

The Reliability of Framework for Teaching Scores in Kindergarten

Journal of Psychoeducational Assessment
2020, Vol. 38(7) 831–845
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0734282920910843
journals.sagepub.com/home/jpa



Helen Patrick¹ , Brian F. French² ,
and Panayota Mantzicopoulos¹

Abstract

We evaluated the score stability of the Framework for Teaching (FFT), a prominent observation instrument used for teacher evaluation. Three raters each scored 200 reading and mathematics lessons taught by 20 kindergarten teachers. Using Generalizability theory analyses, we decomposed the FFT's Classroom Environment, Instruction, and Total scores into potential sources of variation (teachers, lessons, raters, and their interactions). The scores' variances attributable to differences among teachers were 71% and 76% for Classroom Environment, 49% and 37% for Instruction, and 69% and 66% for the Total score, for reading and mathematics, respectively. Reliability estimates (G) ranged from 0.92 to 0.96 for Classroom Environment and Total scores; they were 0.87 and 0.79 for reading and mathematics Instruction. Decision studies indicated that two raters, each scoring three reading lessons or four mathematics lessons, are necessary to achieve sufficiently reliable Total scores. For Instruction scores, three raters each scoring seven readings lessons are needed; more than four raters each scoring eight lessons are needed for mathematics.

Keywords

teacher evaluation, teacher accountability, classroom observation, generalizability theory, Framework for Teaching (FFT)

Classroom observations have long been the cornerstone of teacher evaluation. In response to both concerns that teacher evaluations were superficial and unreliable (e.g., Weisberg et al., 2009), and the federal requirements that ensued (U.S. Department of Education, 2009, 2011), states began adopting formal observation systems with established protocols and rating systems. Most teachers and other educational professionals are now evaluated with observation instruments, often in concert with other indices of effectiveness such as standardized test scores (Steinberg & Kraft, 2017). Despite the prominence of many of these observation instruments, however, there is little evidence of the reliability, or stability, of their scores and, relatedly, the

¹Purdue University, West Lafayette, IN, USA

²Washington State University, Pullman, USA

Corresponding Author:

Helen Patrick, Department of Educational Studies, Purdue University, 100 North University Street, West Lafayette, IN 47907, USA.

Email: hpatrick@purdue.edu

minimum number of observations needed for accurate assessments. Without this information, school administrators cannot make informed decisions when selecting an instrument and deciding on an observation schedule. Although this paucity of reliability evidence is concerning in general, it is of particular concern for the teachers whose evaluations are based almost exclusively on their observed instruction.

We address the need for reliability data by examining scores from the observation instrument used most widely in the United States, the Framework for Teaching (FFT; Danielson, 2013). We focus on kindergarten, a grade level where teachers are evaluated predominantly with classroom observations (Cohen & Goldhaber, 2016; Garrett & Steinberg, 2015). We use Generalizability (G) Theory to partition the sources of variance in teachers' FFT scores, separately for the two central subject areas taught in kindergarten—reading and mathematics. We then conduct a series of Decision (D) studies to estimate differences in reliability with iterative increases in the number of lessons and raters.

Classroom Observations for Evaluating Teachers and Other Educational Professionals

Teachers and other educational professionals are evaluated with multiple measures, although the weight given to each typically varies across grade levels and content areas. In general, teachers in third through fifth grades and English and mathematics teachers in later grades, whose students take state standardized tests, are evaluated with a combination of student achievement and classroom observation scores. However, the 70% of teachers who teach grade levels or subjects outside the standardized testing program are evaluated primarily with observations (Steinberg & Kraft, 2017), as are special educators and school psychologists (e.g., Pennsylvania Department of Education, 2014).

In an effort to ensure that observations are transparent and reliable, most school districts use a formal observation instrument chosen from a selection approved by the state. Broadly, these practices address creating a supportive classroom environment, managing student behavior, and specific instructional practices (Kane & Staiger, 2012). Despite the concern that evaluation systems are fair and reliable, though, there is a conspicuous paucity of empirical evidence across observation measures to guide and justify their use. Importantly, this includes evidence of reliability.

Reliability of Observation Scores

The reliability of observations is typically considered in terms of rater agreement. Although consistent scoring across raters is crucial, an equally important aspect of reliability that receives much less consideration is the stability of scores. That is, for an instrument to reflect differences among teachers' instruction, beyond any single instance, its scores need to be consistent across lessons. The stability of an instrument's scores has implications for the number of observations necessary to form an accurate assessment; greater stability requires fewer observations.

Without reliability data to guide decision-making, selecting an instrument to use and determining the number of observations to conduct are not informed by research, but rather left to intuition. Many, but not all, states require teachers be observed twice per year; some states, though, allow less frequent observations for some teachers (e.g., those with tenure or previous high ratings) (National Council on Teacher Quality, 2019). However, policies are not tied to a specific instrument or based on the stability of its scores. This is an issue of concern for all teachers who are evaluated with observation instruments. It is especially problematic for teachers whose evaluations are based primarily on observations, because inaccuracies cannot be countered by achievement data.

Measuring Score Stability With Generalizability Theory

An assumption about scores used to evaluate teachers is that they reflect stable differences among teachers, and not variance stemming from other factors, such as a teacher's lesson-to-lesson variability or inconsistencies among raters in their scoring of particular teachers (i.e., Rater \times Teacher interaction). One method of identifying the stability of observation scores is Generalizability (G) theory (Brennan, 2001; Shavelson & Webb, 1991). Analysis involves identifying the different sources of score variance, such as teachers, lessons, raters, and interactions between them, and the relative contribution of each source (Brennan, 2001; Shavelson & Webb, 1991). The resulting G coefficient is an estimate of stability, and is interpreted similarly to Cronbach's alpha (Webb et al., 2006).

G-study estimates may also be used to conduct Decision (D) studies, whereby calculations indicate the projected iterative improvements in stability that accrue as the number of raters or observations increase. Thus, D-studies indicate the number of raters and lessons needed to obtain observation scores that are sufficiently stable to warrant their use for research or individual accountability purposes.

G-studies have been conducted with many classroom observation instruments. In general, studies indicate that scores do not meet acceptable reliability standards (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014) when they are based on the number of observed lessons typical in teacher evaluations (e.g., Ho & Kane, 2013; Praetorius et al., 2012). In addition, ratings of instructional practices are less stable, and therefore require more observations, than those of classroom climate (Meyer et al., 2011; Praetorius et al., 2014). Two features of this research are relevant to our present study.

First, most G-studies (e.g., Hill et al., 2012; Praetorius et al., 2014) employ researcher-developed instruments that are not approved or recommended by any state for evaluating teachers. Therefore, although of interest to researchers, the results are unlikely to inform educational decisions.

Second, most G-studies with instruments used to evaluate teachers were conducted in either upper elementary and middle school (Ho & Kane, 2013; Kane & Staiger, 2012) or high school (Mashburn et al., 2014) classrooms; one study (Meyer et al., 2011), though, did not identify the grade level(s) involved. Consequently, there is woefully little research on the stability of observation instruments used to evaluate teachers in the early elementary grades. This is an important omission, given evidence that observation scores vary systematically across grade levels (Mihaly & McCaffrey, 2014). Thus, results from one grade level cannot be assumed to apply to another. We respond to the critical need for evidence about the reliability of classroom observation scores in the early grades by using G-theory to examine the most prevalent observation system used for teacher evaluation, the FFT (Danielson, 2013).

The FFT

The FFT is the most prominent classroom observation instrument in the United States, recommended by 26 states and the District of Columbia for evaluating teachers and other educational professionals (Center on Great Teachers and Leaders, 2013). It is rooted in the PRAXIS III, a performance assessment developed by Educational Testing Service (ETS) for evaluating and licensing beginning teachers (Danielson, 2007). Danielson (2007), who contributed to the development of the PRAXIS III, modified and re-named it, then promoted the FFT as a measure of in-service teachers' effectiveness. The FFT is "grounded in the constructivist approach" (p. 17), "does not endorse any particular teaching style" (p. 25), and "is a generic instrument, applying to all disciplines" (Danielson, 2013, p. 6). Its components measure "those aspects of a teacher's

responsibilities that have been documented through empirical studies and theoretical research as promoting improved student learning” (Danielson, 2013, p. 3).

The FFT comprises four domains of practice: Preparation and Planning, Classroom Environment, Instruction, and Professionalism (Danielson, 2013). Two domains—Classroom Environment and Instruction—involve observing teachers; therefore, they are the focus of our study. When considering the observation-based components of the FFT, researchers usually combine teachers’ Classroom Environment and Instruction scores to create a composite score (e.g., Garrett & Steinberg, 2015; Polikoff & Porter, 2014), mirroring the single score that teachers receive. However, researchers sometimes consider the stability of practices reflecting the classroom environment and those specific to instruction separately (e.g., Meyer et al., 2011; Praetorius et al., 2014).

Generalizability Studies With the FFT

We located three G-studies involving the FFT. Two were conducted as part of the Measuring Effective Teaching (MET) project and included primarily middle school English and mathematics teachers. In the first study (Kane & Staiger, 2012), researchers rated four lessons from each of 1,333 teachers, whereas the second (Ho & Kane, 2013) used FFT scores from 67 teachers (four lessons each) for whom lessons were rated as part of one county’s evaluation system. The variance in FFT Total scores attributable to teacher differences was similar for both studies (37% and 39%). The reliability of scores from one lesson was low (0.37). Although reliability increased to 0.67 with four lessons (Kane & Staiger, 2012), it did not meet acceptable standards (AERA et al., 2014).

It is not clear whether the MET G-study results apply to other grade levels. More observations may be necessary for reliable scores in kindergarten, when children are just learning what the student role involves, compared with the later grades when students have learned school norms and routines. Whether or not the results from the middle grades are replicated with our sample of kindergarten teachers is an issue we explore in this study. We also consider two important issues not addressed in the MET project studies: (a) stability of the same teachers’ scores for English and mathematics lessons separately, rather than aggregated across content areas as in the MET studies, and (b) score stability for the Classroom Environment and Instruction domains individually, in addition to the aggregated Total score.

The third FFT G-study involved science lessons in 10 kindergarten classrooms, and considered the two domain scores separately but not in aggregate (Mantzicopoulos et al., 2018). The variance between teachers’ Classroom Environment scores was 36% but was considerably less (16%) for Instruction.

It is possible that the stability of teachers’ Instruction scores differs across content areas, reflecting the different curricula, instructional resources (e.g., manipulatives for mathematics), predominant instructional formats (e.g., more small groups for reading), and pedagogical content knowledge involved (Grossman et al., 2004). Scrutiny of teachers’ instruction, time allocated to teaching, and professional development also vary according to the relative valuing of content areas. In the early elementary grades reading is viewed as most important, followed by mathematics, with all other subjects considered much less important (Grossman et al., 2004). If score stability differs across content areas—a question that has yet to be addressed—information about reading and mathematics instruction is most relevant to teachers and their evaluators. Therefore, we build on Mantzicopoulos et al.’s (2018) study of science lessons by examining scores for reading and mathematics.

The Present Study

We investigated the stability of kindergarten teachers’ Classroom Environment, Instruction, and Total FFT scores for reading and mathematics lessons, given that evaluation practices assume

score stability despite a paucity of evidence. We conducted a series of G-theory studies to decompose the scores' reliability into potential sources of variance, and then, using a series of D-studies, we estimated improvements in reliability if there were additional lessons and raters. We examined scores for reading and mathematics lessons separately, because score stability may differ given that instruction differs by content area; there was insufficient research on which to base content-specific hypotheses. Based on extant research, we hypothesized that Instruction scores would be less stable than Classroom Environment and Total scores. We evaluated score stability in reference to reliability standards for decision-making about individual teachers (AERA et al., 2014).

Method

Participants

Participants were 20 kindergarten teachers (19 female and 1 male; 18 White and 2 Hispanic) in six public schools in Indiana. We received consent from 23 teachers (72%) in seven schools; however, 3 teachers recorded insufficient lessons for this study. The 20 teachers comprised: all 3 kindergarten teachers in each of two schools, all 6, 2 of the 3, 3 of the 4, and 3 of the 6 kindergarten teachers in the other four schools. Teachers' experience ranged from 1 to 33 years ($M = 16$ years).

We received informed consent to collect data on 79.4% of students. Most students (63.2%) were White; 22.8% were Hispanic, 9.1% were Black, and 4.6% were Multiracial or Other; 53% of students received free or reduced-cost lunch (FRL). The schools were academically, ethnically, and socioeconomically diverse. Across schools, 30% to 73% of students received FRL. Schools were located in rural areas, small towns, small cities, and the urban fringe of a large city; state report card grades ranged from A to C.

Lessons

Teachers used a researcher-issued iPad and stand to record one reading and one mathematics lesson each week for 10 weeks during the spring, and uploaded lessons to a secure website. Teachers chose the 10 weeks in which to record; we did not require teachers to record in consecutive weeks. We told teachers we were interested in their regular lessons, and encouraged them to upload lessons even if they did not occur as anticipated or teachers were dissatisfied with them. We asked teachers to record entire lessons of at least 20 min; different activities were usually included within the lesson. Each teacher's set of lessons included both whole group instruction and individual seat work; some teachers also included lessons involving centers. We selected five reading and five mathematics lessons randomly from each teacher ($N = 200$ lessons). Reading lessons averaged 24 min 57 s ($SD = 7:47$) and mathematics lessons averaged 23 min 57 s ($SD = 7:35$).

Observation Instrument and Procedure

The FFT (Danielson, 2007, 2013) observation measure of teacher practices comprises two domains with four components in each. The Classroom Environment components are (a) creating an environment of respect and rapport, (b) establishing a culture for learning, (c) managing classroom procedures, and (d) managing student behavior. The Instruction components are (1) communicating with students, (2) using questioning and discussion techniques, (3) engaging students in learning, and (4) using assessment in instruction. At the end of the observation period, raters score each component on a 4-point scale (1 = *unsatisfactory*, 2 = *basic*, 3 = *proficient*, 4 = *distinguished*), then average component scores within each domain.

Evidence about the structure of FFT scores is inconsistent, therefore so are researchers' procedures for creating scores. We use two approaches. Specifically, we consider Classroom Environment and Instruction scores separately, which allows us to investigate whether they are similarly stable, in addition to creating a Total score by aggregating the two domains, therefore allowing comparison with the MET studies in middle school.

Raters. The raters in the present study were part of a larger group of eight FFT-trained and -certified raters engaged in a larger project investigating numerous observation instruments. The three raters were educational psychology graduate students. Two were former teachers (one elementary and one secondary) with either a bachelor's or master's degree in education; none of the raters had administrative experience.

Rater training. Raters completed Teachscape's Focus for Observers, an on-line, self-paced FFT training and certification program used by school district evaluators, and were certified Proficient by passing the set of two ETS-administered tests (Teachscape, n.d.). Tests were on-line and took approximately 3 hr each. Results are reported as Proficient or Not Proficient. Although criteria to be certified as proficient are not available, "the passing score is based on overall performance on the multiple choice and video scoring sections, in which a user can gain full and partial credit" (Growth Through Learning, n.d.).

Rater calibration. Prior to scoring the lessons used in this study, the three raters engaged in calibration activities with the five other FFT-trained and -certified project members. Activities involved viewing four recorded lessons not part of the current study, scoring lessons individually, discussing scores assigned, and calculating inter-rater agreement. Exact agreement ranged from 74% to 83% across the lessons ($M = 80\%$).

Observing and scoring lessons. The three raters independently observed and scored each lesson. They followed different schedules, to ensure that lessons were not scored sequentially or grouped by teacher or content area.

Analysis Plan

Raters scored all lessons, therefore there were no missing data.

G-theory model. We used a two-facet (Lessons, Raters), partially nested (lessons within teachers), random design to decompose the variance in FFT scores (Shavelson & Webb, 1991). Lessons were not identical across teachers; therefore, we could not estimate a main effect for Lesson; Lesson is confounded with the Teacher \times Lesson interaction (Brennan, 2001). That is, we cannot determine whether teachers' scores differed from lesson to lesson (interaction effect) or whether different practices were associated with specific lessons. Because each rater scored every lesson, the Rater facet was crossed with Lessons (i.e., Rater \times Lesson interaction effect).

Our model follows Brennan's (2011) G-theory model estimation guidelines (i.e., at least two levels per facet, many tasks, and at least two raters). Of note, we included Lesson as a facet. Not incorporating lesson (i.e., occasion) would misrepresent the relative contributions of facets and error variances by overestimating reliability and underestimating error.

For each score, we partitioned the variance (σ^2) into the following components:

1. Teacher (t, σ_t^2): Variance attributed to differences across teachers.
2. Rater (r, σ_r^2): Variance attributed to differences across raters.

Table 1. Descriptive Statistics and Correlations for FFT Scores.

Domain	M	SD	Minimum	Maximum	Reading			Mathematics		
					1	2	3	4	5	6
Reading										
1. Classroom Environment	2.72	0.43	1.75	3.15	—					
2. Instruction	2.22	0.34	1.52	2.78	.93	—				
3. Total	2.47	0.38	1.65	2.97	.99	.98	—			
Mathematics										
4. Classroom Environment	2.58	0.48	1.22	3.02	.91	.87	.91	—		
5. Instruction	2.16	0.28	1.67	2.88	.80	.90	.86	.81	—	
6. Total	2.37	0.36	1.46	2.95	.91	.92	.93	.98	.92	—

Note. FFT = Framework for Teaching. All correlations significant at $p < .01$.

3. Teacher \times Rater ($t \times r, \sigma^2_{tr}$): Variance attributed to inconsistencies between raters in evaluating a particular teacher's practices.
4. Lesson: Teacher ($l:t, \sigma^2_{l,t}$): Variance attributed to inconsistencies in teacher practices from lesson to lesson.
5. Lesson: Teacher \times Rater ($l:t \times r, e, \sigma^2_{rl,tl,e}$): Residual variance comprising unmeasured effects and random events affecting the measurement.

Consistent with models used with other studies of classroom observation measures (e.g., Hill et al., 2012; Mashburn et al., 2014), the teacher, rater, and lesson effects were random. The random effects model reflects educational practice and is based on the assumption that teachers, raters, and lessons are replaceable with equivalent sets drawn from our universe of teachers, lessons, and raters (Shavelson & Webb, 1991). In practice, teachers are not all evaluated on the same lesson. Thus, in our study Lesson is a sampling of lessons typically taught by kindergarten teachers during the spring.

We used EduG (Swiss Society for Research in Education Working Group, 2006) to estimate the models. We report the five variance components and their standard errors, to provide information on the generalizability of the findings given the sample and conditions. We also report the relative reliability estimates (G), which range from 0 to 1.0; higher estimates represent greater dependability of the measurement procedure (Shavelson & Webb, 1991). This index can be interpreted like coefficient alpha, and used when relative decisions about teachers (e.g., ranking performance) are being made (Cronbach et al., 1972).

Evaluation criteria. The criteria for acceptable levels of score stability differ, depending on how scores will be used. Specifically, scores used for high stakes decisions require a higher degree of stability than is required for research purposes (AERA et al., 2014; Nunnally & Bernstein, 1994). Thus, following recommendations of Nunnally and Bernstein, and standards for educational and psychological measurement (AERA et al., 2014) we required estimates to be at .90 or above to provide evidence of score stability for decisions at the individual teacher level.

Results

Descriptive Statistics

Descriptive statistics and correlations are shown in Table 1. Across domains and content areas, average scores ranged from 2.16 to 2.72. Scores within each content area were correlated highly

($r_s > 0.80$), as were comparable domain scores across content areas ($r_s \geq 0.90$). Across domains and content areas, skewness and kurtosis values ranged from -1.36 to 1.83 , supporting the normality assumption (i.e., values < 2.0 ; Lomax & Hahs-Vaughn, 2012). Also, review of Q-Q plots did not suggest severe departures from normality. The evidence did not raise concern of normality violations, especially given the robustness of variance components analysis to minor deviations from normality.

Reliability Estimates

The relative reliability (G) estimates for teachers' Classroom Environment scores were 0.95 and 0.96 for reading and mathematics lessons, respectively. For Instruction, G estimates were 0.87 and 0.79 for reading and mathematics, respectively. The G estimates for the Total reading and mathematics scores were 0.94 and 0.92, respectively. The values for Classroom Environment and Total scores meet reliability criteria for making decisions about individuals, however, Instruction scores do not.

Sources of Variance in FFT Scores

The decomposition of variance in FFT scores is shown in Table 2.

Classroom Environment. The partitioning of variance in Classroom Environment scores was similar for reading and mathematics lessons. For both subjects, approximately three-quarters (71.2% & 76.4%) of the variance in scores was attributed to differences between teachers, whereas approximately 11% was due to teachers' lesson-to-lesson variation. Rater variability was a very small component of the overall variance; together, Rater and Rater \times Teacher interaction comprised less than 2% of the variance in scores. The residual variances were 15.2% (reading) and 10.7% (mathematics).

Instruction. Considerably less variance in Instruction scores was attributed to the teacher: 48.7% in reading and 37.3% in mathematics. Conversely, a greater proportion of variance was across teachers' lessons: 24.5% in reading and 29.5% in mathematics. As with Classroom Environment, rater variance was small (0.10% & 1.5%) for each content area; however, there was substantially more Rater \times Teacher variance: 3.0% and 6.5% for reading and mathematics, respectively. The residual variances, at 23.7% (reading) and 25.2% (mathematics), were also larger than for Classroom Environment, suggesting there may be additional facets that can explain variance beyond random error.

Total score. The partitioned variances in the Total score were similar for both content areas. Approximately two thirds (68.7% & 66.1% for reading and mathematics, respectively) of the variance was between teachers and 14% to 17% was among teachers' lessons. Variance associated with raters (including Rater \times Teacher) was low: 2.3% for reading and 4.0% for mathematics. The residual variances were 14.6% and 13.1%, for reading and mathematics, respectively.

Decision (D) Study Estimates for the Optimal Number of Raters and Lessons

We conducted D-study analyses to estimate the reliability coefficients should different numbers of lessons (from 1 to 8) be scored by different numbers of raters (from 1 to 4). Thus, the D-study uses the G-coefficients to extrapolate beyond the five lessons and three raters used in this study. In line with standards for educational and psychological measurement (AERA et al., 2014), we considered a coefficient of ≥ 0.90 to indicate sufficient reliability for individual evaluation.

Table 2. Variance Components of FFT Scores for Reading and Mathematics Lessons.

Domain	Source of variance	Reading		Mathematics	
		Variance component (SE)	% of total variance	Variance component (SE)	% of total variance
Classroom Environment	Teachers (t)	0.175 (0.056)	71.2	0.218 (0.070)	76.4
	Raters (r)	0.001 (0.001)	0.5	0.001 (0.001)	0.3
	Lessons: Teachers (l:t)	0.029 (0.006)	11.8	0.031 (0.006)	11.2
	Teachers × Raters (tr)	0.003 (0.002)	1.3	0.004 (0.002)	1.4
	Residual (l:tr,e)	0.037 (0.004)	15.2	0.030 (0.003)	10.7
	Total		100.0		100.0
	Relative Error Variance	0.009 (0.096)		0.002 (0.099)	
Instruction	Teachers (t)	0.102 (0.036)	48.7	0.060 (0.023)	37.3
	Raters (r)	0.0002 (0.0001)	0.10	0.002 (0.002)	1.5
	Lessons: Teachers (l:t)	0.051 (0.010)	24.5	0.047 (0.009)	29.5
	Teachers × Raters (tr)	0.006 (0.003)	3.0	0.010 (0.004)	6.5
	Residual (l:tr,e)	0.049 (0.005)	23.7	0.040 (0.004)	25.2
	Total		100.0		100.0
	Relative error variance	0.015 (0.12)		0.015 (0.12)	
Total	Teachers (t)	0.135 (0.036)	68.7	0.120 (0.040)	66.1
	Raters (r)	0.001 (0.0001)	0.40	0.001 (0.001)	0.6
	Lessons: Teachers (l:t)	0.028 (0.010)	14.4	0.030 (0.006)	16.7
	Teachers × Raters (tr)	0.003 (0.003)	1.9	0.006 (0.002)	3.4
	Residual (l:tr,e)	0.028 (0.005)	14.6	0.023 (0.003)	13.1
	Total		100.0		100.0
	Relative error variance	0.008 (0.093)		0.009 (0.098)	

Note. FFT = Framework for Teaching.

Results are shown in Table 3. Estimates reached 0.90 for reading Classroom Environment with four raters each scoring two lessons, and for mathematics Classroom Environment with three raters each scoring two lessons (see Figure 1); more lessons and raters are needed for reliable Instruction scores (see Figure 2). Estimates for reading Instruction were 0.90 with three raters each scoring seven lessons or two raters each scoring eight lessons. More than four raters, each scoring eight lessons, are needed to estimate mathematics Instruction reliably.

Figure 3 displays the estimated coefficients for the Total score with different configurations of raters and lessons. For reading lessons, estimates reached 0.90 with four raters each scoring two lessons, or two raters each scoring three lessons. With mathematics lessons, three raters each scoring three lessons, or two raters each scoring four lessons, are needed. If one rater is used, it is necessary to observe four reading lessons or seven mathematics lessons.

Discussion

In the present study, we examined the reliability (i.e., stability) of FFT scores applied to kindergarten reading and mathematics lessons. We add to the information about the FFT in three ways. First, we found greater variance between teachers in their Total score compared with that found for middle grade teachers (Ho & Kane, 2013; Kane & Staiger, 2012). Second, when examining the two domain scores separately, we found greater lesson-to-lesson variability for Instruction than for Classroom Environment scores, consistent with findings with

Table 3. Reliability Estimates (G) for FFT Scores With Three Raters Scoring Different Numbers of Lessons.

Domain	Number of lessons ^a							
	1	2	3	4	5	6	7	8
Reading								
Classroom Environment	.80	.89	.92	.94	.95	.96	.96	.96
Instruction	.59	.75	.81	.85	.87	.88	.90	.91
Total	.80	.89	.92	.94	.95	.96	.97	.97
Mathematics								
Classroom Environment	.83	.91	.93	.95	.96	.96	.96	.97
Instruction	.48	.64	.72	.76	.79	.81	.83	.84
Total	.75	.85	.90	.92	.92	.94	.95	.95

^aCoefficients for five lessons were calculated from G-studies; other coefficients come from D-studies. All estimates are based on three raters scoring each lesson.

other observation instruments. And third, we found evidence that the stability of scores in kindergarten differs across the content areas observed. Our findings (a) identify psychometric issues regarding observation instruments that are critically important for those affected by their use; and (b) add to the literature that is critical to researchers, teachers, school psychologists, educational diagnosticians, special educators, and school administrators with interests in the classroom contexts in which students learn and develop.

In decomposing the variance of the kindergarten teachers' Total scores, the portion attributable to differences between teachers was considerably greater than was found in the middle grades; 67% to 69% versus 37% to 39% (Ho & Kane, 2013; Kane & Staiger, 2012). Of note, our analyses were based on a similar number of lessons: five per teacher compared with four in the MET studies. Interestingly, there was also more variability across teachers' lessons in our kindergarten sample than was found with middle grade teachers. These differences between kindergarten and middle school suggest that the FFT may not be similarly sensitive across grade levels; this is an area where further research is needed.

Our finding of considerably greater stability in ratings of teachers' practices reflective of their classroom environment compared with ratings of their instruction replicates those found with other observation instruments (e.g., Meyer et al., 2011; Praetorius et al., 2014) and with the FFT in kindergarten (Mantzicopoulos et al., 2018). This may be because practices rated in the Classroom Environment domain, such as keeping talk respectful or communicating high expectations, are appropriate for every lesson, whereas those in the Instruction domain, such as assessing student work or engaging in discussions, are not always relevant to a specific lesson. Furthermore, relational skills needed for a high Classroom Environment score may be less difficult to develop than comparably rated skills in the Instruction domain (e.g., matching the cognitive challenge of higher-order questions to students' ability).

Finally, our results provide evidence to refute the claim of the FFT's content independence. Specifically, teachers' reading instruction scores were considerably more consistent than for mathematics, with the between-teacher variance in reading exceeding the between-teacher variance in mathematics by more than 11%. Moreover, to establish reliable estimates of kindergarten teachers' instruction, slightly more raters or lessons are needed for mathematics than reading. This is considerably fewer than for evaluating kindergarten science instruction, however; six raters and nine lessons were insufficient for reliable Classroom Environment or Instruction scores in science (Mantzicopoulos et al., 2018). Accordingly, teacher evaluations are likely most reliable, at least in kindergarten, if observations are focused on reading.

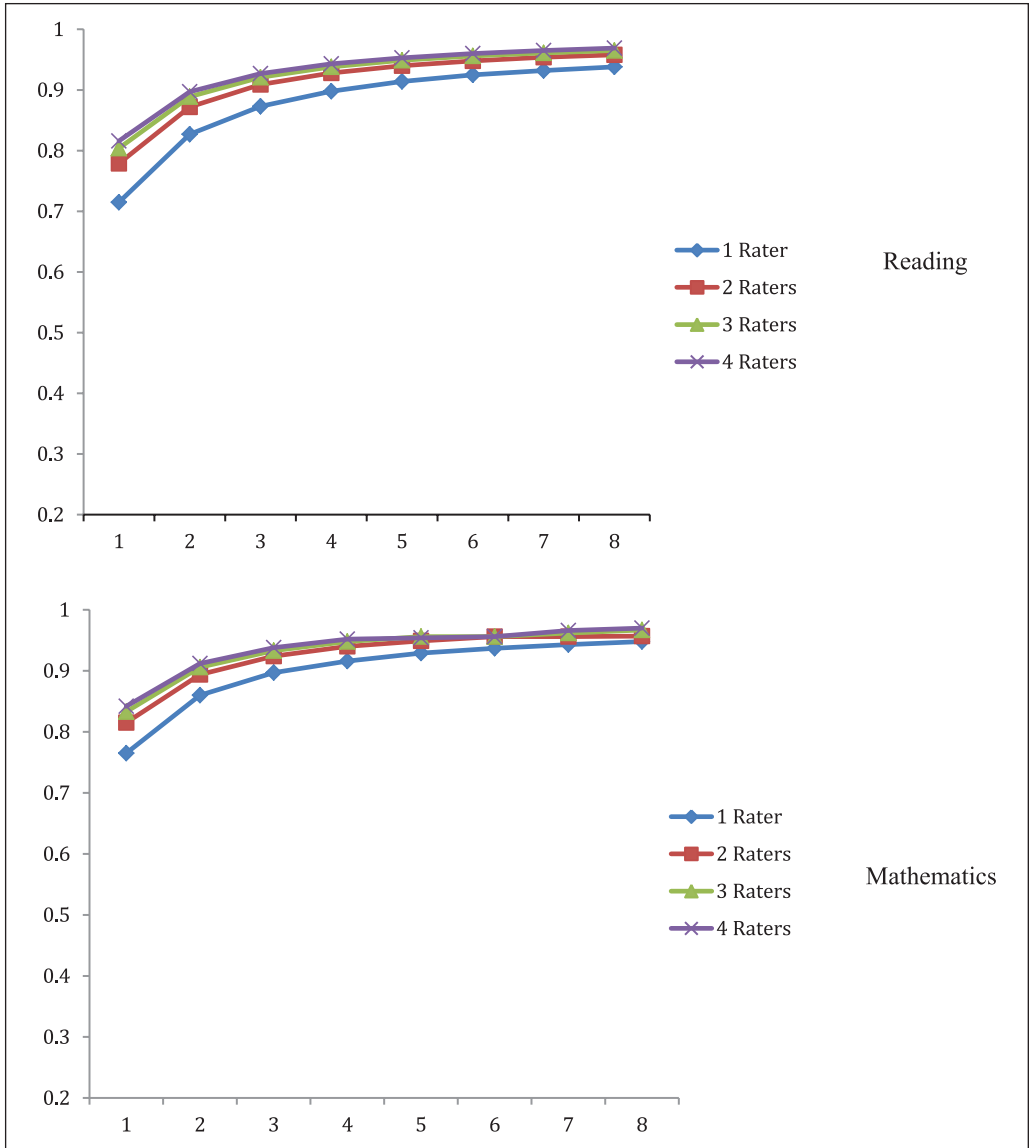


Figure 1. Classroom Environment reliability estimates (G) [on the y-axis] for combinations of number of raters by number of lessons scored [on the x-axis] for reading and mathematics lessons.

Limitations and Directions for Future Research

A potential limitation of our study was the mismatch between our study and teacher evaluation in practice; that our teachers were not being evaluated and our raters were not teacher evaluators may have affected our findings. It is noteworthy, however, that the two MET project FFT G-studies—one with district evaluators scoring lessons (Ho & Kane, 2013) and the other using project researchers to score lessons not used for evaluation (Kane & Staiger, 2012)—found comparable score stability.

Another limitation involves our small sample of teachers—considerably smaller than those of the MET project studies, although larger than comparable analyses with similar instruments (e.g.,

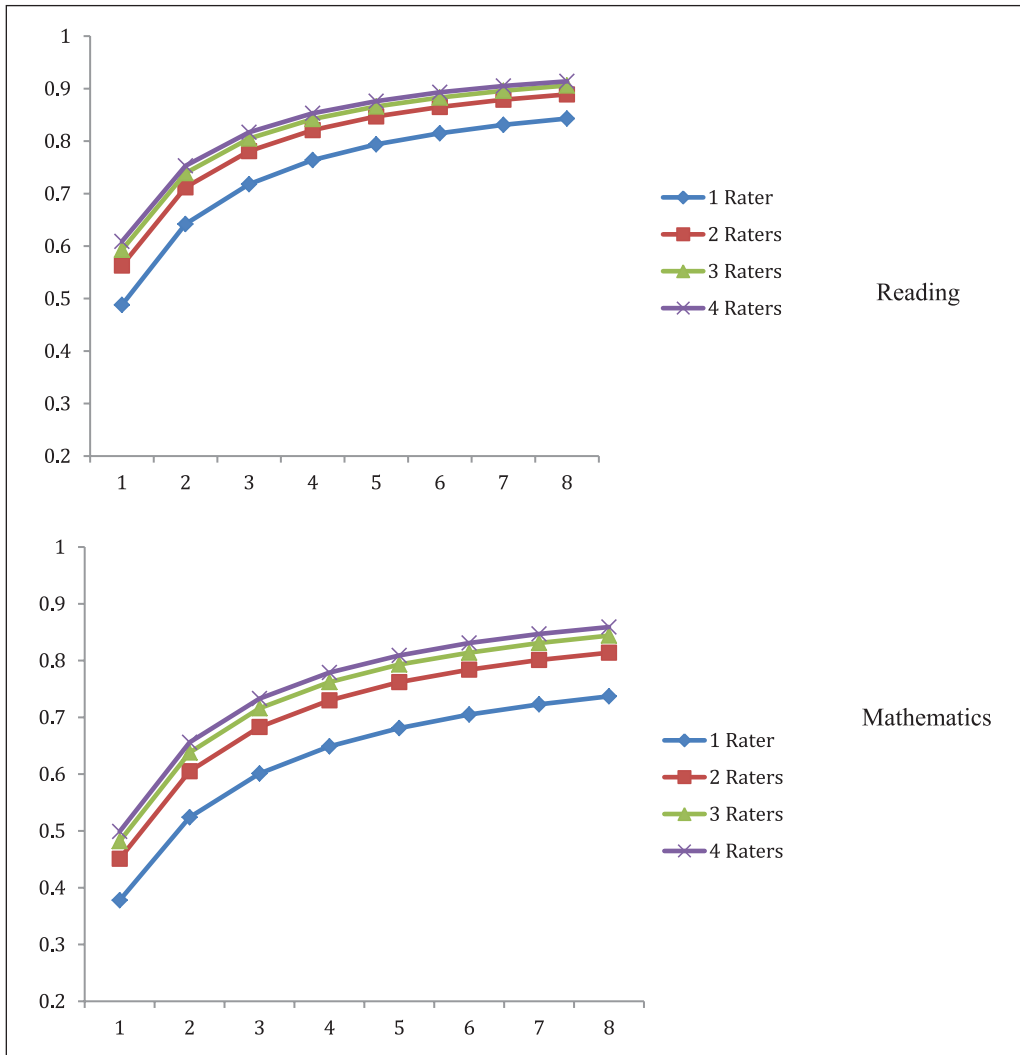


Figure 2. Instruction reliability estimates (G) [on the y-axis] for combinations of number of raters by number of lessons scored [on the x-axis] for reading and mathematics lessons.

eight teachers; Hill et al., 2012). Nevertheless, the small sample size may limit the generalizability of our findings.

Our study examined only kindergarten classrooms, which is both a strength and a limitation. Evidence to support the claim that the FFT is equally appropriate for all grades between kindergarten and 12th grade (Danielson, 2013) is needed. This requires that researchers examine grade levels separately, rather than aggregating scores across grade levels, as is typical (e.g., Kane & Staiger, 2012; Polikoff & Porter, 2014; Tyler et al., 2010); combining grades undoubtedly masks differences. Accordingly, our results reflect kindergarten classrooms but not those in other grades, where similar research is needed.

Another area where further research is crucial involves addressing whether FFT scores differ for various content areas, particularly when they involve the same teacher and students. Many

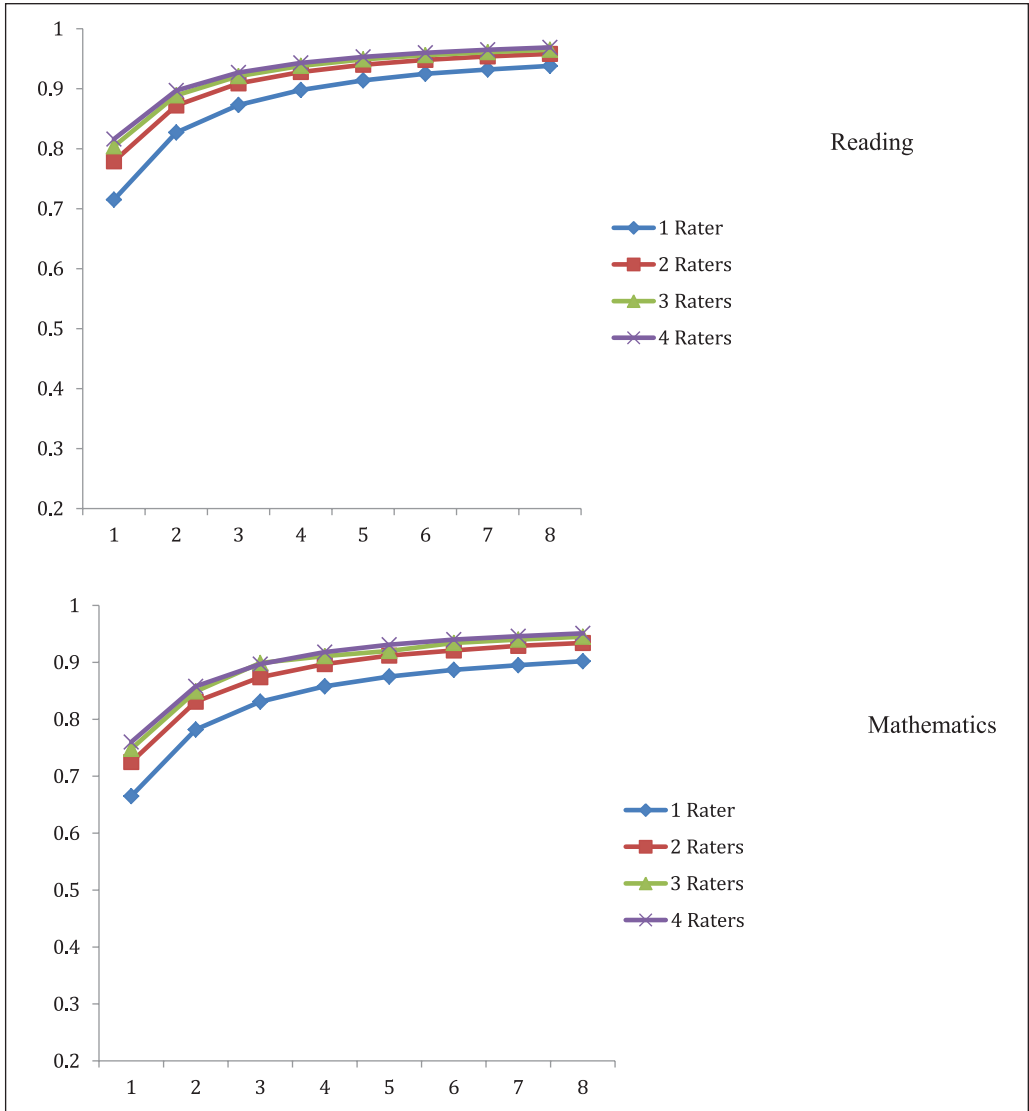


Figure 3. Total FFT reliability estimates (G) [on the y-axis] for combinations of number of raters by number of lessons scored [on the x-axis] for reading and mathematics lessons. Note. FFT = Framework for Teaching.

researchers aggregated scores from different content areas (e.g., Kane & Staiger, 2012; Tyler et al., 2010); however, like with grade level, claims that the FFT is content independent must be scrutinized empirically rather than assumed to be correct. It is also possible that content area differences vary among grade levels; this is a further issue for investigation.

Finally, questions about grade level and content area equivalence apply to other observation instruments that are used to evaluate teachers, beyond the FFT. There is startlingly little published psychometric evidence involving most observation instruments approved for teacher evaluation, particularly in the early elementary grades. This documentation is necessary in order for any evaluation system to be reliable, fair, and transparent.

Acknowledgment

We greatly appreciate the involvement of the kindergarten teachers and their students. We also thank Inok Ahn, Chorong Lee, Qian Li, ChangChia James Liu, Yaheng Lu, Hyejeong Oh, and Sam Watson for their assistance.

Declaration of Conflicting Interests


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

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A140664 to Helen Patrick, Panayota Mantzicopoulos, and Brian F. French. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

ORCID iDs

Helen Patrick  <https://orcid.org/0000-0001-7352-1729>

Brian F. French  <https://orcid.org/0000-0002-3896-7888>

References

- American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*.
- Brennan, R. L. (2001). *Generalizability theory*. Springer.
- Brennan, R. L. (2011). *Using generalizability theory to address reliability issues for PARCC assessments: A white paper*. Center for Advanced Studies in Measurement and Assessment, University of Iowa.
- Center on Great Teachers and Leaders. (2013). *Databases on state teacher and principal evaluation policies*. <http://resource.tqsource.org/stateevaldb/Compare50States.aspx>
- Cohen, J., & Goldhaber, D. (2016). Building a more complete understanding of teacher evaluation using classroom observations. *Educational Researcher, 45*, 378–387.
- Cronbach, L. J., Gleser, G., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. Wiley.
- Danielson, C. (2007). *Enhancing professional practice: A framework for teaching* (2nd ed.). Association for Supervision and Curriculum Development.
- Danielson, C. (2013). *The Framework for Teaching evaluation instrument* (2013 ed.). The Danielson Group.
- Garrett, R., & Steinberg, M. P. (2015). Examining teacher effectiveness using classroom observation scores: Evidence from the randomization of teachers to students. *Educational Evaluation and Policy Analysis, 37*, 224–242.
- Grossman, P. L., Stodolsky, S. S., & Knapp, M. S. (2004). *Making subject matter part of the equation: The intersection of policy and content*. Center for the Study of Teaching and Policy, University of Washington.
- Growth Through Learning. (n.d.). *Teachscape FAQ*. <https://growththroughlearningillinois.org/Support/TeachscapeFAQ.aspx>
- 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*, 56–64.
- Ho, A. D., & Kane, T. J. (2013). *The reliability of classroom observations by school personnel*. Bill & Melinda Gates Foundation.
- Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching*. Bill & Melinda Gates Foundation.
- Lomax, R. G., & Hahs-Vaughn, D. L. (2012). *An introduction to statistical concepts*. Routledge.
- Mantzicopoulos, P. Y., French, B. F., Patrick, H., Watson, J. S., & Ahn, I. (2018). The stability of kindergarten teachers' effectiveness: A generalizability study comparing the Framework for Teaching and the Classroom Assessment Scoring System. *Educational Assessment, 23*, 24–46.

- 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*, 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*, 227–243.
- Mihaly, K., & McCaffrey, D. F. (2014). Grade-level variation in observational measures of teacher effectiveness. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing teacher evaluation systems: New guidance from the Measures of Effective Teaching project* (pp. 9–49). Jossey-Bass.
- National Council on Teacher Quality. (2019). Frequency of evaluation and observation national results. *State Teacher Policy Database* [Data set]. <https://www.nctq.org/yearbook/national/Frequency-of-Evaluation-and-Observation-95>
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.
- Pennsylvania Department of Education. (2014). *Guiding questions for evaluator and certified school psychologist*. <https://www.education.pa.gov/Documents/Teachers-Administrators/Educator%20Effectiveness/Non-Teaching%20Professionals/Guiding%20Questions%20For%20Certified%20School%20Psychologist.pdf>
- Polikoff, M. S., & Porter, A. C. (2014). Instructional alignment as a measure of teaching quality. *Educational Evaluation and Policy Analysis, 36*, 399–416.
- Praetorius, A., Lenske, G., & Helmke, A. (2012). Observer ratings of instructional quality: Do they fulfill what they promise? *Learning and Instruction, 31*, 2–12.
- Praetorius, A., Pauli, C., Reuseer, K., Rakoczy, K., & Klieme, E. (2014). One lesson is all you need? Stability of instructional quality across lessons. *Learning and Instruction, 31*, 2–12.
- Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. SAGE.
- Steinberg, M. P., & Kraft, M. A. (2017). The sensitivity of teacher performance ratings to the design of teacher evaluation systems. *Educational Researcher, 46*, 378–396.
- Swiss Society for Research in Education Working Group. (2006). *EduG user guide*. Institute for Educational Research and Documentation.
- Teachscape. (n.d.). *Framework for Teaching proficiency system*. <https://growththroughlearningillinois.org/Support/TeachscapeFAQ.aspx>
- Tyler, J. H., Taylor, E. S., Kane, T. J., & Wooten, A. L. (2010). Using student performance data to identify effective classroom practices. *American Economic Review: Papers and Proceedings, 100*, 256–260.
- U.S. Department of Education. (2009). *Race to the Top program: Executive summary*.
- U.S. Department of Education. (2011). *Fact sheet: Bringing flexibility and focus to education law*.
- Webb, N. M., Shavelson, R. J., & Haertel, E. H. (2006). 4 reliability coefficients and generalizability theory. *Handbook of Statistics, 26*, 81–124.
- Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness*. The New Teacher Project.