Using Classroom Observations in the Evaluation of Special Education Teachers

Nathan D. Jones

Courtney A. Bell

Mary Brownell

Yi Qi

David Peyton

Daisy Pua

Melissa Fowler

Steven Holtzman

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

Author Note

Correspondence concerning this article should be addressed to Nathan Jones, BU Wheelock College of Education & Human Development, Boston University, 2 Silber Way, Boston MA 02215. E-mail: ndjones@bu.edu

Abstract

We examine the degree to which the Framework for Teaching (FFT), an observation tool that has been widely adopted across the country, reliably and validly captures special education teachers' instruction for the improvement of teaching and for human capital decisions. Drawing on approximately 200 videotaped lessons from 51 special education teachers in Rhode Island, we investigate FFT scores created by well-trained raters using standardized scoring procedures, and we compare these scores to other sources of data. Our findings suggest that the tool's psychometric properties when used with special educators are in line with previous studies of FFT. However, our evidence raises serious concerns about whether observation tools like FFT can be used to improve special education teaching quality.

Key words: special education, teacher evaluation, classroom observations, teaching quality, teacher quality, validity

Introduction

Validated observation tools have figured prominently in the teacher evaluation and development landscape over the past decade. At present, all 50 states require 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, 2015).

These observation systems are being used for two critical purposes -- to inform human capital decisions and to guide teacher improvement efforts. When included as a component of formal teacher evaluation, observation scores provide information on the quality of a teacher's professional practice and complement information used in the district's human capital management enterprise, such as the academic growth of students and the fulfillment of teacher's non-instructional professional responsibilities.[i] These observation systems can also be used to improve teaching by providing critical information to both school and district administrators and teachers about the specific instructional practices to target for improvement.

In addition to being used for similar purposes, the observation systems used across states and districts share similar features. For scaling and cost reasons, adopted observation systems tend to be general (content-agnostic) and not targeted to specific grade levels or student populations (Hill & Grossman, 2013).[ii] Additionally, these tools are modeled on typical general education instruction. As such, they reflect perspectives of teaching and learning that are commonly espoused in the general education community (Hill & Grossman, 2013; Jones & Brownell, 2014), and they reflect general education instructional settings, in which a teacher

provides academic instruction to a classroom of students around a common, shared classroom learning objective.

How well such observation tools work for populations of teachers who fall outside of "typical" instruction – those teaching in subject areas or formats different from core academic instruction – warrants further investigation. These teachers include, for example, special education teachers, teachers of English learners, career and technical educators, and reading specialists, among others. In the state of Rhode Island, for example, this category of "non-typical" teacher accounts for roughly 36% of the teaching workforce. We are not aware of any existing studies that inform the question of whether the decision to use general observation tools across these groups of teachers matters for the purposes of either evaluating them or helping them improve.

In this study, we investigate the validity of using general tools with one group of non-typical teachers, special educators. Special education serves as a compelling case of this broader phenomenon for several reasons. They make up approximately 12% of teachers in schools. Unlike general educators, special educators are typically charged with providing intensive, explicit, teacher-directed instruction to small groups of students. Often, this instruction is defined based on the specific, individualized needs of students with disabilities (SWDs), which may be academic, social and emotional, or functional in nature. For these reasons, researchers have raised questions about the appropriateness of using general observation tools with special educators (e.g., Jones & Brownell, 2014; Mathews et al., 2020). Yet most states have not taken any steps to refine observation systems (or provide supplemental guidance) to support the evaluation of special educators (Gilmour & Jones, 2020).

To understand how general observation systems function for special educators, we focused our attention on Danielson's Framework for Teaching (FFT). FFT is the most widely-used observation system in the United States and is implemented in at least 29 states at present (American Institutes for Research, 2012). We examine the degree to which scores from FFT reliably and validly capture special education teachers' classroom teaching. We do this by treating validity as an evidence-based argument (Bell et al., 2012; Hill, Kapitula & Umland, 2011; Kane, 2013) in which evidence is gathered and then evaluated against a set of validity criteria. The criteria used come from what Kane and Wools (2019) refer to as the measurement perspective and what Cronbach (1988) referred to as the functional perspective on validity. While the two perspectives are related and overlap, they each emphasize different validity concerns raised by the use of FFT scores to guide human capital decisions and improve teaching. The measurement perspective emphasizes the reliability, accuracy, and generalizability of FFT scores; the functional perspective emphasizes the degree to which FFT scores have "appropriate consequences for individuals and institutions" (Cronbach, 1988, p.6) and do not result in adverse consequences.

Drawing on approximately 200 videotaped lessons from 51 special education teachers in Rhode Island, we investigate FFT scores created by well-trained raters using standardized scoring procedures. In line with the measurement perspective, we examine overall scoring patterns across the scoring scale, the distribution of scores across teachers and across lesson types, and the reliability of the scores. To consider the degree to which FFT captures special education teaching quality as defined by the special education community, we compare FFT scores to scores on an observation system that reflect the special education community's consensus views of high-quality teaching practices: the Quality of Classroom Instruction (QCI)

(Doabler et al., 2015). Finally, taking a more functional perspective, we examine whether scores produced by trained raters are similar to those created by local administrators within Rhode Island's teacher evaluation system. Our findings provide the first comprehensive empirical evidence of how a popular observation system functions for special education teaching.

FFT's Conceptualization of Teaching

Charlotte Danielson's Framework for Teaching (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 appendix).

The FFT is based on a constructivist conception of teaching. It posits that effective instruction requires teachers to play the role of facilitator and enact strategies that deeply engage students in content, promote critical thinking and reasoning skills, and encourage students in intellectual argumentation (Danielson, 2013). High scores on the instructional domain of FFT are undergirded by two important assumptions (Matthews et al., 2020). First, high-quality teaching values students' construction of their learning. 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. Second, the FFT assumes that advanced teaching is based upon the extent to which student learning is anchored to conceptual knowledge. 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 (Mathews et al., 2020). While this view of instruction may have substantial benefits for students in general education settings, prior research in special education suggests that the degree to which these approaches are effective is

less well known. In the next section, we examine the potential consequences of using the FFT in special education, a field that approaches learning from perspectives more in line with cognitivism and behaviorism.

Using FFT with Special Educators

Scholars in special education have raised concerns about the degree to which FFT is a suitable tool for assessing effective special education teaching (Jones & Brownell, 2014; Sledge & Pazey, 2013); these concerns center on how FFT conceptualizes effective teaching and how general versus special education instruction is enacted.

Different Conceptions of Teaching

Brownell and Jones (2014) and Sledge and Pazey (2013) argue that FFT privileges an approach to teaching and learning that is not well-aligned with views of effective special education instruction. FFT describes and measures general features of teaching practice that reflect mostly constructivist views of teaching and learning. Specifically, high performing teachers are described as facilitators rather than directors of student learning; they foster discussions of content that help students make sense of what they are learning, encourage them to formulate questions, initiate topics, and challenge their peers' thinking.

Effective special education teaching, in contrast, is grounded in behavioral, social learning, and information processing theories. These theories have been translated into intervention research, which demonstrates repeatedly that students with disabilities and other low achieving students learn best when their instruction is intensive, explicit and systematic. Such instruction (a) involves extensive teacher modeling and controlled practice, (b) is highly interactive, (c) uses multiple practice opportunities and consistent feedback, and (d) incorporates ongoing progress monitoring to see if changes should be made to instruction (Archer & Hughes,

2010; Doabler et al., 2015; Swanson & O'Connor, 2009; Vadasy, Sanders, & Peyton, 2005; Vaughn, Gersten, & Chard, 2000). Further, teachers apply these instructional processes systematically – progressing from less to more complex strategies, skills, and concepts in order to facilitate students' application, generalization, and transfer of new learning. Special education researchers argue that this intensive, explicit and systematic instruction helps to reduce cognitive load and thereby support the learning of students with disabilities, many of whom have working memory and cognitive processing difficulties.

A recent content analysis of FFT's instruction domain demonstrates that theoretical perspectives underlying practices articulated in FFT fundamentally differ from those underlying long-standing conceptions of effective teaching in special education (Mathews et al., 2020). In a systematic content analysis of the instrument, Mathews and colleagues found that FFT operationalized effective instruction in ways that were aligned with constructivist and social constructivist theories and either disadvantaged or overlooked views of effective special education teachers. Attributes of effective special education instruction, such as explicit and systematic instruction, were not present in 93% of the sentence units describing different levels of performance in the instruction domain. Further, some aspects of effective special education instruction – such as providing multiple practice opportunities – were never mentioned in the tool.

Unique Nature of Special Education Instruction

The FFT was developed to assess general education instruction as it frequently occurs in schools – a single teacher assigned to a large group of students around a broad instructional topic (e.g., 2nd grade math, 6th grade science, etc.). Some characteristics of special education instruction differ from this general education context. First, whereas general education teachers

typically provide instruction to large groups of students designed to help them achieve grade level standards in a content area, special education instruction is designed to meet students' individual needs as articulated in an individualized educational plan (Fuchs, Fuchs, & Stecker, 2010). Second, special education teachers often target particular skills and strategies within a content area, which are often focused on building students' fluency and automaticity with foundational content (e.g., decoding or operations with fractions). As such, special education teachers may not have the opportunity to demonstrate the broad range of instructional behaviors general education teachers exhibit when teaching, focusing instead on a narrower segment of the curriculum. For instance, if a special education teacher targets decoding and fluency instruction for an individual student because it best aligns with their individual needs, the teacher may intentionally limit the number of higher order questions they ask. Finally, special education teachers often share the instructional load with other educators and paraprofessionals, and observers might find it challenging to evaluate the special education teacher's instruction isolated from instruction provided by the other adults (Hock & Isenberg, 2017). When observing special education teachers' practice, each of these differences may influence how raters assign scores to lessons.

In sum, there are two primary reasons why the instruction observed in special education might deviate from the vision of effectiveness outlined in FFT. First, the two fields draw on two distinct theoretical traditions surrounding learning and pedagogy, with special education privileging instruction that draws on principles of behaviorism and cognitive processing, where the teacher plays a more direct role in supporting novice learners as they build expertise. Second, students with disabilities have educational needs that require additional instruction and practice in building foundational skills, which often results in a more teacher-directed role. In other

words, students' needs mandate that teachers take on the responsibility of designing learning opportunities that guide students toward mastery. The relevant distinction between FFT and special education teaching then is in the role of the teacher in defining and guiding students through learning activities. We would not expect students to know how to design activities to address memory or processing challenges, particularly when they are young. Thus, across general education and special education, the instructional goal is ultimately the same: to ensure that students are actively engaged in using higher-order thinking. But for students with disabilities, a highly skilled educator is one who can rapidly get them to mastery so that they can engage in these activities. The analyses in this paper provide evidence about the degree to which these theoretical and empirical concerns about differences between special and general educators' work threaten the validity of FFT scores.

Conceptual Framework

Examining Validity Evidence from Measurement and Functional Perspectives

Given questions about the appropriateness of using FFT to capture special education teaching practices, we follow Kane and Wools' (2019) framework for assessing the validity of classroom assessments. Specifically, we investigate the observation tool'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). 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 while assessments may be "truthful", i.e., they have adequate measurement characteristics, they may or

may not be worthwhile. For example, if 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. Further, 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 FFT's validity in the evaluation of special educators.

The Measurement Perspective

A measurement perspective has been used in other validation arguments of observation tools such as the *Classroom Assessment Scoring System*, the *Mathematical Quality of Instruction*, and the *English Language Learner Classroom Observation Instrument* (Baker et al., 2013; Bell et al., 2012; Mantzicopoulos, French, & Patrick, 2018). 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 the 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 (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, or 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, Charalambous, & Kraft, 2012). These facets could be raters, lessons, or different classroom compositions. Generalizability studies frequently provide relevant evidence on this point (c.f., 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 is 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 teachers, 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 helps achieve specific purposes. In the case of FFT, we are concerned with how useful scores are in evaluating and improving teaching and the degree to which undesirable outcomes are minimized. To establish evidence of FFT's worth, we conduct empirical assessments in two ways -- by considering the degree to which FFT is aligned to the special education community's views of high-quality teaching and by considering the consequences of FFT scores in Rhode Island.

Aligned to the community's views. The worth of FFT to the two stated purposes depends in part on how well it aligns the community's views of teaching quality. Various communities have different views on teaching. Administrators' views are likely to align with the general education perspectives detailed previously; special educators likely have views more aligned to those of the special education community. To investigate the worth of FFT scores, we focus on the degree to which FFT captures what special educators consider high quality teaching, but we also consider how those views relate to general educators' views empirically.

Appropriate consequences. To improve teaching quality, observation scores are often used to inform leadership, certification, and probationary decisions (the human capital management goal) and 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 and then those scores are used to assign overall teacher evaluation scores that are themselves associated with specific consequences. Those consequences include being eligible for specific teacher leadership opportunities and being put on probationary status, which requires more observations and a professional learning plan (e.g. Rhode Island Department of Education, 2016). The observation scores are also used as a part of the formalized reflection and improvement conversations between administrators and teachers, as dictated by the teacher evaluation policy. The appropriateness and fairness of these consequences should be evaluated empirically and logically. If, for example, research-based practices used by special education teachers score lower on an observation tool, that might mean teachers using these practices will score lower on their overall teacher evaluation ratings and receive more negative consequences than general education teachers. Such an outcome would undermine the worth of the scores. Our current study is a research study so there were no consequences for teachers, however, we compare the research-based FFT scores to FFT scores assigned by administrators and to overall teacher evaluation scores, which have consequences associated with them. These comparisons provide some speculative information about the potential consequences of FFT scores.

To summarize, as we have suggested, there are a number of concerns regarding policies that govern FFT's use with special education teachers. Validating FFT for use in the current policy environment will require measurement and functional evidence, or, information supporting the tool's truthfulness and worth. Therefore, the goal of this study is to investigate, through a series of analyses on classroom videos from approximately 50 special education teachers in Rhode Island, the validity evidence supporting FFT's use with special educators. To

address this goal, we consider three research questions. The first of these draws on the measurement perspective of validity; the second and third draw on the functional perspective.

1. To what degree does FFT provide accurate and reliable estimates of teaching quality among special education teachers?
2. To what degree does FFT provide similar information about teaching as compared with n observation system designed to measure teaching quality in special education classrooms?
3. To what degree do we see similar patterns in our teachers' FFT scores collected by administrators in formal teacher evaluation?

Method

The current study was designed to validate the FFT for identifying effective special educators in teacher evaluation systems. Specifically, we adopted Kane's validity argument approach to assess evidence supporting the proposed interpretation and use of FFT scores. Three research questions were posed. First, to what degree does FFT provide accurate and reliable estimates of teaching quality among special education teachers? We hypothesized that components in the Instruction domain of FFT would be problematic when scoring special educators.

Second, to what degree does FFT provide similar information about teaching as compared with a tool that is designed to measure teaching quality in special education classrooms? We hypothesized that FFT scores on components that contradict explicitness (i.e., 3B, 3C, 3D) would have the lowest correlations to QCI scores.

Third, to what degree do we see similar patterns in our teachers' FFT scores collected by administrators in formal teacher evaluation? It was expected that there would be weak to

moderate correlations between the two sets FFT rankings, for the similar reasons already laid out above.

Teacher Sample

To assess the questions in this study, we recruited a sample of elementary and middle school special educators (N =51) from Rhode Island. 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

As many scholars have suggested (Casabianca, Lockwood, & McCaffrey, 2015; Cash & Pianta, 2014; Mashburn, Meyer, Allen, & Pianta, 2014; Meyer, Cash, & Mashburn, 2011), careful consideration should be paid to how the sample of lessons selected for analysis map 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 timepoints too close to one another. In order to represent participating teachers' actual practice in the field, data collection mirrored the teaching assignments of the teachers in the sample. For example, if a teacher spent close to 75% of their time in a co-teaching setting and 25% of their time in a resource room setting, members of the research team conducted 3 co-teaching lesson observations and 1 resource room observation. This decision was made given that judgment about teaching quality is made at the teacher level. 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, and near testing windows to limit bias associated with those times of year.

Approximately 52% of the teachers in our sample are elementary teachers. About 54% teach reading or English language arts, and 45% teach mathematics. Seventy-three percent of the teachers are in resource room or self-contained classroom and about 17% are in a co-taught setting.

Measures

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

Framework for Teaching (FFT)

On the FFT, teachers receive ratings on each of eight dimensions, along a four-point scale which ranges from unsatisfactory (1) to distinguished (4). The framework is intended to reflect the range of instructional and other teaching responsibilities, containing 22 component criteria across four domains. These domains are: Planning and Preparation (Domain 1), Classroom Environment (Domain 2), Instruction (Domain 3), and Professional Responsibilities (Domain 4); each component is further broken down into elements. Many states, including Rhode Island, use a variation of FFT that includes only the Classroom Environment and Instruction domains. To reflect the evaluation system used in Rhode Island, ratings were only assigned across these two domains.

In line with existing conventions for creating observation scores (Bell, Dobbelaer, Klette, & Visscher, 2018), FFT ratings were applied to 15-minute segments of a lesson. For example, a 60-minute lesson would consist of 4 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, 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. The FFT rating of all 206 videos (with each double-scored) took place over 9 weeks and included 12 raters. An ETS statistician generated rater assignments using a design in which raters were paired, and each teacher was rated by 8 different raters. All lessons were double-scored, and random assignments were made with constraints on various variables to achieve a balanced distribution in terms of grade level, subject matter, and special education classroom settings (co-teaching, resource room, and substantially separate). One rater removed herself from the project during week five of scoring, and her remaining videos were reassigned amongst the other 11 raters.

Among the 11 raters who scored through the entire scoring period, each rater scored between 8 to 10 percent of the videotaped lessons. In order to ensure standardization of scores, raters scored a calibration video each week of scoring, the goal of which was to ensure that raters continued to score in ways similar to how they were trained. Calibration scores were used as quality control and to inform rater trainers to improve rater reliability, and reduce the chances of rater drift (Bell et al., 2012).

Scoring also included validation videos, which were used to enable rater trainers to monitor raters' scoring accuracy during the scoring process. Five pre-scored videos were embedded into each rater's assigned video queues, without being identified as validation videos. The order of validation videos was randomized across raters. Across all validation videos, rater

agreement with master ratings was exact 74.2% of the time, exact or adjacent 98.8% of the time, higher than master ratings 14.6% of the time, and lower than master ratings 11.2% of the time.

Quality of Classroom Instruction (QCI)

The Quality of Classroom Instruction (QCI) was used to assess the validity of scores for special education teachers on FFT. The QCI data around 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. Scores are assigned at the end of each lesson on a score of 1 (low quality) to 3 (high quality).

QCI Scoring Procedures. The rating of the QCI instrument took place over the course of 11 months. In all, 206 videos, 20% which were doubled coded, were completed by 6 different raters. Each rater was allotted an equal number of videos at the outset of coding assignments.[iv] The remaining videos were redistributed amongst the remaining raters. On average raters took 80 minutes per video to complete the QCI. Raters uploaded their completed codes sheets via a secure server and their scores were assessed by master raters to assure accuracy and completeness of observation forms prior to data entry.

As in our FFT training, raters completed calibration exercises for the purposes of quality control. Calibration videos were assigned either after raters completed 10 videos in their queue or every two weeks, and continued for the duration of their coding assignments. As a secondary check, we also used validation videos, where master raters randomly selected three coded videos from each rater, master rated the video, and compared agreement.

Rhode Island Teacher Evaluation Data

To compare FFT scoring patterns in our research context to the FFT scores created by teachers' administrators, we leveraged formal teacher evaluation data from the Rhode Island

Department of Education (RIDE). The state's Rhode Island Model Evaluation and Support System includes four measures: 1) Professional Practice: Classroom Environment (25% of total rating), 2) Professional Practice: Instruction (25%), 3) Professional Responsibilities (20%), and 4) Student Learning (30%). Teachers' overall effectiveness is rated on a four-point scale from ineffective to highly effective. Classroom observations make up 50% of an educator's evaluation, with the first two measures corresponding with the Classroom Environment and Instruction domains of FFT. Rhode Island requires a minimum of three observations, one announced and two unannounced; principals or assistant principals serve as the primary evaluator, but districts may elect to have complementary evaluators (e.g., those with specialized content knowledge) conduct some observations. The Professional Responsibilities domain consists of a portfolio of work-related tasks (e.g., lesson plans) that are assessed by the primary evaluator. Evidence of student learning is gathered either through student growth percentiles, which are calculated from the state assessment, or student learning objectives, created locally for untested grades and subject areas. Like other states, Rhode Island requires that teachers be formally evaluated at least once every three years, with more frequent evaluations for provisionally-certified teachers or those who receive low total effectiveness scores in the prior year. For our purposes, we include data from any year from 2015-16 through 2017-2018 to ensure that we include as many of our teachers as possible. Our analyses first examine whether the FFT scores we created are correlated with teachers' end-of-year ratings on Rhode Island's two Professional Practice domains. We also examine our FFT scores with teachers' student learning scores. As a final step, we examine whether, across the state, there are general patterns in which special educators are scoring in ways that differ from general educators. Here, we look overall and restrict the comparisons to teachers within the same schools.

Data Analysis Plan

To investigate the appropriateness of using FFT in special education, a first starting point is ensuring that raters are able to score lessons (i.e., operationalize the scales on the instrument) in a way that is psychometrically sound. A first step in our analysis was to look at how raters used the scoring scales, as well as whether the FFT factor structure conformed to theory. We examined the factor structure of FFT scores by using multi-level 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, keeping the within-teacher factor structure constant at two factors and modeling either one or two between-teacher factors to assess 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 four-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 (Brennan, 2001; Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Shavelson & Webb, 1991) for the FFT. We estimated the variance among teachers' average

value on the overall FFT score, lesson variability, segment variability, rater variability, 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 also did not include variance components related to specific aspects of teachers (e.g., grade level) or lessons (e.g., co-teaching vs. resource).

Then, to empirically assess the function of FFT scores (or whether they lend evidence to support intended uses of the scores), we included two additional sources of data. To address the question of whether FFT adequately reflects the true variation across special educators' lessons, we compared the means and distributions on FFT to a tool that has been established to more accurately reflect special educators' instructional quality, the QCI. Finally, we also investigated whether we would see similar patterns in our teachers' FFT scores collected by administrators in formal teacher evaluation. By expressly specifying our inferences and presenting our evidence along these claims, we can inform the question of the degree to which FFT scores can reasonably be used for the intended purposes (Kane, 2006).

Findings

To what degree does FFT provide accurate and reliable estimates of teaching quality among special education teachers?

Mean and Distribution of FFT Scores

A first strategy for assessing the appropriateness of 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 1, 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 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 FFT for capturing 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.

Factor Structure of FFT

A second step in assessing FFT from a measurement perspective is to examine whether FFT's factor structure conforms to theory. We would expect to see that the tool's factor structure – five components aligned with Classroom Environment and five with Instruction – holds when applied to our sample of special education lessons. Table 2 presents fit statistics for multilevel confirmatory factor analyses (CFA), run for one-factor and two-factor models. 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 two. We see that the fit indices are virtually identical across the one-factor and two-factor models. Both models have comparative fit indices (FCI) and Tucker-Lewis indices (TLI) of 0.89 and 0.86 respectively, with both failing to meet the >0.90 threshold of acceptable model fit. The AIC and 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. While two of the eigenvalues are greater than one, the first is by far the highest. Further, follow up analyses of the one-factor model show that all loadings are sufficiently high, ranging from 0.77 to 1.00. Additionally, the fit statistics are best for the one-factor model.

Accuracy and Reliability of FFT Scores

An additional way to assess 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, 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 the exact and adjacent agreement rates are in line with existing conventions for score accuracy. To assess rater consistency, inter-rater agreement was assessed across all lessons, which were each scored by two raters. In Table 3, 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 Classroom Environment averaged 75.4% across all components, and agreement rates ranged from 50.6% on *Culture* to 86.3% on *Space*. For Instruction, the average exact agreement was 52.0% and ranged from 47.0% on *Assessment* to 56.5% on *Responsiveness*. As with the analyses of the mean scores across the sample, the accuracy and consistency analyses

suggest that the Classroom Environment domain functions better than the Instruction domain, calling into question whether highly-trained raters could consistently apply scoring rules.

Finally, we assessed the extent to which raters' scores appear to be free from bias; that is, they do not exhibit patterns of scoring that reflect some aspect of the rater themselves (e.g., raters are systematically harsher in their scoring than others) or the teachers they are observing. 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 scores 14.6% of the time and lower than master scores 11.2% of the time. These results are not significant enough to warrant concern about this source of rater bias.

Generalizability of FFT Scores

We conducted a series of generalizability studies (g-studies) to understand teacher, lesson, and rater variability, and residual error, and interactions between each of these sources of variation. (Brennan, 2001; Cronbach, Gleser, Nanda, & Rajaratnam, 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 rater's severity. We estimated the variance components using standard random effect analysis of variance methods (Searle, Casella, & McCullough, 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 4. Our findings suggest that across all components, teacher-specific variation is relatively low, ranging from 2% to 15%. In line with previous studies using FFT, the residual variance (not explained by teachers, raters, or

lessons) is the largest variance component. The second highest variance component is the rater by lesson component. This too is in line with previous studies. It is not simply that some raters are more or less strict than each other. The variance is introduced at the level of the individual lesson when two raters observe the same instruction.

To what degree does FFT provide similar information about teaching as compared with an observation system 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 FFT functions with special educators. To test this, we compared FFT scores to 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 (see Table 1 for a summary of means and standard deviations for QCI). We summarize correlations between FFT Scores and QCI Scores in Table 5. In general, there are some positive, small correlations between the two measures. At the FFT domain 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 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 FFT's Instruction domain, despite considerable variation on the QCI. As we argue in the Discussion, this finding suggests that while FFT seems to rank order lessons in appropriate ways, it is likely limited in the extent to which it can provide appropriate feedback, given that it does not seem to reflect differences across lessons based on their use of high-quality special education teaching practices.

To what degree do we see similar patterns in our teachers' FFT scores collected by administrators in formal teacher evaluation?

In addressing this question, we wanted to better understand the extent to which the FFT scores created in the research context would be reflected in the data collected on teachers in the context of formal teacher evaluation. We addressed this question in two ways. First, we examined the correlation between FFT scores assigned by researchers, averaged across teachers' four lessons, with year-end evaluation scores on the Professional Practice domains. We used our overall FFT scores (averaged across Classroom Environment and Instruction) because the state collapsed scores across these two domains. A limited number of teachers in our sample had more than one year of teacher evaluation data; in these cases, we took the mean of their scores. Overall, the correlation between our research scores and the evaluation scores was 0.37, with 48 of our 51 teachers having evaluation scores available. We also summarize the scores in Figure 4. All of our teachers were rated between 3 (effective) and 4 (highly effective) in their formal evaluations, raising questions about whether the low observation scores we saw in research were likely to mirror what we saw in practice.

It could be the case that, while the special educators in our sample were scoring high overall on their formal teacher evaluation, there might be gaps between the performance of special educators and general educators in the state more generally. In other words, gaps might still be present if everyone in the state was shifted higher in the scoring distribution. If true, this could suggest that the tool was systematically biased against special educators or if the two populations – general educators and special educators – differed in other ways. Fortunately, we see no such gaps in the data across the state. From 2015-2016 to 2017-2018, there were 1,655 special education teachers and 10,058 general education teachers with evaluation scores. Ninety-eight percent of special education and general education teachers were assigned final evaluation ratings of highly effective or effective. The percentage of teachers within each group who received each rating is shown in Figure 4. Fisher's exact test did not provide evidence of a statistically significant relationship between teacher type and rating category ($p<.001$). That is, the final evaluation ratings were not dependent on whether someone was classified as a special education teacher or a general education teacher.

Discussion

In our study, we examined the degree to which a commonly used observation tool, the FFT, captures "atypical" instruction, focusing specifically on the case of special education. The examination of tools like the FFT, the most widely used observation system in the United States, is critical. Practitioners and policymakers are relying on information from observation systems like FFT to improve teaching and provide evaluative information about teachers. Most concerning is that universal use of these systems is predicated on the assumption that instruction provided in different contexts (e.g., special education, arts education) is sufficiently similar and that all students benefit from similar instruction (Gilmour & Jones, 2020).

Our study offers some of the first empirical evidence surrounding the trade-offs of applying systems developed for general education teachers to special education. We have organized our investigation around building two complementary forms of validity evidence (Kane & Wools, 2019). First, from a measurement perspective, we need confidence in FFT's technical properties. We need to know if scores are defensible in order to ensure stakeholders' trust in the evaluation and improvement processes. Second, from a functional perspective, we need to know if FFT provides useful information for achieving two specific outcomes: distinguishing between effective and ineffective teachers and helping teachers improve. In the case of teacher evaluation, where state and district policies mandate that observation scores drive decision-making, this latter source of evidence must be considered.

The psychometric properties of FFT in our study – including results from the reliability analyses, factor analysis and generalizability studies – suggest that the measurement properties of FFT in our study are in line with previous studies that assessed FFT's validity for general education teachers (Liu et al., 2019). The scores provide little reason to question the *truthfulness* of raters' scores. Our most striking finding, however, relates to mean lesson level scores in the instructional domain. These scores are almost universally low compared to 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. It seems unlikely special educators are universally worse at instruction than general education teachers and therefore, this finding raises questions about how well FFT scores capture special education teaching quality. It does not, however, undermine the validity of the scores from a strictly psychometric perspective. Other psychometric results also support use of FFT. As in previous studies of FFT (e.g., Liu et al., 2019), factor analyses support two factors – one for the learning

environment domain and one for the instructional domain. Exploratory factor analyses, however, support a single-factor model like we have seen in previous research uses of FFT (e.g., Liu et al., 2019). Similarly, results of our generalizability study align closely with previous studies that applied FFT to general education teachers (Kane & Staiger, 2012; Liu et al., 2019; Lockwood, Savitsky, & McCaffrey, 2015). The g-study findings point to high variation attributable to the residual and to the rater x lesson facets. In other words, the variation appears not to be between raters generally but as an interaction with the individual lesson, where two observers are less likely to agree on what they see.

But what about how the scores function for special educators, or what Cronbach (1988) would call their *worth*? Here, the evidence raises more serious concerns. We addressed this question by comparing FFT scores with two additional sources of information on our special educators' instruction – lesson-level scores on an observation system more aligned with principles of effective special education (QCI) and administrators' assessment of teachers' instructional quality on year-end summative teacher evaluation. Our findings suggest that FFT and QCI are only modestly correlated; correlations between FFT's Instruction domain and the QCI indicators ranged from 0.24 to 0.38. Lessons where teachers scored highly on QCI were also ones more highly rated on FFT's two domains. However, these correlations are not nearly as strong as we have seen in previous research on studies employing multiple observation tools. The MET study, for example, examined scores on five different observation tools (see Kane and Staiger (2012) for more information). FFT and CLASS (both general observation tools) were correlated at 0.88, and correlations between FFT and the other tools were similarly high.[1] We were more likely to see scores normally distributed on the QCI, whereas, scores on FFT were

[1] In the MET Study, FFT's lowest correlation was with the Mathematical Quality of Instruction at 0.67.

clustered on the low end of the scoring scale. These findings suggest that even though FFT appeared to do reasonably well 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 FFT's instruction domain that showed that FFT scoring criteria and descriptions do not fully capture effective special education instruction (Matthews et al., 2020). Our examination of state evaluation data, however, suggest that current uses of FFT by administrators do not show this same shift downward in scores on the instruction domain. For our sample of teachers, FFT scores were correlated with state evaluation scores at 0.04. Across the state, nearly all teachers (98%) were rated effective (3) or highly effective (4) on a 4-point scale, and we saw no systematic differences in how special educators and general educators scored.

Limitations

Two primary limitations of this study involve our sample of teachers and the sample of lessons within teachers. Our sample of 321 lessons across 51 teachers – although large for a special education sample – is likely at the low end of what would be acceptable to conduct confirmatory factor analysis. We also are limited in our ability to conduct generalizability studies that focus on more nuanced sources of variation, such as instructional setting by rater interactions. That said, no existing large-scale studies of FFT have included special educators, thus ours lends important empirical evidence surrounding its use.

Our sample is also a convenience sample, consisting of teacher volunteers from over twenty districts. On the one hand, this allows us to capture experiences of teachers across a broad

range of school contexts, but on the other, we do 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. While 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 in the state administrative data.

It is also worth questioning whether our sample of lessons adequately map onto the universe of lessons that special educators provide across a school year. As we see in our sample, it is typical that a special educator works across a variety of subjects and a variety of instructional settings. Because special education is by definition individualized instruction, it is common for their 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 of how a teacher's time was distributed; if a teacher reported spending 75% of their time co-teaching, we observed 3 out of 4 lessons in co-teaching). 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, the principals systematically chose to not observe co-teaching because of the challenges associated with tracking on two teachers' instruction.

One way our study would have been strengthened is if we were able to leverage student outcome data to inform the question of how FFT functions with special educators. Ideally, we would be able to examine whether differences in instructional quality (as measured by FFT) 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 teachers up for evaluation in that year. Further, because special educators often do not have an adequate number of students to reliably calculate student growth percentiles (Jones et al., 2013), many of our teachers had student learning scores based on district-derived student learning objectives. In short, we are limited in what we can say about whether our scores are associated with student learning gains. See Johnson and colleagues (2020) as an example of how future studies may make connections between observation tools and academic outcomes for SWDs.

Finally, we acknowledge that the FFT scoring patterns we see in our data are not directly applicable to 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, and received more rigorous training and ongoing monitoring activities than actual raters received in either of our two states. Plus, there were no stakes 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, suggesting that 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 offered little in the way of guidance for using 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 observation system is not functioning in the way that researchers or practitioners would desire.

Implications for Research and Policy

Our study provides initial validity evidence on using FFT in special education, and a rationale for the kind of research we need moving forward. Our study focused on the technical properties of FFT, but we still need a complementary line of research that examines how administrators use observation tools like FFT in practice. We find some underlying problems in how 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 how they use FFT in special education settings in ways that are beneficial, or demonstrate that they do not know the criteria they should use to differentiate between special education teaching quality? Or do principals score in ways that sacrifice accuracy in order 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 finally how teaching improves over time. There are examples of studies that do explore the validity of how principals assign scores. Research by Briggs, Dadey, and Kizil (2015), 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 tools and teachers' improvement (e.g., Hamre et al., 2012), though this work has not been done with principals as observers.

Our study also has an important implication for how observation tools like the FFT are used in practice. Can FFT support both human capital management and improvement purposes? If administrators were to apply FFT scores in ways consistent with what we see in the research context, the low scores would have negative consequences for special educator teacher evaluation. The teacher could be treated unfairly in the evaluation system because she would be at higher risk for earning a lower overall teacher evaluation score than should have been earned if the observation scores reflected best practices for special education. For a population of teachers that suffers from chronic shortages and high rates of attrition out of the profession or into general education (Boe, 2014; Dewey et al., 2017), the field of special education cannot sustain losses of its best teachers. Fortunately, for special educators, it does not appear that administrators in Rhode Island evaluated these teachers in ways that were consistent with how teachers were rated in our research context.

In the case of teacher improvement, the potential gap between how a teacher scores on FFT and their actual effectiveness at teaching SWDs could be more problematic. Rating a teacher lower than best practice standards could result in a special educator wasting time and resources learning to implement teaching practices that do not benefit their students, or worse, are harmful. Erroneously low scores could also lead to discouragement and frustration, particularly if teachers are told they are using inappropriate instructional practices. Whether districts rank order teachers or try to identify the specific practices that need improvement, the initial validity evidence from this study suggests that care is warranted in using the scores in the instructional domain.

Given the current policy environment – where states have begun to deemphasizes the human capital management goals of teacher evaluation and instead prioritize professional learning, the critical concern is whether general observation systems like FFT can be used to help special educators improve. Our study suggests that the answer is likely no. FFT scores did not reflect the variation in teachers' use of evidence-based practices for SWDs. It is hard to see how principals could use FFT to give teachers sufficiently targeted feedback; that is, unless they have background knowledge in special education or receive some other additional support. To better address special educators' needs, schools or districts could elect to train administrators to help them recognize differences in special education teaching quality. Or, they could allow special educators to be evaluated on adapted FFT scales in the instructional domain. Adapted scales could be specified for special education settings and based on effective special education instruction research. This would allow states and districts to keep using FFT, while making it more likely that its scales can drive improvement.

Although we chose to focus on special educators in this study, it is easy to imagine results like ours having implications for other "atypical" populations of teachers. For any one of these groups, an implication of our research is that districts might adopt entirely different observation systems more in line with what we know from each respective area of research. But this too presents challenges, including those related to cost and administrator capacity. And while a decision like this could be seen as an avenue for addressing unfairness, it could present its own issues related to equity, as some observation systems might be easier or harder to be rated highly by an administrator.

While separate instruments tailored to specific research bases and teaching demands is certainly an option for states, we close by recommending smaller steps that might help special

educators more immediately. First, districts could consider using alternative scoring criteria on (or foregoing altogether) one to two of the Instruction components where special educators consistently score low: *Using Questioning and Discussion Techniques* and *Using Assessment*. 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 small step would be for districts to train special education administrators or other leaders with certification in special education to conduct observations. At the very least, future investigations could productively train principals to recognize the use of effective special education practices and to provide feedback that is more likely to support teacher development. In the absence of such policies, our concern is that district policies will fail to address the goals of improving instruction for a critical subgroup of teachers, which in turn will only further deepen the long existent divide between general education and special education.

---

[i] See the Rhode Island Model Evaluation & Support System as one example: https://www.ride.ri.gov/Portals/0/Uploads/Documents/Teachers-and-Administrators-Excellent-Educators/Educator-Evaluation/Guidebooks-Forms/Teacher_Guidebook_2015-16.pdf

[ii] Some states such as New Jersey and Michigan allow districts or local education authorities (e.g., charter school authorizers) to select from a list of "validated" observation systems that have been approved by the state (Michigan Department of Education, 2017; New Jersey Department of Education, 2017). Other states such as Idaho and Rhode Island, mandate the use of a single observation system or that all observation systems are derivatives of a single observation system (Youde, 2017).

[iii] 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.

[iv] Two raters completed abbreviated assignments. One rater had increased responsibilities at her place of work and removed herself from the project during month three. A second rater requested a reduced number of videos due to unforeseen family matters during month four.

References

American Institutes for Research. (2012). Database on state teacher evaluation policies. Retrieved from http://resource. tqsource.org/stateevaldb/Default.aspx.

Archer, A., & Hughes, C. (2010). Explicit instruction— effective and efficient teaching. New York: Guilford.

Baker, S. K., Gersten, R., Haager, D., & Dingle, M. (2013). Teaching practice and the reading growth of first-grade English learners: Validation of an observation instrument. *The Elementary School Journal*, *107*(2), 199-220.

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.

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.

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). New York, NY: Routledge.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer.

Briggs, D.C., Dadey, N., & Kizil, R.C. (2015). Comparing student growth and teacher observation to principal judgments in the evaluation of teacher effectiveness. Report commissioned by the Georgia Department of Education. Boulder, CO: Center for Assessment, Design, Research and Evaluation (CADRE).

Casabianca, J. M., Lockwood, J. R., & McCaffrey, D. F. (2015). Trends in classroom observation scores. *Educational and Psychological Measurement*, *75*(2), 311-337.

Cash, A. H., & Pianta, R. C. (2014). The role of scheduling in observing teacher–child interactions. *School Psychology Review*, *43*(4), 428-449.

Cronbach, L. J. (1988). Five perspectives on validation argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3–17). Hillsdale, NJ: Erlbaum.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, M. (1972). *The Dependability of Behavioral Measurement: Theory of Generalizability for Scores and Profiles*. New York: John Wiiey and Sons.

Danielson, C. (1996). *Enhancing professional practice: A framework for teaching*. Alexandria, VA: ASCD.

Danielson, C. (2013). *The framework for teaching evaluation instrument, 2013 Edition*. Princeton, NJ: 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.

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.

Fuchs, D., Fuchs, L. S., & Stecker, P. M. (2010). The "blurring" of special education in a new continuum of general education placements and services. *Exceptional Children*, *76*(3), 301-323.

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.*

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.

Harris, D. N., & Sass, T. R. (2014). Skills, productivity and the evaluation of teacher performance. *Economics of Education Review*, *40*, 183-204.

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.

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

Hill, H. C., Kapitula, L., & Umland, K. (2011). A validity argument approach to evaluating teacher value-added scores. *American Educational Research Journal*, *48*(3), 794-831.

Hock, H., & Isenberg, E. (2017). Methods for accounting for co-teaching in value-added methods. *Statistics and Public Policy, 4,* 1-11.

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.

Johnson, E. S., Zheng, Y., Crawford, A. R., & Moylan, L. A. (2020). The Relationship of Special Education Teacher Performance on Observation Instruments with Student Outcomes. *Journal of Learning Disabilities*, 0022219420908906.

Jones, N. D., Buzick, H. M., & Turkan, S. (2013). Including students with disabilities and English learners in measures of educator effectiveness. *Educational Researcher*, *42*(4), 234-241.

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.

Kane, M. T. (2013). So much remains the same: Conception and status of validation in setting standards. In *Setting performance standards* (pp. 67-102). Routledge.

Kane, M. T. (2006). Validation. *Educational Measurement*, *4*(2), 17-64.

Kane, M. T., & Wools, S. (2019). Perspectives on the Validity of Classroom Assessments. *Classroom Assessment and Educational Measurement*, 11.

Kane, T. J., & Staiger, D. O. (2012). Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains. Research Paper. MET Project. *Bill & Melinda Gates Foundation*.

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.

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.

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

Mashburn, A. J., Meyer, J. P., Allen, J. P., & Pianta, R. C. (2014). The effect of observation length and presentation order on the reliability and validity of an observational measure of teaching quality. *Educational and Psychological Measurement*, *74*(3), 400-422.

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.

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.*

McCaffrey, D. F., Yuan, K., Savitsky, T. D., Lockwood, J. R., & Edelen, M. O. (2015). Uncovering multivariate structure in classroom observations in the presence of rater errors. *Educational Measurement: Issues and Practice*, *34*(2), 34-46.

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.

Patrick, H., Mantzicopoulos, P., Samarapungavan, A., & French, B. F. (2008). Patterns of young children's motivation for science and teacher-child relationships. *The Journal of Experimental Education*, *76*(2), 121-144.

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.

Rhode Island Department of Education (2017). Rhode Island Model Evaluation & Support System Guidebook – Teacher. Retrieved from: http://www.ride.ri.gov/Portals/0/Uploads/Documents/Teachers-and-Administrators-

Excellent-Educators/Educator-Evaluation/Guidebooks-Forms/Teacher_Guidebook_2015-16.pdf.

Searle, S. R., Casella, G., & McCulloch, C. E. (2009). *Variance components* (Vol. 391). John Wiley & Sons.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer* (Vol. 1). Sage.

Sledge, A., & Pazey, B. L. (2013). Measuring teacher effectiveness through meaningful evaluation: Can reform models apply to general education and special education teachers? *Teacher Education and Special Education*, *36*(3), 231-246.

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.

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.

Swanson, H. L., & Hoskyn, M. (2001). Instructing adolescents with learning disabilities: A component and composite analysis. *Learning Disabilities Research & Practice*, *16*(2), 109-119.

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*, 364–380. doi:10.1177/00222194050380041401

Vaughn, S., Gersten, R., & Chard, D. (2000). The underlying message in LD intervention: Findings from research syntheses. *Exceptional Children, 67,* 99-114.

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.

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

Tables and Figures

Table 1
Means and Standard Deviations of FFT and QCI scores

Dimension	N	Mean	SD
FFT			
FFT Domain 2: Environment	206	2.73	0.24
FFT2a: Respect	206	2.84	0.32
FFT2b: Culture	206	2.26	0.39
FFT2c: Procedures	206	2.79	0.35
FFT2d: Behavior	206	2.85	0.29
FFT2e: Space	206	2.94	0.17
FFT Domain 3: Instruction	206	2.07	0.34
FFT3a: Communication	206	2.29	0.39
FFT3b: Questioning	206	1.82	0.45
FFT3c: Engagement	206	2.14	0.40
FFT3d: Assessment	206	1.94	0.44
FFT3e: Responsiveness	206	2.16	0.35
FFT Overall Mean	206	2.40	0.27
QCI			
QCI Principle 1	237	2.23	0.70
QCI Principle 2	237	2.33	0.64
QCI Principle 3	235	2.25	0.75
QCI Principle 4	237	2.16	0.68
QCI Principle 5	237	2.20	0.72
QCI Principle 6	237	2.30	0.70
QCI Principle 7	219	2.19	0.74
QCI Principle 8	237	2.18	0.70
QCI Overall	237	2.16	0.69

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

Figure 1
Frequency of Assigned FFT Scores on Domains 2 and 3 (Lesson Level)

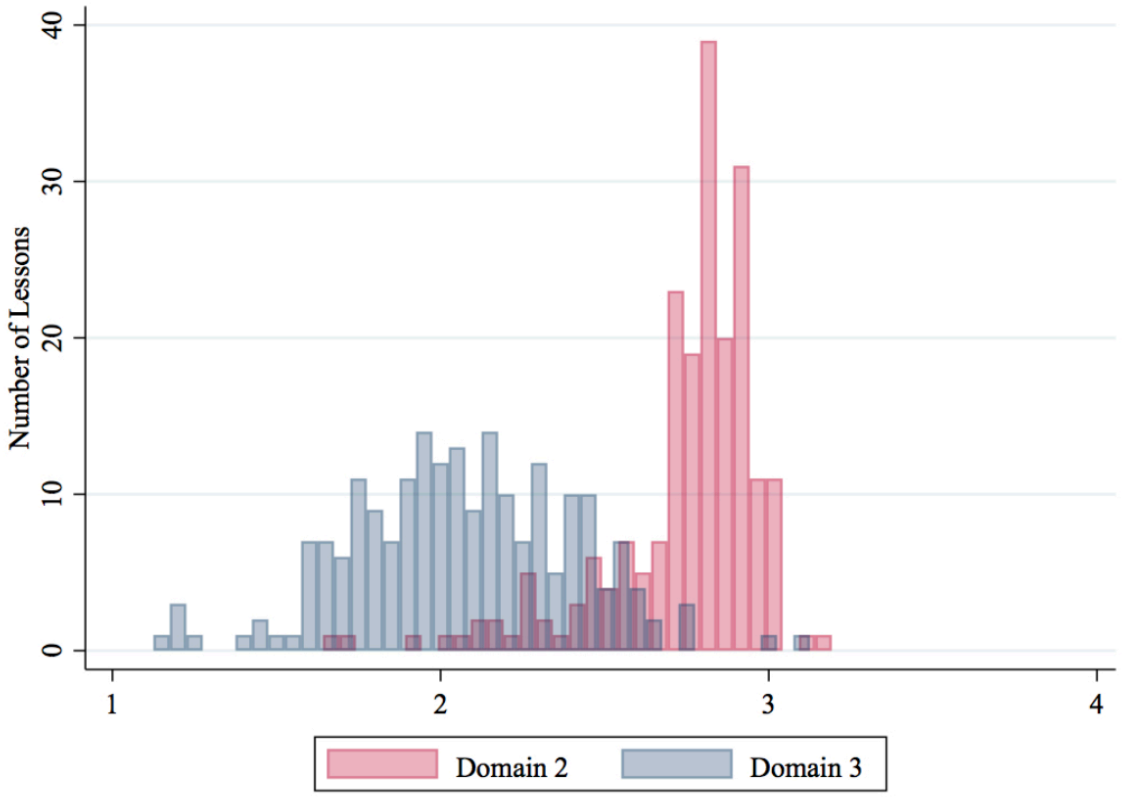


Table 2
Reliability Estimates for FFT

Dimension	Double-scored Segment Sample Size	Percent Exact Agreement	Percent Exact or Adjacent Agreement	Simple Kappa	Linear Weighted Kappa	Quadratic Weighted Kappa
FFT2a	526	79.09%	99.62%	0.34	0.39	0.45
FFT2b	526	50.57%	96.58%	0.07	0.09	0.14
FFT2c	526	76.62%	98.10%	0.32	0.36	0.42
FFT2d	526	84.60%	99.05%	0.42	0.43	0.45
FFT2e	526	86.31%	99.62%	0.05	0.06	0.06
FFT3a	526	49.43%	97.53%	0.04	0.08	0.15
FFT3b	526	52.47%	96.39%	0.15	0.20	0.26
FFT3c	526	54.75%	98.29%	0.09	0.14	0.22
FFT3d	526	46.96%	97.15%	0.04	0.11	0.21
FFT3e	526	56.46%	97.34%	0.08	0.10	0.13

Table 3
Fit Statistics for Multilevel CFA Models Fit to 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

Notes: Both models were run with 2 within-teacher factors. Acceptable model fit is indicated by values of CFI (comparative fit index) and TLI (Tucker-Lewis index) >.90; RMSEA (root mean square error of approximation) and SRMR (standardized root mean square residual) <.05 (Hu & Bentler, 1999).

Figure 2
FFT Multilevel CFA Between-Teacher Eigenvalues

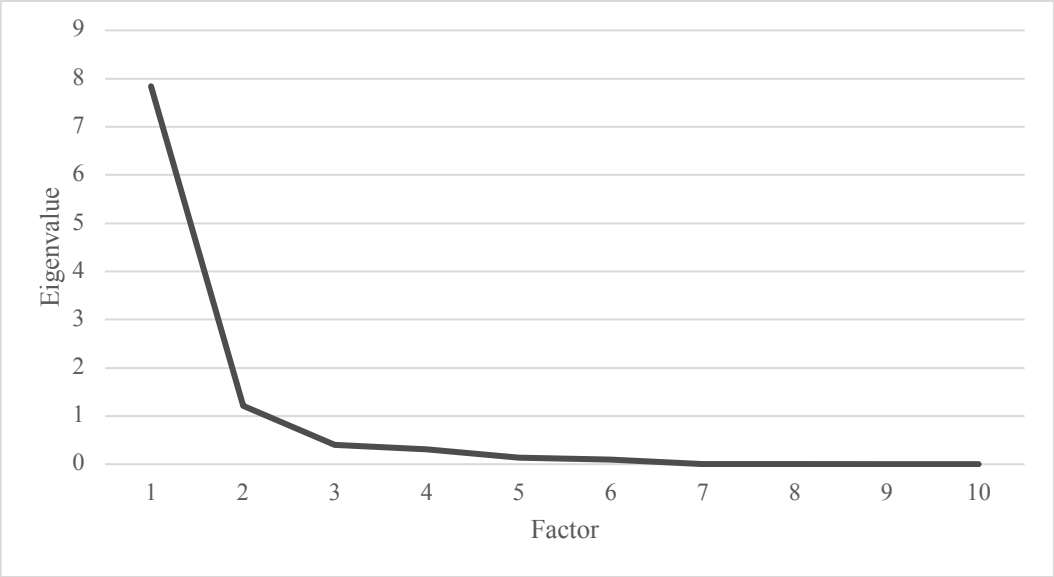


Table 4
Generalizability Study Results: Percent Variance by Source

FFT Component	2A	2B	2C	2D	2E	3A	3B	3C	3D	3E
Teacher	19%	14%	13%	16%	2%	15%	14%	14%	10%	11%
Lesson x Teacher	14%	3%	14%	18%	10%	5%	8%	6%	6%	3%
Observer	2%	16%	2%	1%	5%	13%	10%	7%	17%	11%
Observer x Lesson	19%	31%	26%	22%	40%	29%	27%	30%	30%	30%
Segment x Lesson	11%	3%	13%	12%	0%	4%	7%	4%	7%	1%
Residual	35%	32%	33%	30%	44%	34%	34%	39%	30%	44%

Table 5
Pearson Correlations between FFT Scores and QCI Scores

	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**

Figure 3
Comparing Lesson-Level Means on FFT Domain 3 and the QCI

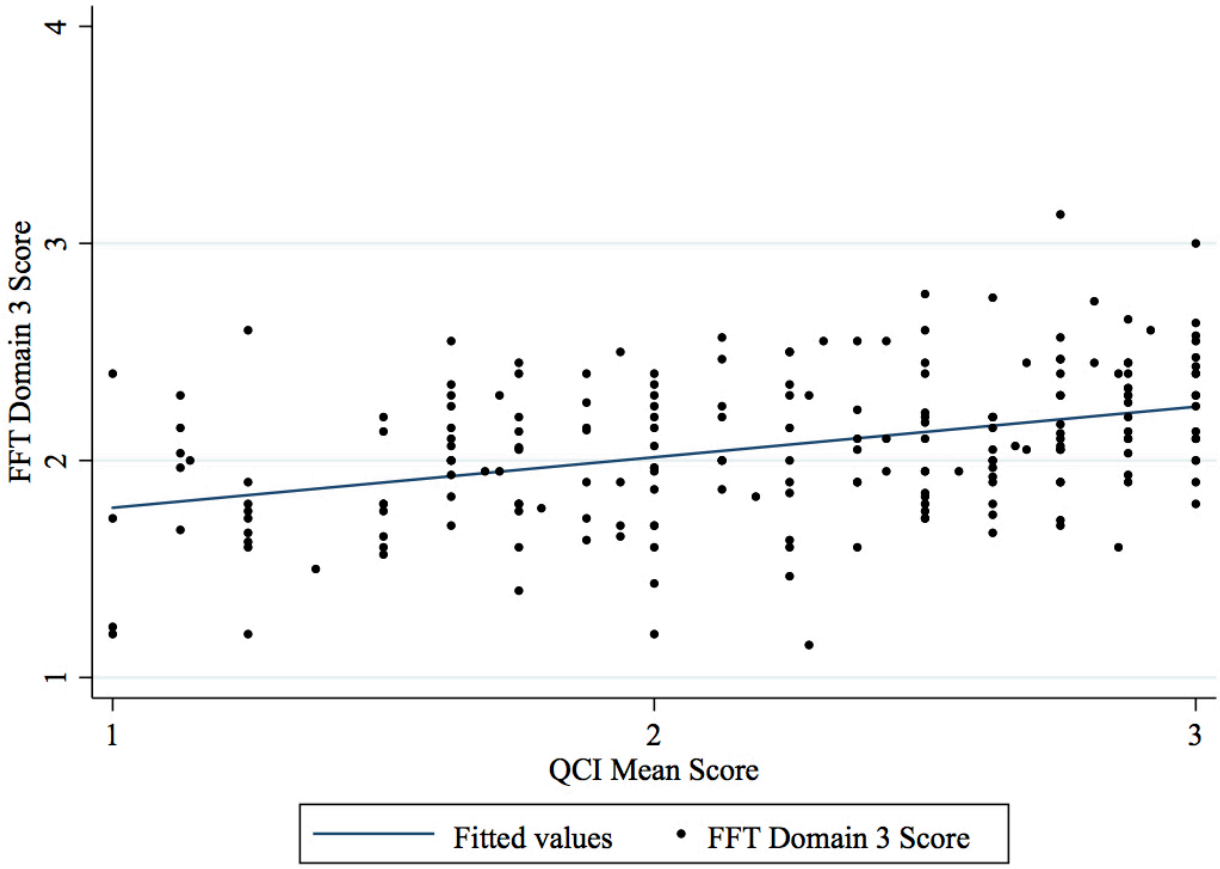
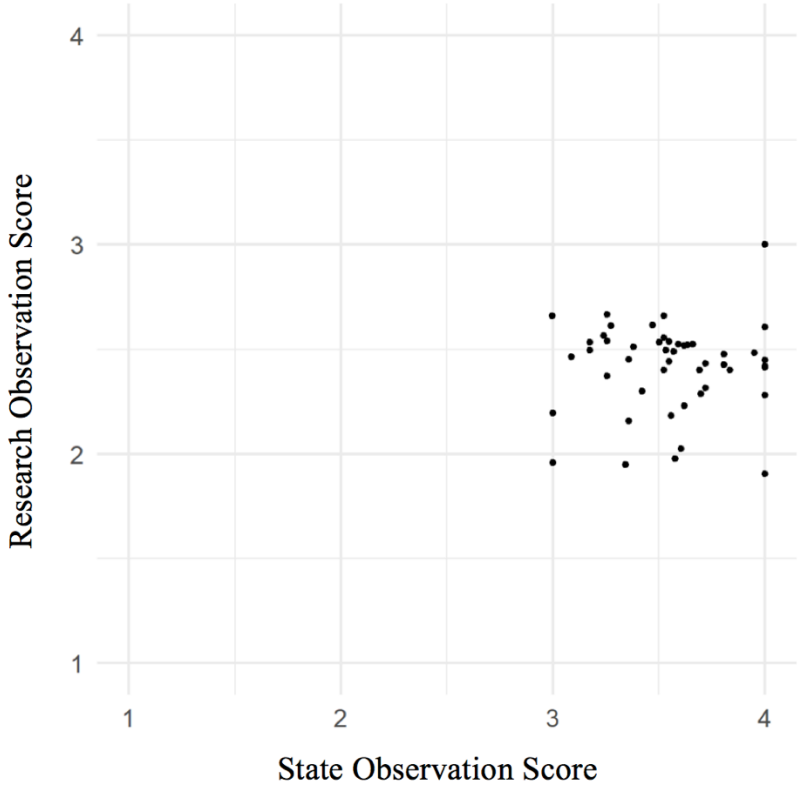
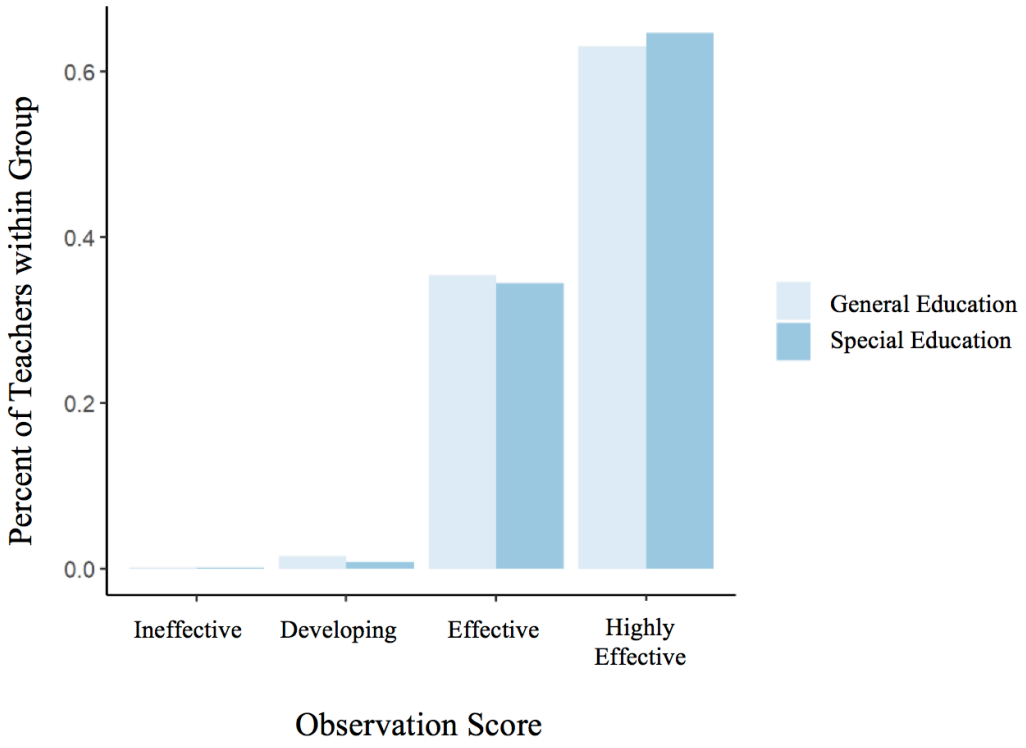


Figure 4
Comparing Special Educators Research and State Evaluation Observation Scores



Note: N=48

Figure 5
Comparing General Education and Special Education Teachers' Observation Ratings Across Rhode Island



Note:
Special Education N=1,655
General Education N=10,058

Appendix

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)

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