree Does the FFT Provide Accurate and Reliable Estimates of Teaching Quality Among Special Education Teachers?

Mean and Distribution of FFT and QCI Scores. A first strategy for assessing the appropriateness of the FFT for measuring special education teaching quality is to examine the distribution of scores across the sample. If implemented appropriately, all four categories of instructional quality, from *unsatisfactory* to *distinguished*, should be represented in a sample of lesson scores. As outlined in Table 2, which presents mean scores for lessons, the Classroom Environment components generally performed as expected. Across the 201 lessons, the mean scores ranged from 2.26 on *Culture* to 2.94 on *Space*, corresponding to a range of developing to effective. No lessons scored at the distinguished level (4 on the 4-point scale) on any of the components. Average scores for the Instruction domain, in contrast, were consistently low across all components, with average scores ranging from 1.82 to 2.29. Two components had average scores less than 2.00 (*Questions* and *Assessment*) putting the sample average between unsatisfactory and developing. For each of these two components, teachers scoring at the 90th

TABLE 2

Means and Standard Deviations of FFT and QCI Scores

Dimension	<i>N</i>	<i>M</i>	<i>SD</i>
FFT	206	2.40	0.27
FFT Domain 2: Environment	206	2.73	0.24
FFT 2a: Respect	206	2.84	0.32
FFT 2b: Culture	206	2.26	0.39
FFT 2c: Procedures	206	2.79	0.35
FFT 2d: Behavior	206	2.85	0.29
FFT 2e: Space	206	2.94	0.17
FFT Domain 3: Instruction	206	2.07	0.34
FFT 3a: Communication	206	2.29	0.39
FFT 3b: Questioning	206	1.82	0.45
FFT 3c: Engagement	206	2.14	0.40
FFT 3d: Assessment	206	1.94	0.44
FFT 3e: Responsiveness	206	2.16	0.35
QCI	237	2.26	0.57
QCI 1: Modeling	237	2.23	0.70
QCI 2: Checking for Understanding	237	2.33	0.64
QCI 3: Appropriate Response Time	235	2.25	0.75
QCI 4: Student Engagement	237	2.16	0.68
QCI 5: Ensuring High Rates of Success	237	2.20	0.72
QCI 6: Encouraging Effort from All Students	237	2.30	0.70
QCI 7: Efficient Transitions	219	2.19	0.74
QCI 8: Good Pacing	237	2.18	0.70

Note. All lessons are double scored in 15-minute segments and aggregated across segments and across raters. Complete FFT component names are included in the appendix. FFT = Framework for Teaching; QCI = Quality of Classroom Instruction.

percentile in the sample still averaged only 2.50. As summarized in Figure 1, while Classroom Environment components appear to reflect variation across the scoring scale, the Instruction component averages largely hover at the low end of the scoring scale.

It could be the case that the instrument is working as intended, and the special educators in our sample are simply less proficient than desired at the Instruction components. However, coupled with the questions raised about the appropriateness of the FFT for evaluating special education teaching, it may also be that some FFT components are not able to capture the range and quality of instructional strategies valued in special education. This possibility is supported by differences in the QCI distribution relative to the FFT distribution. In contrast to the consistently low scores on the FFT's Instruction domain, the QCI scores were higher overall. As reported in Table 2, the means on individual dimensions ranged

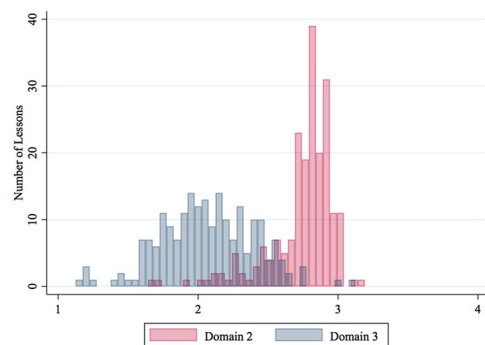


FIGURE 1. *Frequency of assigned lesson-level FFT scores on Domains 2 (Classroom Environment) and 3 (Instruction).*

Note. FFT = Framework for Teaching.

from 2.16 (Principle 4: Student Engagement) to 2.33 (Principal 3: Checking for Understanding), all scored on a 1 to 3 scale. Also, the QCI appears to pick up on variation in lesson scores that was

TABLE 3

FFT Means in Reading and Math Lessons

Domain	Reading		Math	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Domain 2	2.75	0.21	2.69	0.23
FFT 2a: Respect	2.86	0.28	2.80	0.31
FFT 2b: Culture	2.28	0.34	2.20	0.33
FFT 2c: Procedures	2.78	0.30	2.76	0.33
FFT 2d: Behavior	2.87	0.22	2.81	0.29
FFT 2e: Space	2.95	0.10	2.91	0.15
Domain 3	2.08	0.30	2.03	0.29
FFT 3a: Communication	2.32	0.35	2.22	0.33
FFT 3b: Questioning	1.85	0.38	1.76	0.37
FFT 3c: Engagement	2.14	0.34	2.11	0.35
FFT 3d: Assessment	1.90	0.36	1.96	0.35
FFT 3e: Responsiveness	2.19	0.27	2.09	0.30
	<i>N</i> = 43		<i>N</i> = 43	

Note. FFT scores range from 1 (poor) to 4 (distinguished). FFT = Framework for Teaching.

not captured in the FFT, as evidenced by the broader range of scores assigned to QCI components in comparison to FFT components.

Variation in FFT and QCI Scores by Classroom Context. In Tables 3 and 4, we present results comparing mean scores in reading and math instruction for the 43 teachers for whom we had lessons in each subject. Means were calculated by first generating person-specific means for both subject areas and then aggregating across the sample. No differences across the two groups of lessons were significant for any component on either the FFT or the QCI. Interestingly, the general pattern of differences on the FFT was reverse to that for the QCI. Reading lessons scored slightly higher than math lessons on the FFT but lower than math lessons on the QCI. As a reminder, given the small number of teachers who had coteaching and resource room lessons, we chose to not include results comparing scores across settings. But, like with the reading and math comparisons, the classroom setting comparisons were minimal.

Accuracy and Reliability of FFT and QCI Scores. One way to assess the FFT from a measurement perspective is to test whether raters accurately and consistently apply scoring rules

and do so in ways that are bias-free. As described in the Method section, at the conclusion of training, the raters in our sample agreed exactly with master raters 74.0% of the time and were exact or adjacent 98.8% of the time, both of which are in line with existing conventions for score accuracy. To assess one source of potential rater bias, we looked at the extent to which raters scored systematically higher or lower than master raters, finding that they scored higher than master raters 14.6% of the time and lower than master raters 11.2% of the time. These results are not significant enough to warrant concern about this source of rater bias.

In operational scoring, all lessons were double scored on the FFT to assess interrater agreement. In Table 5, we summarize exact agreement across raters on all components and Cohen's kappa to better account for chance agreement (a necessary step on scales with a small number of possible scores). Exact agreement on Domain 2: Classroom Environment averaged 75.4% across all components, and agreement rates ranged from 50.6% on *Culture* to 86.3% on *Space*. For Domain 3: Instruction, exact agreement was consistently lower, ranging from 47.0% on *Assessment* to 56.5% on *Responsiveness*. Although these results raise questions about whether highly trained raters can consistently

TABLE 4

QCI Means in Reading and Math Lessons

Principle	Reading		Math	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
QCI 1: Modeling	2.34	0.67	2.28	0.62
QCI 2: Checking for Understanding	2.34	0.56	2.40	0.57
QCI 3: Response Time	2.33	0.70	2.40	0.67
QCI 4: Student Engagement	2.14	0.62	2.20	0.65
QCI 5: High Rate of Success	2.24	0.66	2.27	0.68
QCI 6: Encouraging Effort	2.36	0.60	2.36	0.68
QCI 7: Transitions	2.12	0.64	2.23	0.74
QCI 8: Good Pacing	2.19	0.65	2.25	0.67
QCI Overall	2.26	0.53	2.30	0.51
	<i>N</i> = 43		<i>N</i> = 43	

Note. QCI scores range from 1 (low) to 3 (high). QCI = Quality of Classroom Instruction.

TABLE 5

Reliability Estimates for the FFT

Dimension	Percent exact agreement	Percent exact or adjacent agreement	Simple kappa	Linear weighted kappa	Quadratic weighted kappa
FFT 2a: Respect	79%	99%	0.34	0.39	0.45
FFT 2b: Culture	51%	97%	0.07	0.09	0.14
FFT 2c: Procedures	77%	98%	0.32	0.36	0.42
FFT 2d: Behavior	85%	99%	0.42	0.43	0.45
FFT 2e: Space	86%	99%	0.05	0.06	0.06
FFT 3a: Communication	49%	98%	0.04	0.08	0.15
FFT 3b: Questioning	52%	96%	0.15	0.20	0.26
FFT 3c: Engagement	55%	98%	0.09	0.14	0.22
FFT 3d: Assessment	47%	97%	0.04	0.11	0.21
FFT 3e: Responsiveness	56%	97%	0.08	0.10	0.13

Note. Double-scored segment sample size = 526. FFT = Framework for Teaching.

apply scoring rules to the FFT's Instruction domain, we see similar patterns of interrater agreement on the QCI. As summarized in Table 6, exact agreement on QCI components ranged from 41% to 57%. We examine the implications of these lower-than-ideal agreement rates in the Discussion.

Factor Structure of FFT. A second step in assessing the FFT from a measurement perspective is to examine whether its factor structure conforms to theory. We would expect to see that the system's factor structure—five

components aligned with Classroom Environment and five with Instruction—holds when applied to our sample of special education lessons. Table 7 presents fit statistics for multi-level CFAs, run for one-factor and two-factor models, for the FFT and QCI, respectively. As a reminder, these models of lesson-level scores are adjusted to account for individual rater effects, and we focus our attention on the between-teacher factor structure, holding the within-teacher factor structure constant at 2. We see that the fit indices are virtually identical across the one-factor and two-factor models.

TABLE 6

Reliability Estimates for the QCI

Principle	Percent exact agreement	Percent exact or adjacent agreement	Simple kappa	Linear weighted kappa	Quadratic weighted kappa
QCI 1: Modeling	53%	90%	0.39	0.34	0.17
QCI 2: Checking for Understanding	41%	92%	0.23	0.25	0.27
QCI 3: Response Time	44%	90%	0.11	0.14	0.18
QCI 4: Student Engagement	56%	93%	0.30	0.33	0.37
QCI 5: High Rate of Success	47%	98%	0.19	0.25	0.31
QCI 6: Encouraging Effort	46%	95%	0.1	0.19	0.28
QCI 7: Transitions	54%	92%	0.35	0.41	0.47
QCI 8: Good Pacing	47%	89%	0.24	0.32	0.42

Note. Double-scored segment sample size = 59. QCI = Quality of Classroom Instruction.

TABLE 7

Fit Statistics for Multilevel CFA Models Fit to FFT Lesson Means

Model	CFI	TLI	RMSEA	SRMR _{BETWEEN}	Chi-square stat	df	AIC	BIC
1-Factor	0.89	0.86	0.10	0.16	283.83	69	488.34	680.68
2-Factor	0.89	0.86	0.10	0.14	276.32	68	489.96	686.08

Note. Both models were run with two within-teacher factors. Acceptable model fit is indicated by values of CFI and TLI > .90; RMSEA and SRMR < .05 (Hu & Bentler, 1999). CFA = confirmatory factor analyses; FFT = Framework for Teaching; CFI = comparative fit index; TLI = Tucker–Lewis index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual; AIC = Akaike information criterion; BIC = Bayesian information criterion.

Both models have comparative fit index (CFI) and Tucker–Lewis index (TLI) of 0.89 and 0.86, respectively, with both failing to meet the >0.90 threshold of acceptable model fit. The Akaike information criterion (AIC) and Bayesian information criterion (BIC) are both slightly smaller for the one-factor model. Because the fit tests failed to distinguish between the two models, we ran exploratory factor analysis (EFA) models for both one and two factors. Figure 2 shows the eigenvalues of the correlation matrix for the lesson-level scores on each of the 10 FFT components. Although two of the eigenvalues are greater than 1, the first is by far the highest. Furthermore, follow-up analyses of the one-factor model show that all loadings are sufficiently high, ranging from 0.77 to 1.00. In addition, the fit statistics are best for the one-factor model.

Generalizability of FFT Scores. We conducted a series of G studies to understand teacher, lesson, and rater variability; and residual error; and

interactions between each of these sources of variation (Brennan, 2001; Cronbach et al., 1972; Shavelson & Webb, 1991). The rater variance can be thought of as the proportion of the variance that is attributable to raters. Rater-by-lesson variance can be thought of as the degree to which two raters are likely to rate a specific lesson similarly after accounting for the raters' severity. We estimated the variance components using standard random effect analysis of variance methods (Searle et al., 2009), with random effects for teacher, lesson, rater, rater by lesson, and the residual. We did not calculate rater-by-teacher variation because we purposely distributed lessons to minimize the potential bias associated with the same rater observing the same teacher more than once. Results from our G studies are summarized in Table 8.

Our G study findings suggest that teacher-specific variation is relatively low, ranging from 2% to 15% on individual components. At the domain level, the teacher-specific variation is still low—16% for Domain 2: Classroom

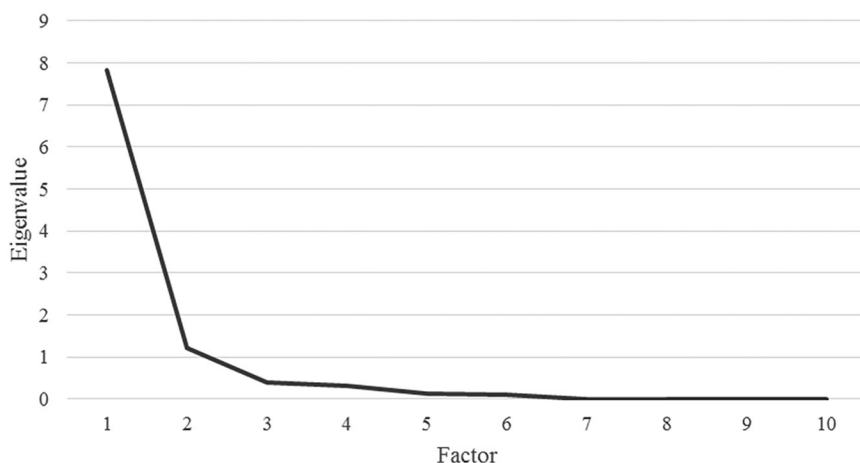


FIGURE 2. *FFT multilevel CFA between-teacher eigenvalues.*
 Note. FFT = Framework for Teaching; CFA = confirmatory factor analyses.

Environment and 13% for Domain 3: Instruction. This suggests that there was not a lot of variation captured between teachers. Lessons also vary. Lesson to lesson variation was estimated to account for 3% to 19% of the variation in scores, depending on the component or domain measured. We see the highest proportions of variance attributable to the rater by lesson component. This means that when similar raters (e.g., similarly strict, similarly lenient) observe the same lesson, they rate it differently. We also see moderately sized variation associated with the residual (variance that is not explained by teachers, raters, or lessons). When interpreting these findings related to lesson-specific variance, it is worth remembering that this study deliberately sampled lessons that mirrored teachers' teaching assignments—if the teacher taught half of her time in a coteaching situation, the lessons captured and rated for her included roughly half coteaching lessons. It is possible that the somewhat higher levels of lesson and rater by lesson variation we document here are due to the ways in which lessons were sampled across types of teaching arrangement. It is also possible that this is due to the ways in which FFT may be easier and harder to rate, depending on the structure of the lesson. And last, this could be because of the specific variance decomposition we were able to investigate. We are unable to sort out these

hypotheses with the current data because this is the first published study we are aware of to carry out this specific variance decomposition on only special education teachers while also using this sampling approach for teaching settings. To be conservative, this finding suggests administrators should carefully consider what lessons are being observed and the fairness associated with specific observation schemes across teachers.

It is important to contextualize these findings. Many others G studies have documented high levels of variation in these same sources, depending on the observation system used and aspect of teaching measured (Mantzicopoulos et al., 2018; Praetorius et al., 2014). We were unable to find an FFT G study that has the exact same design as this one;⁸ two had quite different sample sizes, and none were in special education. Therefore, we contextualize the findings with the caveat that the existing G studies may provide somewhat limited applicability to the current study. In a much larger U.S. study of 1,333 elementary and secondary math, ELA, and science general education classrooms, researchers found that across FFT components, variation at the teacher level ranged from 15% to 30%, variation attributable to raters was 3% to 9%, and lesson variation ranged from 0% to 12% (T. J. Kane & Staiger, 2012). In a smaller Nordic study of 10 kindergarten general education teachers, the G study was

TABLE 8

Generalizability Study Results

FFT component	2A	2B	2C	2D	2E	3A	3B	3C	3D	3E	Domain 2	Domain 3
Percent variance by source												
Teacher	19%	14%	13%	16%	2%	15%	14%	14%	10%	11%	16%	13%
Lesson × Teacher	14%	3%	14%	18%	10%	5%	8%	6%	6%	3%	19%	10%
Observer	2%	16%	2%	1%	5%	13%	10%	7%	17%	11%	6%	8%
Observer × Lesson	19%	31%	26%	22%	40%	29%	27%	30%	30%	30%	31%	39%
Segment × Lesson	11%	3%	13%	12%	0%	4%	7%	4%	7%	1%	8%	7%
Residual	35%	32%	33%	30%	44%	34%	34%	39%	30%	44%	21%	22%
Implied reliability ^a across three designs												
1 lesson, 1 observer	.20	.10	.12	.16	.02	.14	.10	.12	.07	.12	.18	.16
1 lesson, 2 observers	.26	.18	.16	.21	.04	.22	.16	.20	.12	.20	.25	.24
2 lessons, 2 observers	.38	.29	.25	.32	.07	.35	.26	.32	.20	.32	.38	.36

Note. FFT = Framework for Teaching.

^aThis is a generalizability coefficient.

carried out at a higher level of aggregation (domains, not components). In that study, variation attributable to teachers ranged from 16% to 36% on the Classroom Environment and Instruction domains, variation attributable to raters was 0% to 2%, and lesson variation ranged from 25% to 33%. Given these two studies, the findings here have both similarities and differences. Variation attributable to teachers is lower but overlaps with the other two studies. Variation attributable to raters is similar to the other studies; however, there are higher levels of rater variability for some components. And variation attributable to lessons is similar or slightly higher than one study and lower than the other study.

Policy makers often have to make choices about how to allocate scarce resources such as administrators' time. To examine the reliability trade-offs under different scoring designs that might be used in a teacher evaluation system, we also conducted a D study with three possible scoring designs: one lesson observed by one observer, one lesson observed by two observers, and two lessons observed by two observers. These results are also summarized in Table 8.

The D study provides information on the gains to reliability associated with increasing the number of lessons and raters. Others have documented that one FFT observation with a single observer is an unstable estimate of teaching quality and therefore inappropriate for high-stakes use (cf. Ho & Kane, 2013; Mantzicopoulos et al.,

2018). Our findings are consistent with this with implied reliabilities of one lesson with one or two raters ranging from depending on the component being rated. Relative to a single lesson observed by a single rater, adding an additional rater and adding an additional lesson both increased the stability of the D coefficients. However, adding a second lesson improved the reliability to a greater extent than adding a second rater, although reliability remains low.

Research Question 2: To What Degree Does the FFT Provide Similar Information About Teaching as Compared With an Observation System Specifically Designed to Measure Teaching Quality in Special Education Classrooms?

In the introduction, we raised the question of whether the differences in how special education and general education conceptualize effective teaching may impact how the FFT functions with special educators. To test this, we compared FFT scores with a second source of information on special education teachers: scores taken from the QCI observation system, which is believed to more closely reflect the kinds of instructional practices valued by the special education community. We summarize correlations between FFT scores and QCI scores in Table 9. In general, there are some positive, small correlations between the two measures. At the FFT domain

TABLE 9

Pearson Correlations Between FFT Scores and QCI Scores

Domain	QCI1	QCI2	QCI3	QCI4	QCI5	QCI6	QCI7	QCI8	QCI mean
FFT2	0.32**	0.28**	0.36**	0.29**	0.37**	0.30**	0.32**	0.36**	0.42**
FFT2a	0.30**	0.27**	0.27**	0.27**	0.31**	0.28**	0.33**	0.36**	0.39**
FFT2b	0.24**	0.23**	0.27**	0.17*	0.31**	0.23**	0.23**	0.21**	0.29**
FFT2c	0.23**	0.21**	0.31**	0.22**	0.28**	0.22**	0.19**	0.34**	0.32**
FFT2d	0.25**	0.22**	0.24**	0.27**	0.28**	0.22**	0.29**	0.31**	0.33**
FFT2e	0.23**	0.14*	0.33**	0.22**	0.23**	0.16*	0.23**	0.18*	0.28**
FFT3	0.34**	0.30**	0.33**	0.29**	0.38**	0.31**	0.27**	0.24**	0.39**
FFT3a	0.29**	0.27**	0.31**	0.24**	0.31**	0.32**	0.15*	0.21**	0.32**
FFT3b	0.25**	0.22**	0.18**	0.23**	0.27**	0.20**	0.18**	0.13	0.27**
FFT3c	0.32**	0.27**	0.31**	0.26**	0.38**	0.27**	0.28**	0.26**	0.38**
FFT3d	0.27**	0.24**	0.30**	0.23**	0.29**	0.24**	0.28**	0.19**	0.33**
FFT3e	0.28**	0.26**	0.29**	0.26**	0.37**	0.27**	0.22**	0.21**	0.34**
Mean	0.36**	0.32**	0.37**	0.31**	0.41**	0.33**	0.31**	0.31**	0.43**

Note. QCI = Quality of Classroom Instruction; FFT = Framework for Teaching.

* $p < .10$. ** $p < .05$.

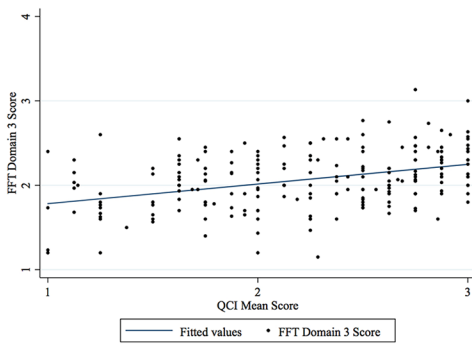


FIGURE 3. Comparing lesson-level means on FFT Domain 3 and the QCI.

Note. FFT = Framework for Teaching; QCI = Quality of Classroom Instruction.

level, Classroom Environment is associated with the QCI overall mean score at 0.42, and Instruction is associated with the QCI mean at 0.39. The correlations between individual components on the FFT and individual components on the QCI range from 0.13 to 0.38. As expected, the lowest correlations are between FFT Instruction components and the QCI principles. A more interesting comparison is when we plot the distribution of lessons as rated by the FFT and the QCI (see Figure 3). The two instruments tend to sort lessons similarly, with higher scores on one instrument generally corresponding with

higher scores on the other. It is difficult to find many cases where a lesson is on the high end of the QCI distribution but the low end of the FFT distribution or vice versa. However, when we look at the *actual* scores on the FFT and the QCI, the results indicate a ceiling on FFT scores between basic and proficient. Regardless of how strong teachers' lessons looked on the QCI, they were unlikely to score high on the FFT. Virtually all of the lessons score < 2.49 overall on the FFT's Instruction domain, despite considerable variation on the QCI.

Another way of demonstrating these patterns is to look at teacher-level scores on the FFT and QCI, with teacher scores organized by quintile (see Table 10). Rank ordering teachers in this summative way more closely approximates how observation scores would be used in evaluative contexts. Here also, we see that a teacher's rank ordering on the FFT corresponds fairly well to their quintile on the QCI; most teachers belong to a QCI quintile that is the same or adjacent to their quintile on the FFT. The FFT seems to capture the ranking of special education teachers relative to one another, but given the truncated range in FFT scores across our sample, the actual distance between teachers' scores on the FFT is quite small. As we argue in the Discussion, these findings would seem to have the greatest ramifications when the FFT is used for improvement

TABLE 10

Teacher Quintile Rankings, FFT and QCI

Principle		QCI mean quintile					Total (%)
		1 (%)	2 (%)	3 (%)	4 (%)	5 (%)	
FFT Domain 3 mean quintile	1	12.3	3.5	1.8	1.8	0.0	19.3
	2	3.5	8.8	8.8	0.0	0.0	21.1
	3	0.0	3.5	5.3	5.3	7.0	21.1
	4	1.8	3.5	1.8	5.3	5.3	17.5
	5	1.8	1.8	1.8	8.8	7.0	21.1
Total		19.3	21.1	19.3	21.1	19.3	100.0

Note. FFT = Framework for Teaching; QCI = Quality of Classroom Instruction.

purposes. Because the FFT does not seem to reflect the differences across lessons and teachers based on the use of high-quality special education teaching practices, the observation system is likely limited in the extent to which it can provide appropriate feedback.

Discussion

This study takes up a question that has largely been unanswered in the literature on the use of classroom observations mandated by teacher evaluation policies: How well do these tools reflect the reality that *instructional quality* depends on individual students' needs? A growing body of evidence suggests that students respond differently to instruction depending on their prior skills. Although all students should be presented with opportunities to develop higher-level skills (and inquiry-based approaches may be most appropriate for fostering these skills), students with weaker initial skill—including many SWDs—are more likely to need additional time engaged in explicit (or teacher-directed) instruction (Connor & Morrison, 2016; Morgan et al., 2015; Sonnenschein et al., 2010; Xue & Meisels, 2004). Many of the observation systems commonly used by districts to assess teaching performance fail to reflect this distinction. We explore this tension by focusing on the most popular observation system in the United States—the FFT (which privileges student-centered instruction)—when applied to special education teachers, who work with students who are more likely to benefit from explicit, teacher-directed instruction. We investigate the trade-offs of this choice.

We have organized our investigation around building two complementary forms of validity evidence (M. T. Kane & Wools, 2019). First, from a measurement perspective, we need confidence in the FFT's technical properties. We need to know if scores are defensible to ensure stakeholders' trust in the evaluation and improvement processes. Second, from a functional perspective, we need to know if the FFT provides useful information for achieving two specific outcomes: distinguishing between effective and ineffective teachers, and helping teachers improve. This latter goal, surrounding whether observation scores can guide professional learning opportunities, will be a particularly important lever for improving teaching quality in special education, where students continue to lag behind their nondisabled peers in math and reading achievement (e.g., Chudowsky et al., 2009; Schulte et al., 2016; Schulte & Stevens, 2015). Teachers of SWDs simply cannot afford to go down improvement paths that do not align with their students' needs, as could be the case if observation systems incorrectly diagnose the areas where they need to improve.

Of our results related to the measurement perspective, our most striking finding relates to mean lesson-level scores in the FFT's Instruction domain. Scores on individual components of Domain 3 were almost universally low compared with scores assigned to general education teachers in other studies; some components had means between Ineffective (1) and Developing (2) and almost all lessons were rated below the Proficient (3) level. If we were to rely on these scores alone, we would conclude that special educators were

performing far below expectations of teacher effectiveness. We had anticipated that we might see differences by teachers' instructional context, but we saw no significant differences across reading and mathematics instruction within the same teacher, and we had insufficient teachers with lessons in cotaught and resource room (or self-contained) settings to facilitate comparisons on special education service delivery models.

We collected several other sources of evidence surrounding the FFT's measurement properties, including the accuracy and reliability of rater scores, rater bias, the FFT's factor structure, and the generalizability of FFT scores. Across each of these categories, we saw that the measurement properties of the FFT when applied to our special education sample were in most cases similar to previous studies assessing the FFT's validity in general education (e.g., T. J. Kane & Staiger, 2012; Liu et al., 2019). Taken at face value, our levels of interrater reliability ranged from ~45% to 55% on FFT Domain 3 components and were lower than desirable. But they also align with reliability estimates from previous studies using the FFT that employed similarly rigorous training regimes as we did. For example, in the Measures of Effective Teaching (MET) study, FFT exact rater agreement for Domain 3 was 55% and 60% for Domain 2 (Bell et al., 2014). But, these scores also predicted differences in student achievement gains and were correlated to student surveys (Ferguson & Danielson, 2014; T. J. Kane & Staiger, 2012).

Also similar to other studies (e.g., Liu et al., 2019), factor analyses supported two factors: one for the learning environment domain and one for the instructional domain. Exploratory factor analyses, however, supported a single-factor model like we have seen in previous research uses of the FFT (e.g., Liu et al., 2019). Results of our G study demonstrated both similarities and differences with previous studies that applied the FFT to general education teachers (T. J. Kane & Staiger, 2012; Mantzicopoulos et al., 2018); however, there was no clear pattern across the teacher, lesson, and rater sources of variation. This type of variation across studies is expected in part because studies rarely parse variation in the exact same way and study populations can vary dramatically, thus reflecting more about the classrooms being studied than the instruments'

function. Looking only within our study, there was high variation attributable to the residual and to the Rater \times Lesson facets. In other words, the variation did not appear to be between raters generally but as an interaction with the individual lesson, where two observers were less likely to agree on what they saw.

But what about M. T. Kane and Wools' (2019) functional perspective? Cronbach (1988) refers to this as an assessment's worth. In our case, it was important to understand whether FFT scores reflected the special education community's conception of good teaching *and* whether scores would likely lead to improvements in teachers' instruction. Here, the evidence raises serious concerns. Our findings suggested that the FFT and QCI were only modestly correlated; correlations between the FFT's Instruction domain and the QCI indicators ranged from 0.24 to 0.38. Lessons where teachers scored high on the QCI were also ones more highly rated on the FFT's two domains. And, when examining teachers' performance relative to others in the distribution, both tools sorted teachers in somewhat similar ways. However, these correlations are not nearly as strong as we have seen in previous studies employing multiple observation systems. The MET study, for example, examined scores on five different observation systems (see T. J. Kane & Staiger, 2012, for more information). The FFT and CLASS (both general observation systems) were correlated at 0.88, and correlations between the FFT and the other systems were similarly high. We were more likely to see scores normally distributed on the QCI, whereas scores on the FFT were clustered on the low end of the scoring scale. These findings suggest that even though the FFT appeared to do reasonably well at sorting teaching quality within a special education sample, it did not distinguish between when teachers did and did not use effective special education practices.

Both systems captured some signal about the teaching quality in special educators' classrooms, but the systems varied in the ways that scores were arrayed and created this signal. These findings support evidence from a content study of the FFT's Instruction domain that showed that FFT scoring criteria and descriptions do not fully capture effective special education instruction (Morris-Mathews et al., 2020).

Limitations

One limitation of this study involved our sample of teachers and the sample of lessons within teachers. Our sample of 206 lessons across 51 teachers—although large for a special education sample—was likely at the low end of what would be acceptable to conduct CFAs. We also were limited in our ability to conduct G studies that focused on more nuanced sources of variation, such as instructional setting by rater interactions. That said, no existing large-scale studies of the FFT have included special educators, thus ours lends important empirical evidence surrounding its use.

Our sample was also a convenience sample, consisting of teacher volunteers from over 20 districts. On one hand, this allowed us to capture experiences of teachers across a broad range of school contexts, but on the other, we did not have sufficient numbers of teachers in any one district to test patterns associated with certain district characteristics. We examined differences between our teachers and the special education teachers in the same districts who did not elect to participate. Although we did not find systematic differences in grade level, years of experience, highest degree, or teacher evaluation scores, we cannot rule out that our sample differed on variables not captured by Rhode Island's administrative data. One plausible scenario is that our teacher volunteers reflect the kinds of teachers who are highly motivated and willing to open themselves up to critique. We feel reasonably comfortable in suggesting that even if our results are upwardly biased because of a sample of high performers, the true performance of teachers in the state would likely align with our overall finding of consistently low scores on FFT's Domain 3 (Instruction).

It is also worth questioning whether our sample of lessons adequately maps onto the universe of lessons that special educators provide across a school year. As we saw in our sample, special educators commonly work across a variety of subjects and instructional settings. Because special education is by definition individualized instruction, it is common for teachers' responsibilities to vary tremendously across a school day. The literature provides no information on how administrators select classes or service

delivery models in which to observe special educators. Thus, we decided to select lessons in ways that were roughly approximate to how a teacher's time was distributed. If a teacher reported spending 75% of their time coteaching, for example, we observed three out of four lessons in a coteaching context. We argue that this is a sensible approach, but it is fair to question whether the kinds of lessons we observed are the kinds of lessons that teachers would be assessed on in their formal evaluation. It could be the case, for example, that principals systematically chose to not observe coteaching because of the challenges associated with tracking on two teachers' instruction.

Our study would have been strengthened if we were able to examine whether differences in instructional quality (as measured by the FFT and QCI) translate to differences in student academic growth across the teachers in our sample. Unfortunately, we did not have access to student-level data that we could link to teachers. Our best option was to rely on student growth percentiles calculated by the state of Rhode Island, but even still, the state changed assessments the year we collected data and did not calculate student growth measures for our sample of teachers. We ultimately elected to look at teachers' student growth in either of the 2 years prior to our study year, provided that our teachers were in the pool of educators up for evaluation in that year. But even still, many of our teachers had student learning scores based on district-derived student learning objectives. Initial analyses revealed minimal correlation between student test scores and scores we assigned teachers on the FFT. In short, we are limited in what we can say about whether our scores are associated with student learning gains. See Johnson and colleagues (2021) as an example of how future studies may make connections between observation systems and academic outcomes for SWDs.

Finally, we acknowledge that the FFT scoring patterns we see in our data are not directly applicable to the FFT use in practice. Our scores are specific to our research context. Our raters were hired by our research team, had no relationships with the teachers they were observing, received more rigorous training, and were subject to more robust ongoing monitoring activities than actual raters. Plus, there were no consequences attached

to our ratings. In previous research, we have seen that these differences across “use cases” matter for score quality (Liu et al., 2019), and recent reviews suggest that the majority of teachers still receive ratings at the high end of the scoring scale because administrators are reluctant to assign ratings that could jeopardize a teacher’s job status (Kraft & Gilmour, 2017). We see similarly high scoring patterns in Rhode Island’s teacher evaluation data.

Despite these limitations, our study is noteworthy because researchers have so far offered little guidance for using the FFT with special educators. We suggest that the overall trends in our data are notable and warrant further attention, as they suggest that even in a highly structured research context, the FFT is not functioning in the way that researchers or practitioners would desire.

Implications

Research Implications. Our study provides initial validity evidence on using the FFT in special education, as well as a rationale for the further research we need moving forward. Our study focused on the technical properties of the FFT; a complementary line of research that examines how administrators use observation systems like the FFT in practice is also needed.

We find some underlying problems in how the FFT functions with special educators, but there may be additional issues studies like ours cannot answer. For instance, we did not directly investigate how principals assign scores, so we cannot understand what criteria principals apply when scoring teaching. Do principals adapt their use of the FFT in special education settings in ways that are beneficial, or do they not know the criteria they should use to differentiate between special and general education teaching quality? Do principals score in ways that sacrifice accuracy to preserve positive relationships with special education teachers? We suggest a need for studies that trace whether and how administrators’ scoring practices lead to changes in teaching practices over time. Research like this would need to explicitly link the scores administrators assign, the feedback they give teachers, how teachers perceive this feedback, and how teaching improves (or doesn’t) over time. There are

examples of studies that do explore the validity of how principals assign scores. Research by Briggs et al. (2014); Harris and Sass (2014); and Jacob and Lefgren (2009) has found that administrators can identify which of their teachers are most and least likely to produce student learning gains. Meanwhile, researchers have documented the connection between targeted coaching on observation systems and teachers’ improvement (e.g., Hamre et al., 2012), though this work has not been done with principals as observers.

Policy Implications. Our results raise concerns that if states and districts intend on using FFT for evaluative purposes *and* if administrators score in accurate ways, then special educators could be treated unfairly. For a population of teachers that suffers from chronic shortages and high rates of attrition (Boe, 2014; Dewey et al., 2017), the field of special education cannot sustain losses of its best teachers. However, in the current policy environment, with states and districts instead prioritizing professional learning (Croft et al., 2018), the critical question surrounding general observation systems like the FFT is whether they can help all teachers improve. Our study suggests that, in the case of special education teachers, the answer is likely no. FFT scores were almost universally low on instructional components, and the system did not distinguish between teachers who looked more or less effective on an observation system capturing effective special education teaching. Even if we acknowledge the reality that very few teachers are likely to receive low observation scores in practice (as has been observed in studies by Kraft & Gilmour, 2017, and others), it is hard to see how principals could use the FFT to give teachers sufficiently targeted feedback; that is, unless they had background knowledge in special education or received some other additional support.

Given findings like these, how should districts proceed to ensure equitable teacher evaluation policies? One path forward might be to encourage states to allow a broader range of different observation systems for different instructional contexts. But this approach is likely not advisable. It is unlikely that districts would be willing to expend additional costs to train district administrators on an additional observation system, and districts would have to confront—across

multiple systems—the challenges in training administrators to score reliably (Bell et al., 2016; Liu et al., 2019). Although a decision like this could be seen as an avenue for addressing unfairness, it could also present its own issues related to equity, as some observation systems might be easier or harder to be rated highly by an administrator.

Other policy strategies could be taken that would be less costly and less ambitious. First, states or districts could consider using alternative scoring criteria on Instruction domain components where special educators consistently score low: *Using Questioning and Discussion Techniques* and *Using Assessment* (or they could forgo scoring on these components altogether). If alternative criteria were used, those scores would not have to be included in the teacher evaluation system until they were deemed appropriate. In the short term, this would give teachers scores on scales that better reflect high-quality special education practice, and it may support richer conversations with administrators that could support teacher learning.

A second approach could be for districts to lean on special education administrators or other leaders with certification in special education to conduct observations of special educators. The logic would be that these administrators would have the necessary background to more accurately assess the quality of a special educator's instruction. And even if their formal observation results did not align with effective special education teaching, the administrators would still be poised to provide feedback that would be

sufficiently targeted to guide teacher improvement efforts.

We argue that, more generally, districts would benefit from looking across their professional learning systems—including their evaluation processes but also other policies and programs (e.g., professional development)—to ensure that these components reflect the range of instructional practices that support students. As we have argued in this article, observation systems such as the FFT do not capture the kinds of instruction that many struggling students need. Studies suggest that all students benefit from explicit, systematic instruction (Clements et al., 2013; Cohen, 2018; Stockard et al., 2018; Taylor, 2018), with such instruction being particularly beneficial among lower performing students (e.g., Connor & Morrison, 2016; Morgan et al., 2015; Sonnenschein et al., 2010). If districts are to ensure that lower performing students have equitable access to instruction that is aligned with their needs, it will likely take a concerted investment of time and resources. Districts will likely need to deepen the professional learning opportunities available to all of their educators to support these students. We are not suggesting that schools and districts abandon their efforts to foster the kinds of inquiry-oriented instruction that the literature suggests leads to deeper, conceptual understanding. Instead, we hope to highlight the urgent need for researchers, policymakers, and practitioners to take on more flexible definitions of *effective teaching*, prioritizing teachers' use of a range of instructional tools that more equitably maps onto the needs of all learners.

Appendix

APPENDIX TABLE A1

FFT Domains and Components

Domain	Component
Domain 2: Classroom Environment	2a. Creating an Environment of Respect and Rapport (Respect) 2b. Establishing a Culture for Learning (Culture) 2c. Managing Classroom Procedures (Procedures) 2d. Managing Student Behavior (Behavior) 2e. Organizing Physical Space (Space)
Domain 3: Instruction	3a. Communicating With Students (Communication) 3b. Using Questioning and Discussion Techniques (Questioning) 3c. Engaging Students in Learning (Engagement) 3d. Using Assessment in Instruction (Assessment) 3e. Demonstrating Flexibility and Responsiveness (Responsiveness)

Note. FFT = Framework for Teaching.

APPENDIX TABLE A2

QCI Principles

Principle	Description
Principle 1	Models skills/concepts appropriately and with ease
Principle 2	Uses timely checks to ensure student understanding
Principle 3	Provides adequate think and response time for students
Principle 4	Engages students in learning throughout the lesson
Principle 5	Ensures high rate of success for students
Principle 6	Encourages effort from all students
Principle 7	Transitions from one activity to the next in an appropriate fashion
Principle 8	Maintains good pacing

Note. QCI = Quality of Classroom Instruction.

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 supported in part by a grant from the Institute of Education Sciences (Award # R324A150231)

ORCID iDs

Courtney A. Bell  <https://orcid.org/0000-0001-8743-5573>

David Peyton  <https://orcid.org/0000-0003-4266-0069>

Notes

1. We use the term observation system when referring to the scoring tools, rating quality procedures, and sampling specifications (Bell et al., 2018), and observation tool when referring to the rubrics used to rate classroom interactions.

2. Special education is, by definition, individualized instruction tailored to students' unique needs. As a consequence, special educators deliver instruction in a variety of settings with a variety of instructional goals. We focus on high-incidence student populations in a limited grade band, but we recognize that in doing so we may not capture the full variation of special education teaching.

3. Although we acknowledge other validity frameworks, including M. T. Kane's (2006) argument approach and Messick's unified theory of construct validity (1995), as being widely used in the field, we

find the more contemporary approach articulated by M. T. Kane and Wools (2019) to be more closely relevant for the policy issue of teacher evaluation.

4. The following procedures were used to recruit teachers. First, we contacted all special education directors in districts across Rhode Island. Upon securing district agreement to participate in the project, we solicited teacher volunteers. Teachers were compensated \$100 for completing all four lessons.

5. Lessons were also scored on the QCI's companion observation system, the Classroom Observation of Student-Teacher Interactions (COSTI; Doabler et al., 2015; Smolkowski & Gunn, 2012). The COSTI uses interval recording to track teachers' frequency of using key practices aligned with effective instruction for SWDs. Observers used the COSTI in determining QCI scores for specific lessons.

6. We drew on separate rater pools for FFT scoring and QCI scoring to ensure that scores across the two measures did not reflect effects of specific raters.

7. Due to unforeseen circumstances, two raters completed abbreviated assignments, thus extending the overall scoring time period. One rater had increased responsibilities at her place of work and removed herself from the project during Month 3. A second rater requested a reduced number of videos due to unforeseen family matters during Month 4.

8. Liu et al. (2019) carried out a G study on FFT ratings from two studies. However, due to the design of the data collection in one of the studies, the variance could only be decomposed into one source and the residual. This does not offer a helpful contextualization for the purposes of this analysis.

References

- Al Otaiba, S., Connor, C. M., Folsom, J. S., Greulich, L., Meadows, J., & Li, Z. (2011). Assessment data-informed guidance to individualize kindergarten reading instruction: Findings from a cluster-randomized control field trial. *Elementary School Journal, 111*(4), 535–560. <https://doi.org/10.1086/659031>
- Baker, S. K., Gersten, R., Haager, D., & Dingle, M. (2006). Teaching practice and the reading growth of first-grade English learners: Validation of an observation instrument. *The Elementary School Journal, 107*(2), 199–220.
- Bays, D. A., & Crockett, J. B. (2007). Investigating instructional leadership for special education. *Exceptionality, 15*(3), 143–161.
- Bell, C. A., Dobbelaer, M. J., Klette, K., & Visscher, A. (2019). Qualities of classroom observation systems. *School Effectiveness and School Improvement, 30*(1), 3–29. <https://doi.org/10.1080/09243453.2018.1539014>
- Bell, C. A., Gitomer, D. H., McCaffrey, D., Hamre, B., Pianta, R., & Qi, Y. (2012). An argument approach to observation protocol validity. *Educational Assessment, 17*(2–3), 62–87.
- Bell, C. A., Jones, N., Lewis, J., Qi, Y., Stickler, L., Liu, S., & McLeod, M. (2016). *Understanding consequential assessment systems of teaching: Year 1 final report to Los Angeles Unified School District* (Research Memorandum No. RM-16-12). Educational Testing Service.
- Bell, C. A., Jones, N., Qi, Y., & Lewis, J. (2018). Strategies for assessing classroom teaching: Examining administrator thinking as validity evidence. *Educational Assessment, 23*(4), 229–249.
- Bell, C. A., Qi, Y., Croft, A. J., Leusner, D., McCaffrey, D. F., Gitomer, D. H., & Pianta, R. C. (2014). Improving observational score quality: Challenges in observer thinking. 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. 50–97). Wiley.
- Bettini, E., Mason-Williams, L., & Barber, B. (2019). Access to qualified, well-supported principals across alternative educational settings and neighborhood schools. *Journal of Special Education, 53*(4), 195–205. <https://doi.org/10.1177/0022466919831302>
- Boe, E. (2014). Teacher demand, supply, and shortage in special education: A national perspective. In P. Sindelar, E. McCray, M. T. Brownell, & B. Lignugaris-Kraft (Eds.), *Handbook of research on special education teacher preparation* (pp. 67–93). Routledge.
- Brennan, R. L. (2001). *Generalizability theory*. Springer.
- Briggs, D. C., Dadey, N., & Kizil, R. C. (2014). *Comparing student growth and teacher observation to principal judgments in the evaluation of teacher effectiveness*. University of Colorado.
- Brownell, M. T., Bishop, A. B., Gersten, R., Klingner, J., Dimino, J., Haager, D., Menon, S., Penfield, R., & Sindelar, P. T. (2009). The role of domain expertise in beginning special education teacher quality. *Exceptional Children, 75*(4), 391–411.
- Casabianca, J. M., Lockwood, J. R., & McCaffrey, D. F. (2015). Trends in classroom observation scores. *Educational and Psychological Measurement, 75*(2), 311–337. <https://doi.org/10.1177/0013164414539163>
- Cash, A. H., & Pianta, R. C. (2014). The role of scheduling in observing teacher-child interactions. *School Psychology Review, 43*(4), 428–449.
- Chudowsky, N., Chudowsky, V., & Keber, N. (2009). *State test score trends through 2007–08, part 4: Has progress been made in raising achievement*

- for students with disabilities? Center on Education Policy.
- Clemens, D. H., Agodini, R., & Harris, B. (2013). *Instructional practices and student achievement: Correlations from a study of math curricula* [NCEE evaluation brief]. U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.
- Cohen, J. (2018). Practices that cross disciplines? Revisiting explicit instruction in elementary mathematics and English language arts. *Teaching and Teacher Education, 69*(1), 324–335. <https://doi.org/10.1016/j.tate.2017.10.021>
- Connor, C. M., Jakobsons, L. J., Crowe, E. C., & Meadows, J. G. (2009). Instruction, student engagement, and reading skill growth in Reading First classrooms. *The Elementary School Journal, 109*(3), 221–250.
- Connor, C. M., Mazzocco, M. M., Kurz, T., Crowe, E. C., Tighe, E. L., Wood, T. S., & Morrison, F. J. (2018). Using assessment to individualize early mathematics instruction. *Journal of School Psychology, 66*, 97–113. <https://doi.org/10.1016/j.jsp.2017.04.005>
- Connor, C. M., & Morrison, F. J. (2016). Individualizing student instruction in reading: Implications for policy and practice. *Policy Insights From the Behavioral and Brain Sciences, 3*(1), 54–61.
- Connor, C. M., Morrison, F. J., Fishman, B. J., Schatschneider, C., & Underwood, P. (2007). Algorithm-guided individualized reading instruction. *Science, 315*(5811), 464–465. <https://doi.org/10.1126/science.1134513>
- Connor, C. M., Morrison, F. J., & Petrella, J. N. (2004). Effective reading comprehension instruction: Examining child x instruction interactions. *Journal of Educational Psychology, 96*(4), 682–698.
- Croft, M., Guffy, G., & Vitale, D. (2018). *The shrinking use of growth: Teacher evaluation legislation since ESSA*. ACT.
- Cronbach, L. J. (1988). Five perspectives on validation argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3–17). Lawrence Erlbaum.
- Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, M. (1972). *The dependability of behavioral measurement: Theory of generalizability for scores and profiles*. John Wiley.
- Danielson, C. (1996). *Enhancing professional practice: A framework for teaching*. Association for Supervision and Curriculum Development.
- Danielson, C. (2013). *The framework for teaching evaluation instrument, 2013 edition*. The Danielson Group.
- Dewey, J., Sindelar, P. T., Bettini, E., Boe, E. E., Rosenberg, M. S., & Leko, C. (2017). Explaining the decline in special education teacher employment from 2005 to 2012. *Exceptional Children, 83*(3), 315–329. <https://doi.org/10.1177/0014402916684620>
- Doabler, C. T., Baker, S. K., Kosty, D. B., Smolkowski, K., Clarke, B., Miller, S. J., & Fien, H. (2015). Examining the association between explicit mathematics instruction and student mathematics achievement. *The Elementary School Journal, 115*(3), 303–333. <https://doi.org/10.1086/679969>
- Donaldson, M. L., & Woulfin, S. (2018). From tinkering to going “rogue”: How principals use agency when enacting new teacher evaluation systems. *Educational Evaluation and Policy Analysis, 40*(4), 531–556. <https://doi.org/10.3102/0162373718784205>
- Ferguson, R. F., & Danielson, C. (2014). How framework for teaching and Tripod 7Cs 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). Wiley.
- Fuchs, L. S., Newman-Gonchar, R., Schumacher, R., Dougherty, B., Bucka, N., Karp, K. S., Woodward, J., Clarke, B., Jordan, N. C., Gersten, R., Jayanthi, M., Keating, B., & Morgan, S. (2021). *Assisting students struggling with mathematics: Intervention in the elementary grades* (WWC 2021006). National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. <http://whatworks.ed.gov/ffuchs>
- Fuchs, L. S., Powell, S. R., Seethaler, P. M., Cirino, P. T., Fletcher, J. M., Fuchs, D., Hamlett, C. L., & Zumeta, R. O. (2009). Remediating number combination and word problem deficits among students with mathematics difficulties: A randomized control trial. *Journal of Educational Psychology, 101*(3), 561–576. <https://doi.org/10.1037/a0014701>
- Gilmour, A. F., & Jones, N. D. (2020). Policies that define instruction: A systematic review of states’ and districts’ recommendations for evaluating special educators. *Educational Researcher, 49*(9), 645–655. <https://doi.org/10.3102/0013189X20935039>
- Hamre, B. K., Pianta, R. C., Mashburn, A. J., & Downer, J. T. (2012). Promoting young children’s social competence through the preschool PATHS curriculum and MyTeachingPartner professional development resources. *Early Education & Development, 23*(6), 809–832. <https://doi.org/10.1080/10409289.2011.607360>
- Harris, D. N., & Sass, T. R. (2014). Skills, productivity and the evaluation of teacher perfor-

- mance. *Economics of Education Review*, 40, 183–204. <https://doi.org/10.1016/j.econedurev.2014.03.002>
- Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher*, 41(2), 56–64. <https://doi.org/10.3102/0013189X12437203>
- 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.
- Ho, A. D., & Kane, T. J. (2013). *The reliability of classroom observations by school personnel* [Research paper, MET project]. Bill & Melinda Gates Foundation.
- Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1–55.
- Jacob, B. A., & Lefgren, L. (2009). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics*, 26(1), 101–136. <https://doi.org/10.1086/522974>
- Johnson, E. S., Crawford, A. R., Zheng, Y., & Moylan, L. A. (2020). Does special educator effectiveness vary depending on the observation instrument used? *Educational Measurement: Issues and Practice*, 40(1), 36–43.
- Johnson, E. S., Zheng, Y., Crawford, A. R., & Moylan, L. A. (2021). The relationship of special education teacher performance on observation instruments with student outcomes. *Journal of Learning Disabilities*, 54(1), 54–65. <https://doi.org/10.1177/0022219420908906>
- Jones, N. D., & Brownell, M. T. (2014). Examining the use of classroom observations in the evaluation of special education teachers. *Assessment for Effective Intervention*, 39(2), 112–124. <https://doi.org/10.1177/1534508413514103>
- Kalyuga, S. (2007). Expertise reversal effect and its implications for learner-tailored instruction. *Educational Psychology Review*, 19(4), 509–539.
- Kane, M. T. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). American Council on Education/Praeger.
- Kane, M. T., & Woolls, S. (2019). Perspectives on the validity of classroom assessments. In S. M. Brookhart & J. H. McMillan (Eds.), *Classroom assessment and educational measurement* (1st ed., pp. 11–26). Routledge.
- Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains*. Measures of Effective Teaching Project, Bill & Melinda Gates Foundation.
- Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist*, 41, 75–86. https://doi.org/10.1207/s15326985ep4102_1
- Klingner, J. K., Brownell, M., Mason, L. H., Sindelar, P. T., Benedict, A. E., Griffin, C., & Park, Y. (2016). Teaching students with special needs in the new millennium. In D. Gitomer & C. Bell (Eds.), *Handbook of research on teaching* (5th ed., pp. 639–717). American Educational Research Association.
- Kraft, M. A., & Gilmour, A. F. (2017). Revisiting the widget effect: Teacher evaluation reforms and the distribution of teacher effectiveness. *Educational Researcher*, 46(5), 234–249.
- Le, V. Y., Stecher, B. M., Lockwood, J. R., Hamilton, L. S., Robyn, A., Williams, V. L., Ryan, G. W., Kerr, K. A., Martinez, J. F., & Klein, S. P. (2006). *Improving mathematics and science education: A longitudinal investigation of the relationship between reform-oriented instruction and student achievement*. Rand Corporation.
- Liu, S., Bell, C. A., Jones, N. D., & McCaffrey, D. F. (2019). Classroom observation systems in context: A case for the validation of observation systems. *Educational Assessment, Evaluation and Accountability*, 31(1), 61–95.
- Mantzicopoulos, P., French, B. F., Patrick, H., Watson, J. S., & Ahn, I. (2018). The stability of kindergarten teachers' effectiveness: A generalizability study comparing the framework for teaching and the classroom assessment scoring system. *Educational Assessment*, 23(1), 24–46. <https://doi.org/10.1080/10627197.2017.1408407>
- Martin, A. J. (2016). *Using load reduction instruction (LRI) to boost motivation and engagement*. British Psychological Society.
- 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. <https://doi.org/10.1007/s1121-012-0357-3>
- Mayer, R. E. (2003). Memory and information processes. In W. M. Reynolds & G. E. Miller (Eds.), *Handbook of psychology* (pp. 47–58). Wiley.
- McCaffrey, D. F., Yuan, K., Savitsky, T. D., Lockwood, J. R., & Edelen, M. O. (2015). Uncovering multi-

- variate structure in classroom observations in the presence of rater errors. *Educational Measurement: Issues and Practice*, 34(2), 34–46.
- Mercer, C. D., Lane, H. B., Jordan, L., Allsopp, D. H., & Eisele, M. R. (1996). Empowering teachers and students with instructional choices in inclusive settings. *Remedial and Special Education*, 17(4), 226–236.
- Meyer, J. P., Cash, A. H., & Mashburn, A. (2011). Occasions and the reliability of classroom observations: Alternative conceptualizations and methods of analysis. *Educational Assessment*, 16(4), 227–243. <https://doi.org/10.1080/10627197.2011.638884>
- Morgan, P. L., Farkas, G., & Maczuga, S. (2015). Which instructional practices most help first-grade students with and without mathematics difficulties? *Educational Evaluation and Policy Analysis*, 37(2), 184–205. <https://doi.org/10.3102/0162373714536608>
- Morris-Mathews, H., Stark, K., Jones, N. D., Brownell, M., & Bell, C. (2020). Danielson's framework for teaching: Convergence and divergence with conceptions of effectiveness in special education. *Journal of Learning Disabilities*, 54(1), 66–78. <https://doi.org/10.1177/0022219420941804>
- Oliveri, M., McCaffrey, D., Ezzo, C., & Holtzman, S. (2017). A multilevel factor analysis of third-party evaluations of noncognitive constructs used in admissions decision making. *Applied Measurement in Education*, 30(4), 297–313. <https://doi.org/10.1080/08957347.2017.1353989>
- Praetorius, A.-K., Pauli, C., Reusser, K., Rakoczy, K., & Klieme, E. (2014). One lesson is all you need? Stability of instructional quality across lessons. *Learning and Instruction*, 31, 2–12. <https://doi.org/10.1016/j.learninstruc.2013.12.002>
- Rhode Island Department of Education. (2017). *Rhode Island model evaluation & support system guidebook—Teacher*. http://www.ride.ri.gov/Portals/0/Uploads/Documents/Teachers-and-Administrators-Excellent-Educators/Educator-Evaluation/Guidebooks-Forms/Teacher_Guidebook_2015-16.pdf
- Schulte, A. C., & Stevens, J. J. (2015). Once, sometimes, or always in special education: Mathematics growth and achievement gaps. *Exceptional Children*, 81(3), 370–387. <https://doi.org/10.1177/0014402914563695>
- Schulte, A. C., Stevens, J. J., Elliott, S. N., Tindal, G., & Nese, J. F. T. (2016). Achievement gaps for students with disabilities: Stable, widening, or narrowing on a state-wide reading comprehension test? *Journal of Educational Psychology*, 108(6), 925–942.
- Searle, S. R., Casella, G., & McCulloch, C. E. (2009). *Variance components* (Vol. 391). John Wiley.
- Semmelroth, C. L., & Johnson, E. (2014). Measuring rater reliability on a special education observation tool. *Assessment for Effective Intervention*, 39(3), 131–145. <https://doi.org/10.1177/1534508413511488>
- Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer* (Vol. 1). SAGE.
- Smolkowski, K., & Gunn, B. (2012). Reliability and validity of the Classroom Observations of Student–Teacher Interactions (COSTI) for kindergarten reading instruction. *Early Childhood Research Quarterly*, 27(2), 316–328.
- Snow, R. (1989). Aptitude-treatment interaction as a framework for research on individual differences in learning. In P. Ackerman, R. J. Sternberg, & R. Glaser (Eds.), *Learning and individual differences* (pp. 13–59). W. H. Freeman.
- Sonnenschein, S., Stapleton, L. M., & Benson, A. (2010). The relation between the type and amount of instruction and growth in children's reading competencies. *American Educational Research Journal*, 47(2), 358–389. <https://doi.org/10.3102/0002831209349215>
- 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
- Steinbrecher, T. D., Fix, R., Mahal, S. A., Serna, L., & McKeown, D. (2015). All you need is patience and flexibility: Administrators' perspectives on special educator knowledge and skills. *Journal of Special Education Leadership*, 28(2), 89–102.
- Stelitano, L., Russell, J. L., & Bray, L. E. (2020). Organizing for meaningful inclusion: Exploring the routines that shape student supports in secondary schools. *American Educational Research Journal*, 57(2), 535–575. <https://doi.org/10.3102/0002831219859307>
- Stockard, J., Wood, T. W., Coughlin, C., & Rasplica Khoury, C. (2018). The effectiveness of direct instruction curricula: A meta-analysis of a half century of research. *Review of Educational Research*, 88(4), 479–507. <https://doi.org/10.3102/0034654317751919>
- Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2), 257–285.
- Taylor, E. S. (2018). Skills, job tasks, and productivity in teaching: Evidence from a randomized trial of instruction practices. *Journal of Labor Economics*, 36(3), 711–742.
- Vadasy, P. F., Sanders, E. A., & Peyton, J. A. (2005). Relative effectiveness of reading practice or word-level instruction in supplemental tutoring: How text matters. *Journal of Learning Disabilities*, 38(4), 364–380. <https://doi.org/10.1177/00222194050380041401>

- Vaughn, S., Gersten, R., & Chard, D. (2000). The underlying message in LD intervention research: Findings from research syntheses. *Exceptional Children, 67*(1), 99–114. <https://doi.org/10.1177/001440290006700107>
- Vaughn, S., Klingner, J. K., Swanson, E. A., Boardman, A. G., Roberts, G., Mohammad, S. S., & Stillman-Spisak, S. J. (2011). Efficacy of collaborative strategic reading with middle school students. *American Educational Research Association, 48*(4), 938–964. <https://doi.org/10.3102/0002831211410305>
- Wanzek, J., Vaughn, S., Scammacca, N. K., Metz, K., Murray, C. S., Roberts, G., & Danielson, L. (2013). Extensive reading interventions for students with reading difficulties after grade 3. *Review of Educational Research, 83*(2), 163–195. <https://doi.org/10.3102/0034654313477212>
- Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness*. The New Teacher Project.
- Xue, Y., & Meisels, S. J. (2004). Early literacy instruction and learning in kindergarten: Evidence from the Early Childhood Longitudinal Study–Kindergarten Class of 1998–99. *American Educational Research Journal, 41*(1), 191–229. <https://doi.org/10.3102/00028312041001191>

Authors

NATHAN D. JONES is an associate professor of special education and education policy at Boston University. His research focuses on teacher quality and teacher development. A particular focus over the last several years has been on the quality of instruction received by students with disabilities in both general education and special education.

COURTNEY A. BELL is a professor of learning sciences and director of the Wisconsin Center for Education Research (WCER) at the University of Wisconsin. Her research focuses on teaching quality, teacher education, and the validity of teaching measures. Her most recent study is a multimodal study of the relationships between teaching and learning in eight countries around the world. In all of her work, she is interested in the intersections of research, policy, and practice.

MARY BROWNELL is a Distinguished Professor of special education at the University of Florida and

director of the Collaboration for Effective Educator Development, Accountability and Reform (CEEDAR Center). Her research interests focus on teaching quality and how to improve it, with specific emphasis on teacher education pedagogy, teacher assessment, and professional development innovations, primarily in the area of reading.

YI QI is a senior research associate at Educational Testing Service. Her work centers around policy and evaluation issues related to teacher education and how best to measure teaching practice. She has worked on large-scale studies on using classroom observation protocols, along with other measures to evaluate teaching quality.

DAVID PEYTON is an assistant professor in special education at Appalachian State University. His research focuses on special education teacher quality, special education teacher preparation, and teacher observation. He is currently studying how special education teaching quality is impacted by access to resources, training, and qualified support personnel.

DAISY PUA is a postdoctoral scholar at Mary Lou Fulton Teachers College, Arizona State University. Her research interests include evidence-based literacy practices for supporting students with disabilities and teacher preparation and training. She is most interested in aligning classroom observations with professional development.

MELISSA FOWLER is a research associate at Educational Testing Service. Her research interests are in education equity, teacher quality, and how social identity impacts student learning. Her current research is centered around performance assessments for teacher licensure.

STEVEN HOLTZMAN is a principal research data analyst at Educational Testing Service. His work at ETS concentrates on using study design, data collection, data management, and data analysis methods to help promote research in education. Recent projects have examined noncognitive assessments, teacher evaluation, and workforce selection assessments.

Manuscript received June 12, 2020

First revision received May 25, 2021

Second revision received August 31, 2021

Accepted November 2, 2021