Hammill Institute on Disabilities

Assessment for Effective Intervention
2019, Vol. 44(2) 69–80

DOI: 10.1177/1534508417745628
aei.sagepub.com
SAGE

Using Observational Assessment to Inform Professional Development Decisions: Alternative Scoring for the Danielson Framework for Teaching

Ryan J. Kettler, PhD[1] and Linda A. Reddy, PhD[1]

Abstract

The Framework for Teaching (FFT) is one of the most widely used observational systems for evaluating teacher effectiveness and driving professional development conversations in schools. This study contrasts reliability and validity evidence relevant to the FFT as traditionally scored with evidence relevant to a composite scoring approach that connects to specific practice feedback. The FFT is typically interpreted at the domain level and at an overall total level using four categories of teacher effectiveness—unsatisfactory, basic, proficient, and distinguished—scored without computational rules governing relationships between the 22 components and four domains. For this study, the composite scoring approach was computed by averaging the components nested within each domain. A sample of 156 teachers and 34 trained school administrators from 12 high-poverty charter schools used the FFT as part of regular evaluation practices, yielding an extant set of de-identified data. Results indicate the composite scores were internally consistent at the domain and total levels. In comparison with traditional scores, the composite scores were more stable across time, as well as more predictive of student growth in reading and mathematics achievement. Implications for professional development and educator evaluation are discussed.

Keywords

instruction, observation, rating scales

Since the No Child Left Behind Act of 2001, large-scale achievement tests in reading and mathematics have been used in Grades 3 through 8, as well as in one grade level in high school, for the purpose of evaluating the effectiveness of teachers, schools, and districts. Using scores from such tests to evaluate teaching is sensible, because the primary purpose of schools is to provide learning, which can be measured through student growth in achievement (i.e., the change in achievement scores from the beginning of a school year until the end). However, best practices in evaluation require focusing on more than solely the primary outcomes of a program; it is equally important to focus on the key processes that lead to the outcomes (McLaughlin & Jordan, 2010). Solely evaluating teachers based on outcomes ignores that there are many factors outside of their control, including variability in student ability, support within and outside school, and school and family stressors, that influence growth in student achievement. For decades, it has been recognized that *how* educators deliver instruction and manage their classrooms profoundly influences student academic engagement and learning (Bennet, 1988; Hattie, 1992; Wang, 1991). Accurately measuring such processes is a focus of the current study.

School districts commonly measure and monitor teaching practices through the use of classroom observation approaches. School administrators gather data during structured and unstructured teaching periods, with varying lengths of time to guide decisions for overall teaching effectiveness, and identify practice strengths and areas for improvement. While there are many different classroom observation tools available, relatively few offer quantitative data that meet acceptable psychometric benchmarks to support inferences and use for effective professional development planning (e.g., Marzano, Frontier, & Livingston, 2011). To evaluate the teaching process, one needs measurement tools that yield reliable scores from which valid inferences about performance may be drawn, so scores attained from such measures are useful. One widely used classroom observational framework, the Framework for

[1]Rutgers, The State University of New Jersey, Piscataway, USA

Corresponding Author:
Ryan J. Kettler, Graduate School of Applied and Professional Psychology, Rutgers, The State University of New Jersey, 152 Frelinghuysen Road, Piscataway, NJ 08854, USA.
Email: r.j.kettler@rutgers.edu

Table 1. Components of the Framework for Teaching by Domain.

Domain		Components
1. Planning and Preparation	1a.	Demonstrating knowledge of content and pedagogy
	1b.	Demonstrating knowledge of students
	1c.	Setting instructional outcomes
	1d.	Demonstrating knowledge of resources
	1e.	Designing coherent instruction
	1f.	Designing student assessments
2. Classroom Environment	2a.	Creating an environment of respect and rapport
	2b.	Establishing a culture for learning
	2c.	Managing classroom procedures
	2d.	Managing student behavior
	2e.	Organizing physical space
3. Instruction	3a.	Communicating with students
	3b.	Using questioning and discussion techniques
	3c.	Engaging students in learning
	3e.	Using assessment in instruction
	3f.	Demonstrating flexibility and responsiveness
4. Professional Responsibilities	4a.	Reflecting on teaching
	4b.	Maintaining accurate records
	4c.	Communicating with families
	4d.	Participating in the professional community
	4e.	Growing and developing professionally
	4f.	Showing professionalism

Note. Source: Danielson, 2013.

Teaching (FFT; Danielson, 2013), has yielded some evidence of scores from which inferences can be made about teacher practices. The current study builds on the previous literature by examining the FFT's traditional domain scoring approach to a composite scoring approach that may be more consistent and connect to more specific teacher practice information, which may in turn enhance professional development planning.

FFT

The FFT is grounded in a research-based set of components of instruction, the Interstate Teachers Assessment and Support Consortium (InTASC) Model Core Teaching Standards (Council of Chief State School Officers, 2011), and a constructivist view of learning and teaching. The framework is designed to apply equally to instruction across content areas covered in school. The FFT employs observations, typically by well-trained school administrators or other supervisors, to evaluate teacher performance. The FFT is composed of 76 elements of effective practice, organized into 22 components within four domains, as depicted in Table 1. The domains include *Planning and Preparing for Student Learning* (Domain 1), *Creating an Environment for Student Learning* (Domain 2), *Teaching for Student Learning* (Domain 3), and *Professional Responsibilities* (Domain 4). In each domain, teachers are rated with one of four effectiveness levels: (1) *unsatisfactory*, (2) *basic*, (3) *proficient*, or (4) *distinguished*. The ratings are based on observations and a portfolio review. The framework is designed to be flexible to fit the needs and preferences of the districts in which it is implemented. Details such as the qualifications needed to be an observer, the number of observations to be used, and the evidence necessary to constitute success may vary by school district. When used properly, the framework promotes performance-based assessment by including guidance for making these decisions, such as a process for training and certifying the evaluators.

The FFT is highly structured and hierarchical. Each domain includes elements and indicators nested within components, as well as a rubric with a description, critical attributes, and possible examples for each level. For example, Component 2a within Domain 2: Classroom Environment includes the elements "Teacher interactions with students, including both words and actions" and "Student interactions with other students, including both words and actions." Definitions of components and elements are included on the FFT. Indicators for Component 2a include "Respectful talk, active listening, and turntaking;" "fairness;" and "politeness and encouragement." The rubric includes multiple critical attributes and possible examples for each level of each component. For example, a proficient (Level 3) rating on Component 2a includes the

critical attribute "Students exhibit respect for the teacher" and the possible example "Students help each other and accept help from each other."

Although the FFT provides guidance for rating each component using a four-level rubric, there is no requirement that the component scores be mathematically linked to the domain scores, nor that the domain scores be mathematically linked to an overall total score. Evaluators are allowed to provide ratings for component, domain, and total scores that are independent of each other. As such, each component, domain, and total score is interpretable as its own one-item score. This traditional domain scoring approach constitutes a significant measurement issue, as evaluators may make inconsistent and invalid inferences across the four domains and 22 components. Use of a nonmathematical scoring approach may reduce the reliability of FFT scores, and result in evaluators providing contradictory and broad feedback during postobservational meetings with educators, rather than specific performance-based feedback.

Alternatively, a systematic mathematical scoring approach may enhance the use of FFT for formative and summative assessments of teaching practices and advance professional development decisions. This study aims to examine the reliability and validity evidence relevant to a composite scoring approach compared with the traditional domain scoring method. To this end, we propose to sum the 22 components as items nesting within each of their corresponding domains and to compute a total score based on the 22 component scores. It is important that an observational assessment such as the FFT is designed to generate scores that accurately evaluate the teaching process and inform meaningful plans for professional improvement. Error in the scores of a tool could lead to inaccurate feedback on instruction or poor human capital management decisions; studies such as the current one are aimed to assess the level of error that may arise from routine evaluation practices. In doing so, the study addresses a gap in the literature on reliability and validity evidence relevant to the FFT, scored using both a mathematical composite method and a traditional method.

Previous Research on the FFT

Several large-scale research studies have been conducted involving the FFT. The FFT was used in the *Measures of Effective Teaching* (MET) Project, a partnership of seven school districts and 21 research entities that used video-based observations. The FFT was evaluated along with four other observational instruments, each designed to focus on specific aspects of teaching practice, as well as to establish common standards for levels of practice (Kane & Staiger, 2012). The subsample for this comparison included 1,333 fourth- through eighth-grade teachers. The reliability of the measures and their relations to student variables were both addressed. Reliability was addressed by decomposing the total variance into that which was explained by teacher, section, lesson, rater, and residual; the implied reliability for the FFT after use for four lessons was .67, meaning that 67% of the variance would be explained by the teachers, if all of the other factors were held constant. Disattenuated correlations between the FFT total score and the total scores from the other observational tools were between the large range (r = .67) and the nearly perfect range (r = .93). Correlations between FFT total scores and growth in English/language arts and mathematics during the previous and current academic years were between the nonexistent range (r = .05) and the small range (r = .19). Differences on the FFT were found between the top quartile and the bottom quartile on a number of other student variables (e.g., student effort, positive emotional attachment); the differences were positive but less than one quarter of 1 *SD* in all cases. In sum, findings offer evidence for the reliability of FFT scores, as well as strong validity evidence based on relations to other observational tools and limited validity evidence based on relations to achievement.

Kane, Taylor, Tyler, and Wooten (2011) analyzed data from Cincinnati's teacher evaluation system from academic year 2003–2004 through 2008–2009, using FFT Domain 2: Classroom Environment and Domain 3: Instruction, based on the rationale that these are the two domains focused on classroom practices. The purpose of the study was to estimate the relationship between classroom practices and student achievement gains on large-scale examinations. The study focused on Grades 3 through 8 due to the availability of large-scale achievement test scores. The researchers determined that 45% of the variation in scores was due to the teacher, 23% was due to the evaluator, and 32% was residual. Depending on the degree to which one assumes students are randomly assigned to teachers, the difference between the top quartile and the bottom quartile on the FFT was estimated to account for between about one tenth of an *SD* and one half of an *SD* in growth in reading and mathematics scores, a relationship that translates to at least 2 to 3 percentile points. The researchers also estimated the correlations between FFT total scores and growth in reading and mathematics achievement for three subsamples defined by the degree to which teachers had participated in the evaluation system. All coefficients were in the small range (r = .13–.15). Similar to the findings from the MET Project (Kane & Staiger, 2012), FFT score relations to student growth in achievement were weak.

Sartain et al. (2011) summarized findings from the 2-year *Chicago's Excellence in Teaching Pilot*, a project designed to improve teacher evaluation systems and instruction. The pilot featured Domain 2: Classroom Environment and Domain 3: Instruction of the FFT used in 44 elementary schools during the first year and 101 elementary schools during the second year. Validity analyses were based on 955 observations of 501 teachers, using evidence of relations to value-added growth in English/language arts and in

mathematics. By including scores from the FFT, value-added growth in achievement was improved above the average amount predicted based on student demographics, daily attendance, and mobility. Across almost all 10 components of the FFT Domain 2: Classroom Environment and Domain 3: Instruction, as well as both content areas, significant differences were found in the direction of teachers rated *basic* being associated with more growth than teachers rated *unsatisfactory*, those rated *proficient* associated with more growth than those rated *basic*, and those rated *distinguished* associated with more growth than those rated *proficient*. Results from this investigation highlight that teachers who perform better on the FFT tend to have students who improve more in achievement.

Lash, Tran, and Huang (2016) evaluated the FFT in Washoe County School District in Reno, Nevada, using domain scores mathematically calculated from the means of their components. The sample included 713 elementary, middle, and high school teachers rated by their principals. The researchers found Cronbach's alphas ranging from .83 to .87, with an estimate of .95 for the 22 components of the total score. Correlations with student growth in achievement were between the small range (r = .18) and the medium range (r = .48) across domains and content areas. The total score correlated with reading growth in the small range (r = .29) and with mathematics growth in the medium range (r = .46). The current study extends this research on composite scores from the FFT using additional forms of reliability (e.g., stability) and validity (e.g., internal structure) evidence.

Multiple investigations have examined the reliability of FFT scores and the validity of ensuing inferences about teaching performance. Results are mixed, and application of FFT scoring has varied. Additional research on the FFT is urgently needed, given its use in schools worldwide and potential impact on evaluating teaching effectiveness, as well as on monitoring changes in teacher practices following professional development.

Research Questions

This investigation builds on the aforementioned literature by further examining the FFT's reliability and validity evidence. The study compares the traditional FFT domain scoring to a new composite scoring approach in urban high-poverty schools. Determining the best method of calculating FFT scores will allow school personnel to make the most accurate decisions about instructional feedback and evaluation. Research questions include the following:

Research Question 1: What is the internal consistency of the FFT using the proposed composite scoring approach?
Research Question 2: What is the stability of FFT for the composite scores versus traditional scores?
Research Question 3: What is the relationship between composite scores and traditional scores for four domains and the total?
Research Question 4: Is there evidence of internal structure validity for the FFT composite scores and traditional scores?
Research Question 5: How related are composite scores and traditional scores to student growth in achievement?

Method

This study involved evaluating reliability and validity evidence for the FFT using indices of classical test theory. The *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Testing, 2014) suggest collecting multiple forms of evidence to support the validity of inferences drawn from test scores. The current study addressed the reliability of scores based on internal consistency and stability, as well as evidence for their ensuing inferences based on internal structure and relations to other variables. All study procedures were approved through the authors' Institutional Review Board.

Participants

Participants in the study included 156 teachers from 12 high-poverty charter schools located across one mid-Atlantic state. The convenience sample was drawn from the U.S. Department of Education–funded School System Improvement (SSI) Project. The SSI Project is a comprehensive school reform grant focused on enhancing educator evaluation and professional development systems that lead to student growth in achievement in urban high-poverty schools. Schools qualified for the SSI Project by having at least 50% of their students eligible for either free or reduced lunch. The teacher sample was predominantly female (75%) and European American (66%), and included representation from African American teachers (22%). The most common highest degree earned was a bachelor's (66%) degree, with a representative subsample having earned a master's (30%) degree. The most common grade band was Elementary (63%), though many teachers taught across grades, as is common in charter schools. The sample included nearly equal representation from reading (51%) and mathematics (49%) teachers. Table 2 provides demographic information for the sample.

The FFT observations were conducted by trained school administers (n = 34) as part of their routine educator evaluation process. The administrator sample was predominantly female (75%) and European American (62%), and included representation from African American administrators

Table 2. Demographic Information for the Teacher Sample.

Variable	Value	*n*	%
Gender	Female	111	75
	Male	37	25
Ethnicity	Asian American	5	3
	African American	33	22
	Latino/Latina American	13	9
	European American	97	66
	Native American	1	1
	Other	14	9
Highest degree	Bachelor	98	66
	Master	44	30
	Doctorate	6	4
Grade band	Elementary	93	63
	Middle	52	35
	High	46	31
Content area	Reading	77	51
	Mathematics	73	49
Total		156	100

Note. The totals may exceed 100% for variables in which some teachers selected more than one category. Demographic information was not reported by eight teachers.

(29%). Most administrators had a master's (82%) degree as the highest degree earned. The mean amount of administrative experience was 2.24 years (SD = 4.58).

Procedures

Teachers in each of the schools were evaluated using the FFT based on observations conducted 3 times (i.e., rounds) per year (i.e., fall, winter, and spring). Students in each of the schools completed achievement testing in the fall and in the spring, and teachers were assigned achievement growth scores that were equal to the means of their students' growth scores. The data were originally used for teacher evaluation. For this study, the data were de-identified and used as an extant set.

Measures

Scores from the FFT were evaluated and used as predictor variables in the current study. The Measures of Academic Progress (MAP; Northwest Evaluation Association [NWEA], 2011) were used as criterion variables to evaluate the relationship between the FFT and student achievement.

FFT. Observer training and calibration for the FFT (Danielson, 2013) include a 6-day intensive workshop addressing the following topics: (a) theoretical background of the FFT, (b) training on how to use the framework for observing teachers, (c) critical observer skills and competencies, and (d) skills necessary to successfully train future observers. In addition to undergoing the workshop, observers for this study participated in the Teaching Proficiency System, an online training and certification program for the FFT. Observers completed 7.5 hr of online training modules and practice coding videos before undergoing an observer certification test that assessed their accuracy in rating teachers using the FFT. Some observers were from schools that used the FFT prior to the SSI Project, so that no additional training was necessary. Observers from schools adopting the FFT as part of the SSI Project were trained as part of induction into the project. Observations were conducted for full-class periods for all three rounds.

For each observation, administrators provided scores for Domain 1: Planning and Preparation, Domain 2: Classroom Environment, Domain 3: Instruction, and Domain 4: Professional Responsibilities, as well as for a total score. In the current study, scores for Domains 2 and 3 were also summed to create a Domain 2 + 3 score, because these domains were specifically used to represent observable classroom practices in previous research (Kane et al., 2011; Sartainet al., 2011). Administrators also provided scores for each component nested within the four domains. Composite scores were calculated by averaging the component scores within each domain, within Domains 2 and 3 for the Domain 2 + 3, and within the full instrument for the Total score.

As noted, reliability and validity evidence for the FFT scores and inferences is promising, though more psychometric evidence is needed, and the current project addresses that need. Lash et al. (2016) found Cronbach's alpha exceeding .80 for Domain scores and at .95 for the Total score. Score variance has been found more attributable to teacher than to other variables (Kane & Staiger, 2012; Kane et al., 2011). Relationships with other observational measures have been in the large range or greater, while relationships with achievement growth scores have varied greatly in magnitude (Kane & Staiger, 2012; Lash et al., 2016; Sartain et al., 2011).

MAP. The MAP assessment system (NWEA, 2011) includes computer-adaptive multiple-choice tests that assess achievement in reading and mathematics for students in Grades K through 12. Each test yields a Growth Index Average that reflects change in achievement from one point in time to another (e.g., fall testing–spring testing). Administration times for the MAP typically range from 15 to 60 min per student depending on grade level. Marginal reliabilities for MAP have been estimated to range from .94 to .97 across grades and content areas (Cronin, 2005). Correlations between MAP scores and scores from large-scale achievement tests in like content areas have historically been in the very large range (i.e., r = .70–.90). Wang, McCall, Jiao, and Harris (2013) found internal structure validity evidence supportive of MAP using confirmatory factor analysis.

Data Analyses

For the majority of research questions in the current study, analyses were conducted using the FFT scores from a single

Table 3. Means and *SD*s of Ratings by Domain for Round 1 and across Rounds.

Domain	Number of Items	Composite Scores			Traditional Scores		
		n	*M*	*SD*	*n*	*M*	*SD*
	Round 1						
1. Planning and Preparation	6	110	3.01	0.28	110	2.99	0.35
2. Classroom Environment	5	150	2.98	0.53	150	2.95	0.64
3. Instruction	5	150	3.05	0.48	149	3.07	0.57
4. Professional Responsibilities	6	65	3.03	0.28	65	3.03	0.30
Domain 2 + Domain 3	10	150	3.01	0.47	149	3.01	0.55
Total	22	63	3.03	0.24	142	3.00	0.61
	Across rounds						
1. Planning and Preparation	6	96	3.11	0.31	52	3.10	0.36
2. Classroom Environment	5	150	3.06	0.43	152	3.05	0.49
3. Instruction	5	150	3.09	0.45	154	3.13	0.44
4. Professional Responsibilities	6	52	3.09	0.37	154	3.09	0.47
Domain 2 + Domain 3	10	150	3.10	0.36	106	3.09	0.44
Total	22	52	3.09	0.37	150	3.11	0.47

Note. Domain 2 + Domain 3 is the mean of Domain 2 and Domain 3. Ratings are on the following scale: 1 = *unsatisfactory*, 2 = *basic*, 3 = *proficient*, 4 = *distinguished.*

round (i.e., Round 1) to reflect the scores useful for professional development, as well as using the FFT scores averaged across Rounds 1, 2, and 3 to resemble scores for teacher evaluation. The exception was Question 2, which addresses the stability of scores across rounds. Internal consistency was estimated using Cronbach's alpha, an indicator of the homogeny of a set of items that constitute a scale. Alphas in the .70s can be considered moderate to fair, in the .80s can be considered moderately high to good, and in the .90s can be considered high to excellent (Murphy & Davidshofer, 2005). Score stability was evaluated using Pearson correlations between Rounds 1 and 2, between Rounds 1 and 3, and between Rounds 2 and 3. Stability was calculated based on means of component scores (i.e., composite scores) and for scores assigned broadly on a 4-point scale at the domain and total levels (i.e., traditional scores). Interpretive standards for score stability depend greatly on the amount of time between observations, in this case about 3 months. Pearson correlations were also used to examine the relations between composite scores and traditional scores, as well as the relations within the FFT featuring scores from both methods, and the relations of both types of scores to MAP scores. Correlations with magnitudes in the .00s may be considered nonexistent, .10s and .20s may be considered small, .30s and .40s may be considered medium, .50s and .60s may be considered large, .70s and .80s may be considered very large, and .90s may be considered nearly perfect (Cohen, 1992; Hopkins, 2001).

Results

Overall, teachers in this study were found on average to perform near the proficient level (FFT mean rating of 3), and taught students who were near the national average in reading and mathematics achievement. Including both the first round and the average across the three rounds, Domain and Total scores, and composite scores as well as traditional scores, the majority of the mean FFT ratings represented in Table 3 were near 3 (*proficient*), ranging from 2.98 (*SD* = 0.53) to 3.13 (*SD* = 0.44). On the MAP, both status (percentile rank = 54 for reading, 51 for mathematics) and growth (percentile rank = 51 for reading, 49 for mathematics) were near the normative national average (i.e., 50).

Internal Consistency

Evidence for internal consistency of composite scores from the FFT was strong. Cronbach's alpha was computed as an indicator of internal consistency for scores for each domain, the Domain 2 + 3 score, and the Total score, for both Round 1 and the average of all three rounds. Based on Round 1 only, alphas for domains ranged from moderate to fair (α = .77 for Domain 1: Planning and Preparation) to high to excellent (α = .90 for Domain 2: Classroom Environment), while both composite scores based on multiple domains were in the high to excellent range (αs = .93–.94). Based on means from across three rounds of data collection, all alphas were in the high to excellent range (αs = .90–.99). Table 4 depicts Cronbach's alpha by round and domain.

Score Stability

Stability for the FFT was acceptable, particularly for the composite scores. Score stability ranged from medium (*r* = .32) to very large (*r* = .87), though 32 of 36 coefficients across observation rounds, domains, and scoring methods

Table 4. Cronbach's Alpha by Round and Domain.

	Round 1			Across Rounds		
Domain	*n*	Number of Items	Cronbach's α	*n*	Number of Items	Cronbach's α
1. Preparation and Planning	104	6	.77	154	6	.99
2. Classroom Environment	147	5	.90	154	5	.94
3. Instruction	144	5	.89	154	5	.90
4. Professional Responsibilities	59	6	.83	154	6	.99
Domain 2 + Domain 3	141	10	.93	154	10	.96
Total	57	22	.94	154	22	.96

Note. Domain 2 + Domain 3 is the mean of Domain 2 and Domain 3.

Table 5. Score Stability by Round, Domain, and Scoring Method.

		R1, R2		R1, R3		R2, R3	
Domain	Scoring	*n*	*r*	*n*	*r*	*n*	*r*
1. Planning and Preparation	**Composite**	**98**	**.50**	**110**	**.39**	**101**	**.68**
	Traditional	98	.38	98	.32	101	.58
2. Classroom Environment	**Composite**	**150**	**.57**	**150**	**.50**	**154**	**.59**
	Traditional	150	.51	150	.40	154	.45
3. Instruction	**Composite**	**150**	**.51**	**150**	**.43**	**154**	**.53**
	Traditional	149	.45	147	.40	152	.54
4. Professional Responsibilities	**Composite**	**52**	**.57**	**54**	**.38**	**58**	**.80**
	Traditional	52	.63	56	.40	57	.74
Domain 2 + Domain 3	**Composite**	**150**	**.58**	**150**	**.51**	**154**	**.60**
	Traditional	149	.51	146	.46	152	.55
Total	**Composite**	**52**	**.74**	**53**	**.58**	**57**	**.87**
	Traditional	125	.40	140	.38	132	.60

Note. R1 = Round 1; R2 = Round 2; R3 = Round 3; Domain 2 + Domain 3 is the mean of Domain 2 and Domain 3. Composite scores are in **bold**. All correlations were significant at α = .05, one-tailed.

were in the medium or large ranges. The general trend, observed in 11 of 12 combinations, was that the correlation between Rounds 2 and 3 was the highest, and the correlation between Rounds 1 and 3 was the lowest. In 15 of 18 comparisons between composite scores and traditional scores, the composite scores shared a greater stability coefficient, although these differences were typically small, equal to .10 or less in 10 of the 15 comparisons. The differences were much greater at the Total score level; composite scores had stability coefficients higher than traditional scores by differences between .20 (Rounds 1 and 3) and .34 (Rounds 1 and 2). Regardless of pairing and scoring method, the correlation coefficients for the Total score at the composite level were the highest, ranging from large (r = .58 between Rounds 1 and 3) to very large (r's = .74–.87). Table 5 depicts score stability coefficients by round, domain, and scoring method.

Composite Scores Versus Traditional Scores

Composite scores and traditional scores appeared more interchangeable if averaged across three rounds rather than derived solely from Round 1. For Round 1, the correlations between composite scores and traditional scores were in the very large range for three coefficients (r's = .75–.79), and were in the nearly perfect range for the other three coefficients (r's = .92–.95). These coefficients are depicted on the diagonal of Table 6. Magnitudes of mean score differences between composite scores and traditional scores were within .03 of each other for all domains (see Table 3).

For scores averaged across the three rounds, the correlations between composite scores and traditional scores were in the very large range for two coefficients (r's = .84–.88), and were in the nearly perfect range for the other three coefficients (r's = .92–.95). These coefficients are depicted on the diagonal of Table 7. Magnitudes of mean score differences between composite scores and traditional scores were within .04 of each other for all domains (see Table 3).

Internal Structure Validity

Internal structure validity evidence was supportive of the FFT and its four domains. Among composite scores for Round 1, correlations for five of six pairs of domains were in either the medium range (r's = .47–.49) or the large range

Table 6. Domain Correlations Among Composite Scores and Traditional Scores for Round 1.

Domain	1. P&P	2. CE	3. IN	4. PR	D2 + D3	Total
1. P&P	**.79**	.53	.68	.59	.71	.75
2. CE	.57	**.92**	.68	.72	.93	.77
3. IN	.67	.77	**.92**	.61	.90	.83
4. PR	.65	.49	.47	**.76**	.76	.65
D2 + D3	.68	.95	.94	.52	**.95**	.87
Total	.87	.90	.85	.75	.95	**.75**

Note. Correlations among composite scores are below the diagonal. Correlations among traditional scores are above the diagonal (shaded). Correlations between composite scores and traditional scores are on the diagonal in **bold**. P&P = planning and preparation; CE = classroom environment; IN = instruction; PR = professional responsibilities; D2 + D3 = mean of Domain 2 and Domain 3. All correlations were significant at α = .05, one-tailed.

Table 7. Domain Correlations for Composite Scores and Traditional Scores Across Rounds.

Domain	1. P&P	2. CE	3. IN	4. PR	D2 + D3	Total
1. P&P	**.92**	.58	.60	.61	.62	.67
2. CE	.81	**.95**	.79	.69	.95	.87
3. IN	.84	.89	**.84**	.69	.94	.90
4. PR	.90	.80	.84	**.88**	.72	.77
D2 + D3	.86	.91	.97	.89	**.92**	.93
Total	.97	.91	.95	.95	.98	**.95**

Note. Correlations among composite scores are below the diagonal. Correlations among traditional scores are above the diagonal (shaded). Correlations between composite scores and traditional scores are on the diagonal in **bold**. P&P = planning and preparation; CE = classroom environment; IN = instruction; PR = professional responsibilities; D2 + D3 = mean of Domain 2 and Domain 3. All correlations were significant at α = .05, one-tailed.

(*r*'s = .57–.67). The correlation between Domain 2: Classroom Environment and Domain 3: Instruction was in the very large range (*r* = .77). Magnitudes of correlations were similar among traditional scores; five of the six correlations between Domain scores were in the large range (*r*'s = .53–.68). The correlation between Domain 2: Classroom Environment and Domain 4: Professional Responsibilities was in the very large range (*r* = .72). Table 6 depicts coefficients shared by Domains for composite scores (below the diagonal) and for traditional scores (above the diagonal) for Round 1.

Among composite scores across rounds, coefficients among the four domains were in the very large range (*r*'s = .81–.89) or nearly perfect range (*r* = .90). Among traditional scores, the pattern was similar to Round 1; five of the six correlations between domain scores were in the large range (*r*'s = .58–.69). The correlation between Domain 2: Classroom Environment and Domain 3: Instruction was in the very large range (*r* = .79). Table 7 depicts coefficients shared by domains for composite scores (below the diagonal) and for traditional scores (above the diagonal) across rounds.

Relations to Other Variables

Concurrent validity of FFT scores with student growth in reading and mathematics was primarily insufficient for Round 1 observations, though many significant correlations were found in combinations using scores averaged across three rounds. Trends were similar across content areas. Across six scores, two content areas, two scoring methods, and two numbers of rounds (i.e., Round 1 vs. all three rounds), 23 of 48 possible correlations were in the nonexistent range (*r*'s = −.08 to .09) or in the small range in either direction (*r* = −.11; *r*'s = .12–.19). Table 8 depicts correlation coefficients between FFT scores and MAP scores across content areas, rounds, and scoring methods.

Eight of the 10 correlations that were in the medium range or stronger were based on composite scores compiled across rounds. Of the 12 correlations between composite scores across rounds and growth in achievement, all were positive and 11 were significant, ranging from *r* = .19 for Domain 2: Classroom Environment relating to mathematics growth to *r* = .67 for the Total score relating to mathematics growth. The magnitudes of nine of these correlations between composite scores of domains were higher than the magnitudes of the corresponding correlations between traditional scores of Domains, with the difference ranging from .01 (Domains 2 + 3 relating to reading growth) to .36 (Total relating to mathematics growth). Of the 12 correlations between traditional scores across rounds and growth in achievement, all were positive and eight were significant, ranging from *r* = .15 for Domain 4: Professional Responsibilities relating to reading growth to *r*'s = .33 for

Table 8. Correlations With MAP Scores by Domain for Round 1 and Across Rounds.

	Reading				Mathematics			
	Round 1		Across Rounds		Round 1		Across Rounds	
Domain	CS	TS	CS	TS	CS	TS	CS	TS
1. Planning and Preparation	−.08	−.08	.29*	.16	.21	.26*	.47*	.28*
2. Classroom Environment	−.03	−.08	.24*	.20	−.01	−.05	.19	.17
3. Instruction	.08	.08	.24*	.31*	.09	.12	.22*	.33*
4. Professional Responsibilities	−.11	−.03	.33*	.15	.40*	.23	.66*	.33*
Domain 2 + Domain 3	.03	.00	.27*	.26*	.04	.04	.26*	.26*
Total	.01	−.01	.32*	.22*	.52*	.04	.67*	.31*

Note. MAP = Measures of Academic Progress; CS = composite scores; TS = traditional scores.
*Significant at α = .05, one-tailed.

Domain 3: Instruction or Domain 4: Professional Responsibilities relating to mathematics growth.

Domain 3: Instruction was the most consistently correlated Domain across content areas and scoring methods (*r*'s = .24–.33). The Total composite score, based on all three rounds, predicted reading growth in the medium range (*r* = .32) and predicted mathematics growth in the large range (*r* = .67).

Discussion

The study used extant data and classical test theory to evaluate the psychometrics of the FFT. Scores were generated from the instrument using both traditional domain scoring and a composite scoring approach that incorporated the mean of components nested within each domain. The composite scores were found internally consistent and more stable than traditional scoring. Composite scores were also found more predictive of growth in reading and mathematics, particularly when attained across three rounds of observations, compared with one round at the beginning of the year. Correlational evidence indicates that when averaged across three rounds, the composite scores and traditional scores share substantial variance, and thus may in many cases be interchangeable. Internal structure validity evidence was sufficient for both scoring methods, although composite scores averaged from three rounds indicated some redundancy across domains.

Reliability of FFT Scores

Reliability was addressed in the current study through estimates of internal consistency and score stability. Internal consistency analyses indicated that three of four domain scores may be appropriate for professional development decisions after a single round. Domain 1: Planning and Preparation was the exception, having internal consistency too low for such purposes. Both scores based on multiple domains, the Domain 2 + 3 score and the Total score, are internally consistent enough to contribute to high-stakes decisions (e.g., tenure, promotion, retention) in combination with other established measures. Averaged across three rounds, all six FFT scores had internal consistency coefficients in a range commensurate with high-stakes decision making (e.g., retention and promotion).

Score stability coefficients were predominantly in acceptable ranges (medium and large) for correlations between rounds spaced several months apart. Coefficients were lowest between Rounds 1 and 3 due to the increased time interval, and were highest between Rounds 2 and 3. It is possible that observers became more accurate after having used the FFT for the first time in a given year, or that observation schedules became compressed, such that there was a shorter lag time between Rounds 2 and 3 than between Rounds 1 and 2. Composite scores were consistently slightly more stable than were traditional scores, which may make such scores more appropriate for monitoring changes in teacher performance over time.

Collectively, these reliability estimates support the use of the composite scores as a justifiable, and perhaps preferable, way to use the FFT. Internal consistency estimates were in the same ranges observed by Lash et al. (2016). The internal consistency estimates exceeded those found by Kane and colleagues, who used designs that disaggregated the variance explained by variables (Kane & Staiger, 2012; Kane et al., 2011). The score stability estimates were lower than those found in previous studies, in terms of variance explained. It is difficult to compare reliability estimates across techniques; the internal consistency and stability estimates in the current study are consistent with expectations based on previous research, taking into consideration the time interval between rounds for stability.

Validity of Inferences From FFT Scores

Validity evidence for inferences from FFT scores across rounds, scoring methods, and domains was mixed. Based on

Round 1, composite scores and traditional scores appeared interchangeable for Domain 2: Classroom Environment and for Domain 3: Instruction, as well as for the combination of those two domains. For Domain 1: Preparation and Planning, Domain 4: Professional Responsibilities, and the Total score, correlations were at magnitudes indicating composite scores and traditional scores were providing unique information, and thus not interchangeable. For these scores, using both a composite score and a traditional score may be preferred, as the two may be providing unique information. Based on an averaging of three rounds of FFT, composite scores and traditional scores correlate at a magnitude indicating the two could be used interchangeably.

Validity evidence based on internal structure, in the form of correlations among domains and other scores, was good for the traditional scores and mixed for the composite scores. Conceptually, the various domains of the FFT are related, and scores should be positively correlated in the medium range or large range, indicating about 9% to about 49% shared variance (i.e., r^2). Magnitudes of correlations in the nonexistent or small ranges may indicate domains that are not related at all, and magnitudes in the very large or nearly perfect ranges may indicate redundancy among domains. Based on Round 1, most of the coefficients were in these ranges, indicating an appropriate level of overlap for the FFT Domains. Based on the average of three rounds, coefficients for the composite scores were too high, potentially indicating redundancy.

Validity evidence based on relations to other variables was addressed by examining the relations of FFT scores to student reading and mathematics achievement as measured by MAP. The FFT scores from Round 1 were not uniformly related to growth in either area, nor were the traditional scores, regardless of being based on Round 1 or on all three rounds. Composite scores based on three rounds were the most related to achievement growth, ranging from about 4% to 44% shared variance across domains and content areas. At the total score level, the composite scores based on three rounds share 10% variance with reading growth and 45% variance with mathematics growth. The correlation magnitudes in the current study exceeded those of previous studies (Kane & Staiger, 2012; Kane et al., 2011; Lash et al., 2016) in both reading and mathematics, and constitute strong evidence that the composite scores of the FFT across three rounds can be used for prediction of growth in achievement.

Practical Implications

Evidence from the current study supports using composite scores from the FFT instead of, or in addition to, traditional scores made at the domain and total levels. The measurement argument for doing so is scores based on more observations tend to be more reliable. Using composite scores also provides a better rationale for addressing component scores nested within a domain that has a low score. To illustrate, consider a teacher who receives a low score in Domain 2: Classroom Environment. Although this score provides the global feedback that the classroom environment should be improved, this information on its own is not very specific compared with a nested component that may also have a low score, such as *Creating an Environment for Student Learning is Creating an Environment of Respect and Rapport*. To further specify, the definition, elements, and indicators of this component can form the basis of a professional development lesson or plan. Although these steps could be taken without mathematically connecting domain scores to component scores, making such a connection strengthens the logic for doing so.

Two primary reasons for using an observational measure of teacher effectiveness are to provide feedback to teachers in a timely fashion and to evaluate educator performance. Psychometric evidence from the current study is primarily more supportive of scores based on an average of three rounds, rather than on a single round in the fall of the academic year, as one would expect given the information gleaned from additional data points. For most domains, scores from a single round appear to be acceptable for low-stakes decisions such as professional development planning and progress monitoring. To make summative decisions, a minimum of three rounds of FFT data should be collected, and additional measures should be used.

The present study builds on the existing literature by examining the FFT's traditional domain scoring approach compared with a mathematical composite scoring approach that connects to more specific teacher practice information, which may in turn enhance professional development. In this investigation, findings revealed that the composite scores were superior to traditional domain scores in terms of stability and relation to student growth in reading and mathematics. These findings are particularly noteworthy because this study was conducted in urban high-poverty schools, where growth in achievement tends to be below national benchmarks and less variable.

Overall, these findings support the use of a composite scoring approach. This approach may enhance the use of FFT for both formative and summative assessments of teaching practices, and in doing so advance professional development decisions. Also, enhancing the reliability of scores and the validity inferences drawn from the FFT is critically important for generating balanced and fair teacher evaluation decisions, as well as for formulating professional development supports that meet the data-specific needs of educators in schools.

Limitations

Convenience sampling is the primary limitation of the current study. Because teachers were evaluated in high-poverty charter schools, generalizability to other school settings is unknown. Also, multiple observations were conducted

within the same schools and by the same administrators, challenging the assumption that observations were independent. However, the process used in this investigation resembles that of routine educator evaluation, offering support for inferences connected to actual practice. Due to a limited sample size, this study did not examine the influence of school types (i.e., elementary vs. middle schools), grade-level assignments, or teacher content area on the psychometric properties of the FFT. Disaggregation of such analyses would enhance future research. Finally, because the schools were part of the SSI Project, initiatives such as professional development programming and a pay-based compensation system were implemented alongside data collection. Although such initiatives are unlikely to systematically affect estimates of the psychometrics of the FFT, it is possible they could, or at a minimum the generalizability beyond the project could be further limited.

Future Research

Additional work should be conducted on the reliability and validity evidence for the FFT. Because it is an observational tool, studies should be conducted estimating interobserver agreement. With a sufficient sample size, it would also be helpful to conduct factor analyses on component scores of the FFT. In another study, relations to other variables could be evaluated using the composite scores along with a second observational tool, extending the work Kane and Staiger (2012) conducted with the traditional scores. Finally, given the FFT scores predict only a proportion of growth in achievement, research should be conducted using multiple measures and multiple methods of capturing teacher effectiveness to further determine the critical aspects.

Conclusion

The results provide support for averaging the component scores of the FFT to form composites having some reliability and validity evidence beyond that of the traditional scores typically used with the measure. Regardless of the scoring method, the reliability and validity indices of the FFT are higher for scores based on multiple rounds of observations. In practice, a single round of observations is likely sufficient for informing professional development, whereas multiple rounds are necessary for evaluating performance. Although no measure should be used alone for such purposes, composite scores from the FFT clearly reflect teacher variables related to growth in reading and mathematics. These findings support the use of the FFT in schools for a variety of inferences and purposes.

Authors' Note

The positions and opinions expressed in this article are solely those of the authors.

Acknowledgment

The authors thank Jiefang Hu, Christopher Dudek, and Kevin Crouse for their excellent research assistance with this study.

Declaration of Conflicting Interests

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

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The current study was implemented as part of the School System Improvement (SSI) Project, a collaboration between multiple universities and charter schools funded by the U.S. Department of Education's Office of Innovation and Improvement as part of the Teacher Incentive Fund program (awarded to Rutgers, The State University of New Jersey; #S374A120060).

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington DC: American Educational Research Association.

Bennet, N. (1988). The effective primary school teacher: The search for a theory of pedagogy. *Teaching and Teacher Education*, *4*, 19–30.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159.

Council of Chief State School Officers. (2011, April). *Interstate Teacher Assessment and Support Consortium (InTASC) Model Core Teaching Standards: A resource for state dialogue*. Washington, DC: Author.

Cronin, J. (2005). *NWEA reliability and validity estimates: Achievement level tests and Measures of Academic Progress*. Lake Oswego, OR: Northwest Evaluation Association.

Danielson, C. (2013). *The Framework for Teaching: Evaluation Instrument*. Princeton, NJ: The Danielson Group.

Hattie, J. A. (1992). Measuring the effects of schooling. *Australian Journal of Education*, *36*, 5–13.

Hopkins, W. G. (2001). *A scale of magnitudes for effect statistics*. Retrieved from http://www.sportsci.org/resource/stats/effectmag.html

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

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

Lash, A., Tran, L., & Huang, M. (2016). *Examining the validity of ratings from a classroom observation instrument for use in a district's teacher evaluation system* (REL 2016–135). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory West.

Marzano, R. J., Frontier, T., & Livingston, D. (2011). *Effective supervision: Supporting the art and science of teaching*. Alexandria VA: Association for Supervision and Curriculum Development.

McLaughlin, J. A., & Jordan, G. B. (2010). Using logic models. In J. S. Wholey, H. P. Hatry, & K. E. Newcomer (Eds.), *Handbook of practical program evaluation* (pp. 55–80). San Francisco, CA: Jossey-Bass.

Murphy, K. R., & Davidshofer, C. O. (2005). *Psychological testing: Principles and applications* (6th ed.). Upper Saddle River, NJ: Prentice Hall.

No Child Left Behind Act, 20 U.S.C.A. § 6301 et seq. (2001).

Northwest Evaluation Association. (2011, January). *Technical manual for Measure of Academic Progress and Measure of Academic Progress for Primary Grades*. Portland, OR: Author.

Sartain, L., Stoelinga, S. R., & Brown, E. R. (2011). *Rethinking teacher evaluation in Chicago: Lessons learned from classroom observations, principal-teacher conferences, and district implementation*. Chicago, IL: Consortium on Chicago School Research at the University of Chicago.

Wang, M. C. (1991). Productive teaching and instruction: Assessing the knowledge base. *Phi Delta Kappan*, *71*, 470–478.

Wang, S., McCall, M., Jiao, H., & Harris, G. (2013). Construct validity and measurement invariance of computerized adaptive testing: Application to Measures of Academic Progress (MAP) using confirmatory factor analysis. *Journal of Educational and Developmental Psychology*, *3*(1), 88–100.