The Mathematical Quality of Instruction (MQI) in Kindergarten: An Evaluation of the Stability of the MQI Using Generalizability Theory

Panayota Mantzicopoulos

Purdue University

Brian F. French

Washington State University

Helen Patrick

Purdue University

Mantziopoulos, P. Y., French, B. F., & Patrick, H. (2018). The Mathematical Quality of Instruction (MQI) in kindergarten: An evaluation of the stability of the MQI using Generalizability theory. *Early Education and Development, 29*, 893-908.

Acknowledgement

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.

Abstract

Research Findings: We evaluated the score stability of the Mathematical Quality of Instruction (MQI), an observational measure of mathematics instruction. Three raters each scored, independently, 100 video-recorded lessons taught by 20 kindergarten teachers in the spring. Using Generalizability theory analyses, we decomposed the MQI's score stability into potential sources of variation (teachers, lessons, raters, and their interactions). The 13-item (three-domain) Ambitious Mathematics Instruction scale and the Whole Lesson scale each explained about one-third of the variance attributed to differences in the main construct of interest (teachers' instructional strategies). The MQI's Errors and Imprecision scale was not relevant at the kindergarten level; there were virtually no errors and/or ambiguities observed across the 100 mathematics lessons. In a series of Decision studies, we examined improvements in reliability with combinations of up to 6 raters and 8 lessons. Only the Richness of Mathematics domain scores and the Whole Lesson scores achieved acceptable reliabilities.

Practice or Policy: The findings have important implications for the use of observation measures to document teachers' mathematics practices in the early years of school.

Keywords:

Mathematical Quality of Instruction (MQI), Teaching Effectiveness Measures,
Teacher Accountability, Observation Measures of Instruction, Generalizability Theory

The Mathematical Quality of Instruction (MQI) in Kindergarten: An Evaluation of the Stability of the MQI Using Generalizability Theory

Evidence that early mathematics competencies consistently predict later mathematics achievement (Watts, Duncan, Clements, & Sarama, 2017) underscores the importance of mathematics education in the early grades of school. There is growing recognition that mathematics deficits begin early and have dire consequences for children's mathematics learning both at the start of school and in the long term (Clements, 2004; Hatchey, 2013; National Research Council, 2001). Thus, attention to the quality of instruction in early mathematics classrooms is crucial as are assessments that document this quality. Unfortunately, though, there is extremely little empirically-supported guidance for documenting the effectiveness of teachers' mathematics practices at the start of school.

The quality of mathematics practices may arguably be addressed by federal accountability policies designed to ensure that all students receive high quality instruction (e.g., Every Student Succeeds Act [ESSA] of 2015; U.S. Department of Education [USDOE], 2011). Driven by these policies, content-general observation protocols are now standard practice across the United States for evaluating teachers' effectiveness (National Council on Teacher Quality, 2015). Widely-used protocols, such as the Framework for Teaching (Danielson, 2013) and the Classroom Assessment Scoring System (Pianta, La Paro, & Hamre, 2008), have been developed on the premise that effective instructional practices are common across subject areas. However, arguments also highlight the need for subject-specific measures (e.g., Charalambous & Praetorious, 2018; Kilday & Kinzie, 2009). Because subject-specific measures are constructed with explicit attention to the norms and practices of a particular content area, these measures may be sensitive to subject-specific aspects of instruction that are not captured by their content-general counterparts. Evidence from a study in pre-kindergarten, where 8 teachers implemented a mathematics curriculum, suggests that a mathematics-specific measure may provide “a more fine-grained analysis of mathematics instructional quality than a content-general measure” (McGuire, Kinzie, Thunder, & Berry, 2016, p. 104). Although preliminary, this finding supports calls for research on mathematics-specific observation protocols (Charalambous & Praetorious, 2018; Kilday & Kinzie, 2009).

One issue that is particularly crucial for assessing effective teaching through observations is the extent to which observation-based scores signal stable differences among teachers. Despite some research on the stability of mathematics instruction in the middle and high-school grades (e.g., Kane & Staiger, 2012; Hill, Charalambous, & Kraft, 2012), there is no published evidence about variability in the quality of teachers’ mathematics strategies in the early school years.

In response to this critical issue, we examine the stability of the Mathematical Quality of Instruction (MQI), a measure developed to assess instruction from kindergarten through the middle school grades (Hill, 2011). To date, the MQI has been used with upper elementary and middle-school mathematics classrooms (Bill & Melinda Gates Foundation, 2013; Kane & Staiger, 2012), however we could not find published data on its use in the early grades. Therefore, our purpose is to contribute evidence with data from kindergarten classrooms. Specifically, we evaluate 100 mathematics lessons taught by 20 kindergarten teachers (5 lessons per teacher) in the spring semester. We provide descriptive data on the MQI’s subscales and use Generalizability theory to document the stability of the measure's scores across teachers, lessons, and raters.

The Case for Using Math-Specific Observation Measures to Document Mathematics Instruction

In mathematics instruction, discipline-specific and pedagogical content knowledge and skills are reflected in the teacher's "specialized fluency with mathematical language, with what counts as a mathematical explanation, and with how to use symbols" (Ball, Hill, & Bass, 2005, p. 21). Yet, only a handful of observation protocols have been developed to assess mathematics instruction from this vantage point, including those appropriate for preschool and the early grades (Charalambous & Praetorious, 2018; Kilday & Kinzie, 2009).

Of interest, measures purporting to be suitable for documenting instruction across grade levels, from kindergarten through at least the middle school grades, have not yet been examined empirically in early elementary classrooms (e.g., Reformed Teaching Observation Protocol; Piburn et al., 2000; U-Teach Observation Protocol [UTOP]; Walkington, Arora, Ihorn et al., 2012; MQI; Hill et al., 2012). Considering the crucial role of early mathematics instruction for student's school success, however, it is imperative that researchers provide this information.

In this study we examine the MQI for several reasons. First, the MQI specifically addresses "the nature of the mathematical content available to students during instruction" (Learning Mathematics for Teaching Project, 2011, p. 30) rather than broad features of the classroom environment such as climate, emotionally supportive practices, or behavior management.

Second, the MQI's framework is aligned with the view that "teaching that fosters the development of mathematical proficiency … can take a variety of forms" (National Research Council, 2001, p. 315). Therefore, the measure is intended to document the mathematical content of a lesson without privileging particular reform curricula or approaches to teaching (e.g., inquiry-driven vs. traditional, child-centered vs. didactic; Learning Mathematics for Teaching Project, 2011). This is in contrast to other measures (e.g., Reformed Teaching Protocol; Piburn et al., 2000; Inside the Classroom Observation and Analytic Protocol; Horizon Research, 2003) that were developed in response to mathematics reform efforts and thus may be less well suited for documenting the quality of a variety of instructional approaches, ranging from mainstream to inquiry-driven.

Third, the MQI was part of the high-profile Measuring Effective Teaching (MET) project (Bill & Melinda Gates Foundation, 2013), which generated evidence on the measure's score reliability in grades 4-8. Additional data on the stability of the MQI have also been reported with a small group of middle school teachers (Hill et al., 2012). However, as noted earlier, data on the psychometric qualities of the MQI in the early grades of school are lacking. Thus, our study's goal is to provide evidence on the stability of the MQI's scores at the start of school. In particular, in addition to evaluating the variability of scores across lessons, teachers, and raters, we address questions about the optimal number of raters and lessons needed to provide accurate estimates of kindergarten teachers' mathematics practices.

Representing Teachers' Mathematics Instruction Reliably

Regardless of the observation assessment used, evidence suggests that multiple observations are needed for high quality information on teachers' instruction (Kane & Staiger, 2012; Whitehurst, Chingos, & Lindquist, 2014). Nonetheless, developers of existing measures provide little to no guidance on the number of observations necessary to achieve stable estimates of teachers' practices.

This information is crucial in light of evidence that the stability of observation-based scores is dependent on the instructional domain being assessed (Curby et al., 2010, 2011; Grimm et al., 2008; Kane, McCaffrey, Miller, & Staiger, 2013; Patrick & Mantzicopoulos, 2016). That is, classroom climate, classroom management, and teachers' emotionally supportive strategies may be reliably documented with fewer observations than instruction-focused strategies involving meaningful questioning, scaffolding, and engaging students in higher-level thinking (e.g., Ho & Kane, 2013; Praetorius, Pauli, Reusser, Rakoczy, & Klieme, 2014). The latter strategies are not only variable from lesson to lesson, but also require multiple points of observation (e.g., $n = 9$) – compared to one or two occasions for classroom management – in order to reflect individual teachers' instruction with high levels of accuracy (Praetorius et al., 2014).

Based on the studies just mentioned, we expect that reliable estimates for the MQI – like other instruction-focused measures – would require multiple observers and lessons. This was borne out in research using the MQI with middle school teachers (Hill et al., 2012). Specifically, a combination of two raters and three lessons reached reliabilities that were acceptable for research purposes (i.e., > .80). However, the reliability estimates were not high enough (i.e., ≥ .95) to support use of the MQI to reliably identify the effectiveness of individual teachers (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014).

It is not clear whether findings about the number of observations necessary to produce stable scores generalize to all grade levels. More observations may be required for reliable scores in kindergarten, when children are just learning what the student role involves, compared to the later grades, when students have become well-acculturated to school. Whether or not the results from middle and high-school grades are replicated with our sample of kindergarten teachers is an issue we explore in this study.

Identifying the Variability in MQI Scores: Unpacking Teacher, Rater, and Lesson Effects

We use Generalizability– (G) theory to identify the unique and joint contributions of factors that are implicated in score variability (Brennan, 2001; Shavelson & Webb, 1991). G-theory permits us to delineate, in an additive manner, the proportion of variance accounted for by teachers, raters, lessons, and the interactions between these components. With instructional strategies being the construct of interest, we expect that measures characterized by high levels of stability will yield scores that identify reliable differences across teachers. In other words, variation among teachers should, ideally, be the largest variance component identified in the analyses. Sources of variance that are irrelevant to the target construct (e.g., variance due to differences in raters' scoring) should yield small estimates, indicating that extraneous sources of bias are kept to a minimum (i.e., rater behaviors do not systematically influence teacher scores). Low levels of variance from lesson to lesson would provide support for the undergirding assumption of teacher accountability policies: Namely, that a teacher's effectiveness score is stable regardless of a particular lesson's characteristics or the time (i.e., occasion) of observation.

G-theory has been used in research with the MQI with upper elementary and middle school classrooms (Hill et al., 2012; Kane & Staiger, 2012), in addition to other observation measures of instruction across different grade levels (Mashburn, Meyer, Allen, & Pianta, 2014; Praetorius et al., 2014). Reliability estimates from these studies, however, indicated that more observations than are usually conducted are required to confidently measure the quality of individual teachers' practices (AERA, APA, & NCME, 2014). Improvements in score reliability have been noted with successive increases in the number of raters, occasions, or both.

G-study estimates, Decision studies (D-Studies; e.g., Hill et al., 2012) contribute additional information on the number of raters and/or observations needed to yield sufficiently stable scores for addressing instructional effectiveness at the individual teacher level. Thus, using the results from the G-study, we also implement a series of D-studies to examine changes in reliability with successive increases in the number of raters and/or occasions.

Summary of Research Aims

We document the MQI's (Hill, 2014) score stability using 100 video-recorded kindergarten math lessons (5 lessons from each of 20 teachers). Each lesson was scored with the MQI by three raters. We conduct a series of G-theory studies to decompose the MQI score reliability into potential sources of variance (teachers, lessons, raters, and their interactions). Next, we conduct D-studies in order to detect improvements in reliability with different combinations of the number of lessons and raters.

Method

Participants

Teachers and Schools. All kindergarten teachers (*N* = 25) in 6 public schools and 4 districts in the midwestern United States were invited to participate. A total of 22 (88%)

teachers provided informed consent. This sample comprised all kindergarten teachers in 3 of the schools and all but 3 teachers in the remaining schools (i.e., one teacher in each of these 3 schools declined participation). However, of the 22 teachers who consented, two teachers were not included because they recorded fewer than the 5 lessons required for inclusion in this study**.** Thus, the final sample of 20 teachers (19 females, 1 male) represented 80% of the kindergarten teacher population in the 6 schools. Eighteen teachers were White; two teachers were Hispanic. Their experience ranged from 1 to 33 years (*M* = 16 years).

To provide a context for the classrooms and schools, we note that we received informed consent for 324 students, representing 79.4% of the kindergarten population in the 20 teachers' classrooms. These students were socioeconomically diverse, with 53.3% receiving free- or reduced-cost lunch. With respect to ethnicity, 63.2% of the students were White; 22.8% were Hispanic, 9.1% were Black, and 5.0% were classified as Other. The schools' report card grades, representing each school's ranking in the statewide assessment system, ranged from A to C. The percent of students eligible for free or reduced-cost lunch ranged from 30.3 to 72.8 across schools.

Lessons

We asked teachers to videotape a regular kindergarten mathematics lesson once a week, for a period of 10 weeks during the spring semester. Teachers chose the day and lesson to record.

We obtained a pool of 211 mathematics lessons, from which we randomly selected 5 lessons per teacher to analyze for this study. These lessons (average lesson length = 24 minutes) were evenly spread through the spring semester and represented typical, public school instruction in kindergarten.

Teacher Measures and Procedure

Mathematical Quality of Instruction (MQI; Hill, 2014). We used the most recent, 2014 version of the MQI (Center for Education Policy Research, n.d.). It involves two sets of scales rated in two separate phases, each with its own procedure and format for scoring lessons. Psychometric information on the 2014 MQI or its previous version is not available for the early grades.

Lesson segment scales. Lessons are first divided into 7½ -minute cycles in order keep rater cognitive load at appropriate levels and in response to "feedback that segments [exceeding this amount of time are] difficult to score" (Hill et al., 2012, p. 58). After viewing each segment, raters stop to first rate a dichotomously scored item (1 = yes; 0 = no) of "whether the focus is on *mathematical* c*ontent* [italics in original] during half or more of the segment" (Hill, 2014, p. 3). Raters then use a 4-point scale (0 = not present, 1= Low, 2 = Mid, 3 = High) to score the teacher's mathematics instruction across four domains, each of which comprises a different number of items. *Richness of Mathematics* includes: (a) 3 items that measure "the extent to which instruction focuses on the *meaning of facts and procedures* [italics in original]*,"* and (b) 3 items that document the use of "*key mathematical practices*" [italics in original] (use of multiple procedures, drawing generalizations, and mathematical language) during instruction (Hill, 2014, p. 4). *Working with Students and Mathematics* items (*n* = 2) assess teachers' responses to students' mathematical contributions, including student questions as well as explanations (both correct and incorrect). *Errors and Imprecision* items (*n* = 3) capture content errors and lack of clarity in the teachers' mathematical language. *Common Core Aligned*

Student Practices (5 items) assesses students' meaningful engagement as reflected in the Common Core State Standards for Mathematics; it includes effortful, sustained, and meaningful engagement with mathematics, reasoning and argumentation, communication of mathematical ideas, use of mathematical tools, and working with meaningful problems (Hill, 2014).

Raters also score each of the four domains with a single *Holistic item* that reflects the overall quality of the practices in that domain. Scores across segments are averaged to create item scores for each lesson; item scores within each domain are then averaged to create *Domain scores*.

Factor analysis with an earlier version of this set of items (excluding the Holistically-scored items) indicates that the MQI has a two-factor structure (Blazar, Braslow, Charalambous, & Hill, 2017). Items on the *Errors and Imprecision* domain form one factor, whereas the 13 items on the remaining scales form a second factor (i.e., *Ambitious Mathematics Instruction*). On that basis, we also computed an Ambitious Math Instruction score by averaging ratings across the 13 items that make up the Richness of Mathematics, Working with Students, and Common Core Aligned Practices scales. Additionally, we created an overall *Holistic* score for Ambitious Math Instruction by averaging the holistic item scores associated with each domain.

Whole lesson scale. At the conclusion of each lesson (i.e., after raters have viewed and scored all 7½ minute segments), the entire lesson is rated with the *Whole Lesson* scale (Center for Education Policy Research, n.d.b). This scale consists of 9 items and 1 holistic item, each scored on a 5-point scale (1 = not at all true of this lesson, 5 = very true of this lesson). The scale's 9 items document the extent to which: mathematics is covered in depth, is meaningful and free of errors, students are engaged, the teacher makes efficient use of time during the lesson, attends to student difficulties, and uses students' ideas in the development of the lesson. The 10th item is a holistic assessment of the overall quality of the lesson; it is not added to the 9-item Whole Lesson scale but is kept as a single *Holistic item*.

Rater training and rater agreement. Three graduate student researchers with significant experience in (a) early education settings, and (b) documenting instruction through observations completed the on-line training for the MQI, administered by the Center for Education Policy Research (n.d.a). They also successfully passed the certification test, administered by the Educational Testing Service, and required for rating mathematics instruction in grades K-9. Successful performance on the test is contingent on raters viewing and coding four videos, each approximately 20-minute long, and achieving agreement with MQI master coders. The level of agreement for raters with the MQI master coders is calculated in terms of the distance from the master coders' score (Hill et al., 2012). Specific information on the procedure for calculating inter-rater agreement is not provided, however.

In the absence of official guidelines for establishing rater agreement (for monitoring raters after certification), and to examine rater calibration/agreement data for our group of raters, we developed a process analogous to that used with other observation measures. Prior to scoring the videos for the present study, the raters watched 10 kindergarten mathematics lessons that were not part of this study. After each lesson, they calculated exact agreement for ratings of segment-level Domain ratings (scored 0-3 every 7½ minutes), the set of 9 Whole Lesson items (scored 1-5 at the end of the lesson), and the associated Holistic item (also scored 1-5 at the end of the lesson). The average exact agreement across pairs of raters was 72%, 58%, and 69%, respectively. For comparison purposes, inter-rater agreement reports (based on exact agreement derived from rater calibration activities post MQI certification) for the MET study ranged from 53.4% to 76.6% (Bell et al., 2014).

Procedures for observing and scoring lessons. Lessons were rated consistent with the MQI's protocol. A scoring schedule was created by numbering lessons and, for each rater, we used a random number generator to assign lessons. Thus, raters independently observed and rated lessons from each of the 20 teachers, but not sequentially or grouped by teacher. Each rater was blind to other raters' scores.

Analysis Plan

G-theory model. We used a two-facet (lessons, raters) partially-nested random design to decompose variance in the MQI scores (i.e., Crocker & Algina, 1986; Shavelson & Webb, 1991). Teachers, the object of measurement, were observed teaching math lessons on different occasions and, as occurs in practice, their lessons were not identical. Therefore, we nested lessons within teachers. Unfortunately, this results in the main effect of lesson not being estimated because the lesson effect is confounded with the Teacher x Lesson interaction effect (Brennan, 2001). Also, because each rater scored each teacher on all 5 lessons, the rater facet was crossed with lessons (i.e., Rater x Lesson interaction effect).

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

For each domain of the MQI, the variance (σ^2_t) was partitioned into the following components:

1. Teacher (t, σ^2_t): Variance attributed to differences across teachers.
2. Rater (r, σ^2_r): Variance attributed to differences across raters.
3. Teacher x Rater (t × r, σ^2_{tr}): Variance attributed to inconsistencies between raters in evaluating a particular teacher's strategies.
4. Lesson: Teacher (l:t, $\sigma^2_{l,tl}$): Variance attributed to inconsistencies in teacher strategies from lesson to lesson. As we noted earlier, the main effect of lesson is confounded with the t x l interaction effect. Therefore, we cannot determine if the rating of a teacher differed from one lesson to another (interaction effect) or if a lesson resulted in different strategies than another lesson.
5. Lesson: Teacher × Rater (l:t × r, e, $\sigma^2_{rl,trl,e}$): Residual variance comprising unmeasured effects and random events affecting the measurement.

Consistent with models used in other studies of observation measures of instruction (Hill et al., 2012; Mashburn et al., 2014), the teacher, rater, and lesson effects were random, rather than fixed. 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, the Lesson facet in our study is a sampling of mathematics lessons typically taught by public school teachers during the spring of kindergarten.

We used the EduG software package (Swiss Society for Research in Education Working Group, 2006) to estimate the models. We report the five variance components and their standard errors, as recommended (Lakes & Hoyt, 2010), 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, with higher estimates representing increases in the dependability of the measurement procedure (Shavelson & Webb, 1991). This index can be interpreted like coefficient alpha (Cronbach, Gleser, Nanda, & Rajaratnam, 1972), and used when relative decisions about teachers are being made, such as the relative ranking of their performance in the classroom given the observation score.

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 scores used for research purposes (Nunnally & Bernstein, 1994). Accordingly, estimates at or above 0.95 are needed to provide evidence of score stability for decisions at the individual teacher level, whereas estimates greater than 0.80 are acceptable for basic research (e.g., aggregating scores across teachers for predicting outcomes, or judging teacher training efforts) (AERA, APA, & NCME, 2014).

Results

Descriptive Statistics & Internal Consistency Estimates for the MQI

Distribution of scores and MQI scale intercorrelations. The activities observed in 82% of the lessons were directly connected to mathematics (as indicated by raters' scores on the item intended to document the percent of time that mathematics content was present in each lesson segment). Teachers' means on the four domains, the Ambitious

Mathematics Instruction scale (i.e., aggregate of three domains), and the Whole Lesson scale are shown in Table 1. Scores for Errors and Imprecision indicated that during math instruction teachers presented the concepts clearly and did not make content errors. However, average scores for the remaining three scales (Richness of Mathematics, Working with Students, and Common Core Aligned Practices) were less than 1, suggesting that the quality of mathematics practices ranged from *not present* to *low*. In terms of the four single, holistic domain items (not shown in Table 1), the average scores for three were in the low-quality range (i.e., .90 -- Richness, 1.03 -- Working with Students, and .49 -- Common Core Aligned Practices); the exception was the Errors and Imprecision holistic item (*M* = .02).

In contrast to the low scores assessed in the 7½ minute segments, end-of-the-lesson scores on the Whole Lesson scale (*M* = 3.09) and its associated holistic item (*M* = 2.80) indicated that the quality of instruction was in the mid-range. Consistent with the MQI criteria, scores in this range are considered characteristic of typical instruction (Center for Education Policy Research, n.d).

Except for Errors and Imprecision, the intercorrelations between domain scores, including scores on the Ambitious Math composite (Table 1) were statistically significant and ranged from .52 to .0.93. The correlations between each domain score and its associated Holistic item (not included in Table 1) were > .91.

Rater accuracy. Even though rater accuracy is evaluated in the G-theory analyses, we include ICC estimates because these permit comparisons with estimates reported for the upper elementary grades (e.g., Bell, et al., 2014). As shown in Table 2, other than Errors and Imprecision, the average ICC estimates were high—at least 0.81 for domain scale scores and at least 0.79 for domain holistic items. The ICC for the Errors and Imprecision domain (0.43) and its associated holistic item (0.23) were low—most likely due to the lack of score variability, rather than significant rater discrepancies.

Internal consistency estimates. As shown in Table 2, internal consistency reliability estimates (α coefficients) for the domains reflecting segment-by-segment ratings ranged from 0.55 (Errors) to .83 (Richness). The reliability of the Whole Lesson scale, which reflects the entire lesson was high (α =0.91). The alpha for all four holistic items (one per domain) was 0.80. Alphas for the 16-item MQI scale and the 13-item Ambitious Mathematics Instruction scale were 0.87 and 0.83, respectively.

Generalizability (G-) Theory Estimates

The relative stability (G-) estimates on the MQI domains ranged from 0.05 (Errors and Imprecision) to 0.78 (Richness of Mathematics). The G-estimate for the Whole Lesson scale was 0.84. These estimates vary widely across domains and are below recommended standards. The G-estimates for the domain-specific Holistic items ranged from 0.71 (Richness of Mathematics) to 0.00 (Errors and Imprecision); the latter is associated with a negative variance estimate. The Whole Lesson holistic estimate was 0.73. Of note, for each domain the estimate for the holistic items is lower than the G estimates for the respective domain score.

The decomposition of variance estimates are summarized in Table 3. Despite low levels of rater variance (< 6%), the MQI score variability is: (a) not comparable across domains of math instruction, and (b) only partially attributable to stable differences across teachers. In each of the four domains, the holistic item identified less between-teacher variability than did the corresponding domain scale scores.

The largest portion of variance attributed to teachers was identified within the Whole Lesson scale (38%). The between-teacher variance accounted for by scores on the Ambitious Mathematics Instruction scale was 27%. Of the domains that make up this scale, Richness of Mathematics yielded the largest portion (34%) of variance attributed to stable teacher differences. The remaining two domains (Working with Students and Common Core Aligned practices) accounted for small portions of teacher relevant variance (14% and 13%, respectively). Teacher variance in Errors and Imprecision was negligible.

Between-lesson variability (within teachers) ranged from 22% to 40% across the MQI domains and scales. The residual variance in domain scores not accounted for by the facets we examined was also high, ranging from approximately 30% for Richness in Mathematics to 70% for Errors and Imprecision.

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

The results for each domain's average score are shown in Figure 1 (Panels A through F). Only the scores for the Richness of Mathematics domain (Panel A) and the Whole Lesson scale (Panel F) achieved at least 0.80 reliability. This level occurs first with 3 raters scoring 5 lessons in the Richness domain, and 3 raters scoring 5 lessons with the Whole Lesson scale. Across all conditions examined (i.e., up to 6 raters and 8 lessons), estimates for the other domains were far below accepted reliability standards (AERA, APA, & NCME, 2014).

Discussion

The Quality of Mathematics Instruction in Kindergarten

Teachers in this study did not make mathematical errors and presented concepts clearly and unambiguously. However, the overall quality of the mathematics lessons that we evaluated, on each of the remaining subscales of the MQI including Ambitious Mathematics Instruction, did not exceed scores in the “low” category. Our findings parallel those reported in other studies with upper elementary (Blazar, 2015) and middle school classrooms using an earlier version of the MQI (Bell et al., 2014). In the MET study over 70% of the lessons taught by middle-grade teachers were rated on the low end of the MQI scale (Kane & Staiger, 2012).

Perhaps these results reflect the concern that teachers, and especially those in the elementary grades, are ill-prepared to teach mathematics (e.g., National Council on Teacher Quality, 2015). However, prior to drawing this inference, it is important to consider additional evidence about early mathematics instruction and curricula. Based on teacher reports, data from the ECLS-K (Early Childhood Longitudinal Study) indicate that teachers in the early grades of school tend to emphasize factual knowledge as well as daily drill and practice as a way to increase procedural fluency and support mathematics learning (Bottia, Moller, Mickelson, & Stearns, 2014). Additional evidence underscores the effectiveness of promoting numerical fluency in the early grades through the provision of daily practice with mathematical concepts and facts (Agodini & Harris, 2016; Aunola, Leskinen, Lerkkanen, & Nurmi, 2004; Guarino, Dieterle, Bargagliotti, & Mason, 2013). These activities, however, do not represent aspects of high quality instruction, at least as measured by the MQI. And, although they may contribute to achievement in the short-term, it’s not clear that they promote conceptual understanding and mathematics reasoning, particularly when they are used in isolation from real-world mathematics activities and meaningful problem solving.

Thus, on one hand, our findings converge with evidence that early childhood teachers need considerable support in order to adopt practices that promote children's use of mathematical language, the ability to connect mathematics concepts, and to engage in mathematics reasoning (McGuire, et al., 2016). On the other hand, attention to a variety of practices that target the development of number concepts may also be needed in measures of early mathematics instruction. It is well established that the acquisition of number concepts and skills forms the basis for later mathematics competencies and should be explicitly supported during instruction (National Research Council, 2001).

Interestingly, the new 5-point Whole Lesson scale (rated at the end of the lesson) indicated that teachers' average rantings were largely consistent with typical instruction, according to the MQI training document (National Center for Education Policy Research, n.d.). Thus, Whole Lesson scores presented a more positive picture about the quality of the math lessons than did the original domain scores (rated on a 4-point scale in 7½ minute-segments). Items on the Whole Lesson scale map directly onto the content of the segment-coded MQI scales; they measure the adequacy of the mathematical content, connections with big ideas, attention to student ideas and remediation of student difficulties, and use of rich mathematics (representations, explanations, multiple solutions). Thus, it is not clear why results from the main MQI scales and the Whole Lesson scale diverge, portraying kindergarten mathematics instruction as either poor (segment-level scores) or adequate (Whole Lesson scale scores). With the caveats discussed next, the Whole Lesson scale may deserve further attention as a potentially promising component of the MQI.

MQI Score Stability

Our findings regarding the stability of kindergarten teachers' instruction across the four domains support several general conclusions. First, the G-coefficients fell short of recommended standards (i.e., > 0.80) for use in basic research (Nunnally & Bernstein, 1994). Second, as expected, G-estimates were lower for the Holistic item scores than for the corresponding domain average scale score. It is not clear what information these single-item scores contribute above and beyond that provided by their respective domain scale. Third, the four MQI domains are not comparable with respect to the teacher variance explained, which fluctuates from 0.6% (Errors and Imprecision to 33.9% (Richness of Mathematics). Fourth, the significance of the MQI scales may vary by grade level. The Errors and Imprecision scale may be less relevant in the early grades, given that it lacked variability and was of low incidence in our kindergarten sample. For upper elementary and middle school teachers, scores on this domain explain approximately one-third of the variation attributed to differences among teachers (e.g., Hill et al., 2012 and are also predictive of student outcomes (Blazar & Kraft, 2017). Fifth, of the MQI scales, the Whole Lesson scale appears to result in the psychometrically strongest scores, at least in kindergarten. Overall, however, the low G-values suggest that it is difficult to consistently capture teacher behaviors specific to the underlying domains assessed by the MQI.

The decomposition of variance in the MQI scores suggests that the measure does not identify stable differences in kindergarten teachers' mathematical practices. For all but one scale there was greater variation in MQI scores among lessons (22% - 40%) than between teachers' practices (13% - 38%). Despite consistently low levels of rater variance (< 6%), considerable variance (> 30%) remained unexplained by the facets we examined.

The findings in our study are consistent with evidence from the MET project (Kane & Staiger, 2012). Like with our study, teacher-relevant variance was small (6% - 16%), and there was considerable lesson-specific variance (Kane & Staiger, 2012). Our and the MET project's results differ somewhat from Hill et al.'s (2012) study of middle school teachers, where between-teacher variance ranged from 32% to 43%, and lesson-specific variance did not exceed 11%. These differences may be attributable to Hill et al.'s smaller sample (24 lessons; 8 teachers, each with 3 lessons) or the diversity of students and schools in both our study and the MET project. Importantly, there was considerably less rater variance in our study compared to the two mentioned; from 6% to 28% in Hill et al.'s (2012) study, and from 2% to 21% in the MET project (Kane & Staiger, 2012).

The new addition to the MQI—the Whole Lesson scale—emerged as the strongest component of the MQI, capturing the largest amount of the total variance (38%) in the Teacher facet. Yet, although rater variance was negligible (< 1%), lesson-to-lesson variability accounted for almost one-quarter of the total variance. Considering that lessons varied in content, it may be helpful in future research to include these as additional facets to explain the relatively large amount of variance in the residual terms.

Limitations

Potential limitations of our study include (a) its generalizability, (b) the inclusion of a range of kindergarten topics in the lessons, and (c) allowing teacher choice in the selection of the lesson that they submitted to us each week of the semester.

Generalizability limitations may stem from the stringency of our methods, compared to actual practice. We observed the teachers on more occasions than is typical in public schools and we engaged in a great deal of rater calibration activities with local data, after our raters were MQI-certified. Additionally, in rating the lessons we controlled for order effects. In practice, when multiple evaluations of a teacher's lessons are conducted, evaluators would document the teacher's mathematics strategies sequentially, which may yield different stability estimates.

Lack of standardization of the lesson content may also be a limitation. At the same time, content differences across teachers from week to week is an aspect of the study that contributes to its ecological validity. In practice, evaluations of instruction are not content-focused, even when teachers in the same subject area are observed.

A third limitation, representing a departure from practice, may be that teachers selected the lesson that they video-recorded each week. Evidence from the MET study (Ho & Kane, 2013), however, suggests that ratings of teacher-selected lessons, compared to non-selected lessons: (a) are slightly higher but are more reliable; (b) more variable across teachers, highlighting the differences between them; and (c) place teachers in the same effectiveness ranking.

Conclusions

In conclusion**,** the variability of the MQI's scores is only partially explained by differences in kindergarten teachers' instructional practices, even when rater agreement is high, keeping rater variance at low levels. Although the Whole Lesson scale yielded the most favorable estimates across all stability criteria, it is premature to suggest that this scale is preferable to the segment-scored MQI scales. Recall that Whole Lesson items are scored after raters have viewed and scored the lesson, in 7½ minute segments, with the items comprising the four main MQI domains. This labor-intensive practice may increase raters' awareness of teachers' instructional behaviors. Perhaps this awareness develops cumulatively over the course of rating the lesson segments and is reflected in judgements made at the conclusion of the lesson, when raters score the Whole Lesson items. The same results may not be evident if raters score only the Whole Lesson scale. This is an important question to address, given the time needed to score lessons.

Another aspect of the MQI requiring further examination is its internal structure. That kindergarten teachers were unlikely to make mathematical errors and omissions suggests that this scale may be less critical to documenting high quality mathematics instruction in the earliest grades of school. Moreover, it's not clear that the MQI items, as grouped within the four separate domains, provide distinct pieces of information about the mathematical quality of instruction. Factor analysis of an earlier version of the MQI used in the upper elementary grades bears out this conclusion (Blazar et al., 2017) and supports a 2-dimensional structure for the measure (Ambitious Mathematics Instruction and Errors). Research is needed to clarify the dimensionality of the MQI and to confirm its invariance across grade levels.

Our findings have important implications for the documentation of mathematics instructional practices at the start of school, an area that remains significantly underexplored. We argue that the need to document the quality of early mathematics teaching must be accompanied by the development of assessments that can inform instructional practice by: (a) being sensitive to mathematics-specific norms, content, and pedagogy; (b) permitting the accurate documentation of individual teachers' practices;

and (c) contributing evidence about the extent to which the instructional context facilitates the development of children’s mathematics competencies and attitudes. Evidence on measures with an explicit focus on pre-K environments (e.g., Clements, Sarama, Spitler, Lange, & Wolfe) and the early grades of school (e.g., Berry, Rimm-Kaufman, Ottmar, Walkowiak, & Merrit, 2010) is much needed to address the National Mathematics Advisory Panel’s (2008) recommendation for “precise measures” that “uncover in detail the relationships among teachers’ knowledge, their instructional skill, and students’ learning.” (p. 38).

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.

Agodini, R., & Harris, B. (2016). How teacher and classroom characteristics moderate the effects of four elementary math curricula. *The Elementary School Journal*, *117*, 216-236.

Aunola, K., Leskinen, E., Lerkkanen, M., & Nurmi, J. (2004). Developmental dynamics of math performance from preschool to grade 2. *Journal of Educational Psychology, 96,* 699-713.

Ball, D. L., Hill, H. C., & Bass, H. (2005). Knowing mathematics for teaching: Who knows mathematics well enough to teach third grade, and how can we decide? *American Educator, 29*(1), 20-22. 43-46.

Bell, C. A., Qi, Y., Croft, A. J., Leusner, D., McCaffrey, D. F., Gitomer, D. H., & Pianta, R. C. (2014). Improving observational score quality: Challenges in observer thinking. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing teacher evaluation systems* (pp. 50-97). San Francisco, CA: Jossey Bass.

Berry, III, R. Q., Rimm-Kaufman, S. E., Ottmar, E. M., Walkowiak, T. A., & Merritt, E. (2010). *The Mathematics Scan (M-Scan): A measure of mathematics instructional quality*. Unpublished measure, University of Virginia.

Bill & Melinda Gates Foundation (2013). *Measures of effective teaching*. http://www.metproject.org/

Blazar, D. (2015). Effective teaching in elementary mathematics: Identifying classroom practices that support student achievement. *Economics of Education Review, 48*, 16-29.

Blazar, D., Braslow, D., Charalambous, C. Y., & Hill, H. C. (2017). Attending to general and content-specific dimensions of teaching: Exploring factors across two observation instruments. *Educational Assessment, 22*, 71-94.

Blazar, D., & Kraft, M. A. (2017). Teacher and teaching effects on students' attitudes and behaviors. *Educational Evaluation and Policy Analysis*, *39*, 146-170.

Bottia, M. C., Moller, S., Mickelson, R. A., & Stearns, E (2014). Foundations of mathematics achievement: Instructional practices and diverse kindergarten students. *The Elementary School Journal, 115*, 124-150.

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

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 (CASMA). University of Iowa.

Center for Education Policy Research (n.d.a). *Mathematical Quality of Instruction (MQI)*. Retrieved from https://cepr.harvard.edu/mqi-access

Center for Education Policy Research (n.d.a). Whole lesson codes [*Training module for the Mathematics Quality of Instruction*]. Unpublished module. Retrieved from https://cepr.harvard.edu/mqi-access

Charalambous, C. Y., & Praetorious, A. (2018). Studying mathematics instruction through different lenses: Setting the ground for understanding instructional quality more comprehensively. *ZDM: Mathematics Education*, https://doi.org/10.1007/s11858-018-0914-8.

Clements, D. H. (2004). Part one: Major themes and recommendations. In D. H. Clements, J. Sarama, & A. M. DiBiase (Eds.), *Engaging young children in mathematics: Standards for early childhood mathematics education* (pp. 1–72). Mahwah, NJ: Erlbaum.

Clements, D. H., Sarama, J., Spitler, M. E., Lange, A. A., & Wolfe, C. B. (2011). Mathematics learned by young children in an intervention based on learning trajectories: A large-scale cluster randomized trial. *Journal of Mathematics Education, 42*, 127-166.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Philadelphia: Harcourt.

Cronbach, L. J., Gleser, G., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.

Curby, T. W, Grimm, K. J., & Pianta, R. (2010). Stability and change in early childhood classroom interactions during the first two hours of a day. *Early Childhood Research Quarterly, 25*, 373-384.

Curby, T. W., Stuhlman, M., Grimm, K., Mashburn, A., Chomat-Mooney, L., Downer, J., Hamre, B., & Pianta, R. C. (2011). Within-day variability in the quality of classroom interactions during third and fifth grade: Implications for children's experiences and conducting classroom observations. *Elementary School Journal, 112,* 16-37.

Danielson, C. (2013). *The Framework for Teaching evaluation instrument* (2013 ed.).

Retrieved from http://www.teachscape.com/frameworkforteaching/home

Every Student Succeeds Act of 2015, Pub. L. No. 114-95 § 114 Stat. 1177 (2015-2016).

Grimm, K. J., Curby, T. W., Pianta, R. C., Mashburn, A. J., Downer, J., Chomat-Mooney, L., & Hamre, B. (2008, March). *Partitioning variance associated with classroom observations*. Paper presented at the annual meeting of the American Educational Research Association, New York.

Guarino, C., Dieterle, S. G., Bargagliotti, A. D., & Mason, W. M. (2013). What can we learn about effective early mathematics teaching? A framework for estimating causal effects using longitudinal survey data. *Journal of Research on Educational Effectiveness, 6*, 164-198.

Hatchey, A. C. (2013). The early childhood mathematics education revolution. *Early Education and Development, 24*, 419-430.

Hill, H. (2011). *Mathematical Quality of Instruction*. Paper presented at the National Center for Teacher Effectiveness Conference. Retrieved from: https://cepr.harvard.edu/ncte-conference-2011

Hill, H. (2014). *Mathematical Quality of Instruction* (MQI: 4-point version). Retrieved from: http://isites.harvard.edu/icb/icb.do?keyword=mqi_training

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*. Seattle WA: Bill & Melinda Gates Foundation. Retrieved from: http://k12education.gatesfoundation.org/resource/the-reliability-of-classroom-observations-by-school-personnel/

Horizon Research (2003). *Looking inside the classroom: A study of K-12 mathematics and science education in the United States*. Retrieved from: http://www.horizon-research.com/insidetheclassroom/reports/looking/complete.pdf

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). *Have we identified effective teachers? Validating measures of effective teaching using random assignment*. Seattle, WA: Bill & Melinda Gates Foundation.

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching*. Seattle WA: Bill & Melinda Gates Foundation.

Kilday, C. R., & Kinzie, M. B. (2009). An analysis of instruments that measure the quality of mathematics teaching in early childhood. *Early Childhood Education Journal, 36,* 365-372.

Lakes, D. K, & Hoyt, W. T. (2010). Applications of generalizability theory to clinical child and adolescent psychology research. *Journal of Clinical & Adolescent Psychology, 38,* 144-165.

Learning Mathematics for Teaching Project. (2011). Measuring the mathematical quality of instruction. *Journal of Mathematics Teacher Education, 14*, 25-47.

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.

McGuire, P. R., Kinzie, M. Thunder, K. & Berry, R. (2016) Methods of analysis and overall mathematics teaching quality in at-risk prekindergarten classrooms. *Early*

Education and Development, 27, 89-109.

National Council on Teacher Quality (2015). *State of the states 2015: Evaluating teaching, leading, and learning.* Retrieved from: http://www.nctq.org/dmsStage/StateofStates2015

National Mathematics Advisory Panel (2008). *Final report of the National Mathematics Advisory Panel*. Washington D.C.: U.S. Department of Education.

National Research Council (2001). *Adding it up: Helping children learn mathematics*. Washington DC: The National Academies Press.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*. (3rd Ed). New York: McGraw-Hill.

Patrick, H., & Mantzicopoulos, P. (2016). Is effective teaching stable? *Journal of Experimental Education, 84,* 23-47.

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *Classroom Assessment Scoring System manual K-3*. Baltimore, MA: Brookes Publishing.

Piburn, M., Sawada, D., Turley, J., Falconer, K., Benford, R., Bloom, I., & Judson, E. (2000). *Reformed Teaching Observation Protocol (RTOP: Reference manual).* ACEPT Technical Report No. IN003. Tempe, AZ: Arizona Collaborative for Excellence in the Preparation of Teachers. Retrieved from http://files.eric.ed.gov/fulltext/ED447205.pdf

Praetorius, A., Pauli, C., Reusser, K., Rakoczy, K., & Klieme, E. (2014). One lesson is all you need? Stability of instructional quality across lessons. *Learning and Instruction, 31*, 2-12.

Shavelson, R. J. & Webb, N. M. (1991). *Generalizability Theory: A primer.* London:

Sage.

Swiss Society for Research in Education Working Group. (2006). *EDUG user guide*. Neuchatel, Switzerland: IRDP.

Walkington, C., Arora, P., Ihorn, S., Gordon, J., Walker, M., Abraham, L., & Marder, M. (2012). *Development of the UTeach Observation Protocol: A classroom observation instrument to evaluate mathematics and science teachers from the UTeach preparation program*. Retrieved from http://cwalkington.com/UTOP_Paper_2011.pdf

Watts, T. W., Duncan, G. J., Clements, D. H., & Sarama, J. (2018). What is the long-run impact of mathematics during preschool? *Child Development, 89*, 539-555.

Whitehurst, G. J., Chingos, M. M., & Lindquist, K. (2014). *Evaluating teachers with classroom observations. Lessons learned in four districts*. Retrieved from: http://www.brookings.edu/~/media/research/files/reports/2014/05/13-teacher-evaluation/evaluating-teachers-with-classroom-observations.pdf

U.S. Department of Education. (2011). *Fact sheet: Bringing flexibility and focus to education law*. Retrieved from http://www.whitehouse.gov/sites/default/files/fact_sheet_bringing_flexibility_and_focus_to_education_law_0.pdf

Table 1

Descriptive Statistics and Correlations for the MQI Scales

MQI Domain	1.	2.	3.	4.	5.	6.
1. Richness of Mathematics [a]						
2. Working with Students [a]	0.65**					
3. Common Core Aligned Practices [a]	0.52*	0.84**				
4. Errors & Imprecision [a]	0.21	0.39	0.18			
5. Ambitious Math Instruction [b]	0.89**	0.93**	0.86**	0.30		
6. Whole Lesson Scale [c]	0.74**	0.79**	0.75**	0.10	0.86**	
M	0.55	0.72	0.49	0.02	0.59	3.09
SD	0.20	0.17	0.13	0.02	0.15	0.32
Minimum score	0.30	0.51	0.32	0.00	0.42	2.41
Maximum score	0.97	1.03	0.84	0.08	0.87	3.66

Note. [a] scored 0-3, [b] average score on items that make up Domains 1-3, [c] scored 1-5.

*$p < .05$; ** $p < .01$

Table 2

ICC's and Internal Consistency Reliabilities (Coefficient alpha) for the MQI Scales across Three Raters and Five Lessons

		Mean (*SD*)			*ICCs*	*alpha*
		Rater 1	Rater 2	Rater 3		
MQI Domains						
Richness of Mathematics	6-item Scale	0.54 (0.22)	0.55 (0.19)	0.55 (0.20)	0.95	0.83
	Holistic Item	0.96 (0.36)	1.06 (0.27)	0.68 (0.35)	0.80	--
Working with Students & Mathematics	2-item Scale	0.74 (0.18)	0.67 (0.20)	0.75 (0.19)	0.87	0.55
	Holistic Item	1.09 (0.23)	0.93 (0.26)	1.09 (0.25)	0.79	--
Errors & Imprecision	3-item Scale	0.03 (0.05)	0.01 (0.02)	0.01 (0.01)	0.43	0.55
	Holistic Item	0.05 (0.10)	0.03 (0.05)	0.02 (0.04)	0.23	--
Common Core Aligned Practices	5-item Scale	0.53 (0.21)	0.41 (0.12)	0.54 (0.11)	0.81	0.73
	Holistic Item	0.65 (0.32)	0.84 (0.26)	0.96 (0.23)	0.80	--
MQI Whole Lesson Scale	9-item Scale	3.10 (0.36)	3.04 (0.36)	3.13 (0.28)	0.94	0.91
	Holistic Item	2.86 (0.58)	2.70 (0.53)	2.83 (0.42)	0.88	

Table 3

Variance Components for a Two-Facet Partially Nested G-study of Teacher Practices using the MQI Scales

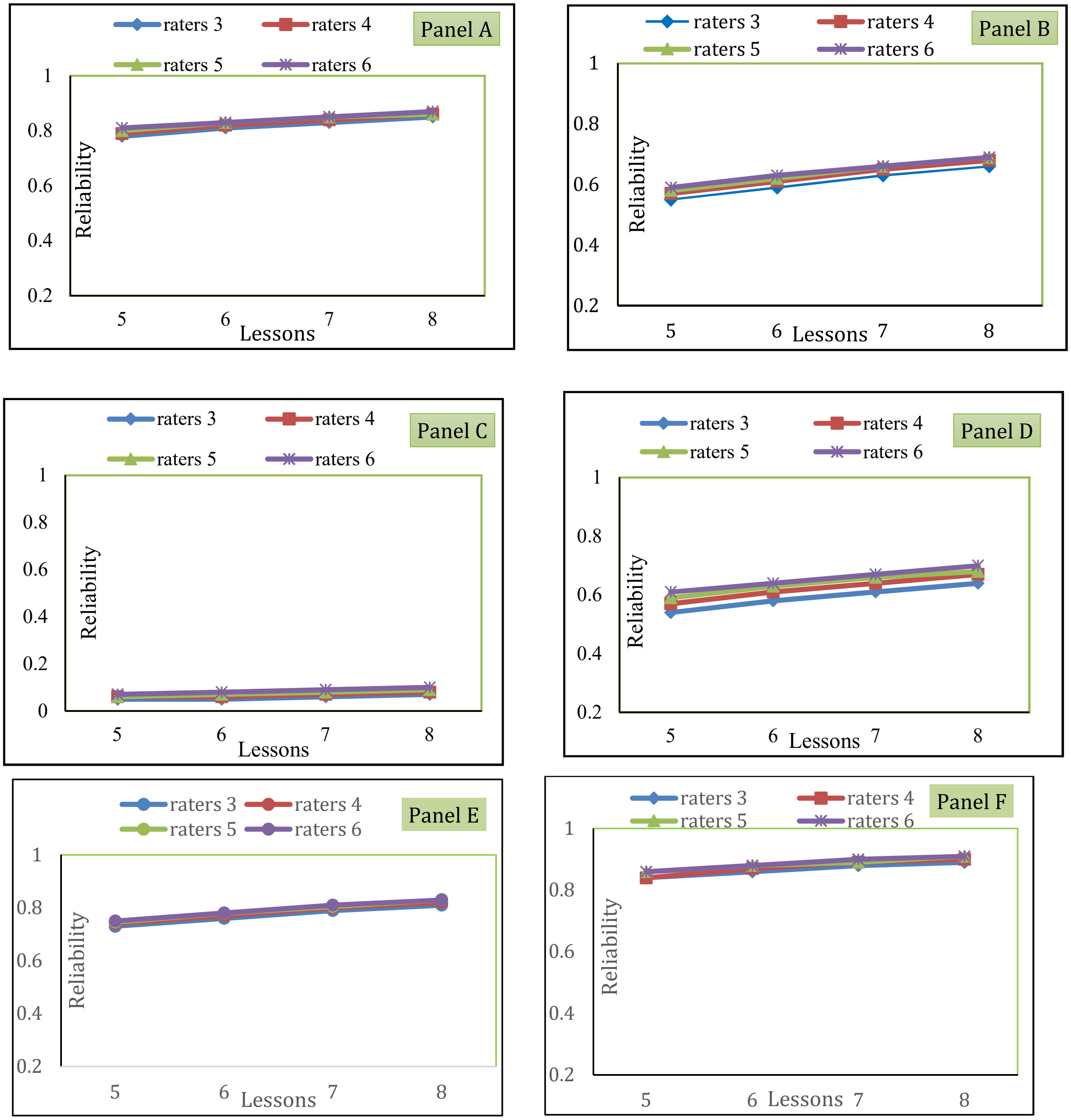
		Average of Items in Each MQI Domain		Holistic Item	
Domain	Source of Variation	Estimated Variance Component (SE)	% of Total Variance	Estimated Variance Component (SE)	% of Total Variance
Richness of Mathematics	Teachers (t)	0.030 (0.012)	33.9	0.065 (0.028)	21.7
	Raters (r)	-0.00 (0.000)*	0.0	0.036 (.027)	12.6
	Lessons: Teachers (l:t)	0.032 (0.010)	35.9	0.881 (0.019)	29.4
	Teachers x Raters (tr)	0.001 (0.001)	0.7	0.004 (0.006)	1.4
	Residual (l:tr,e)	0.026 (0.003)	29.6	0.104 (0.011)	34.9
	Total		100.0		100.0
	Relative Error Variance		0.01		0.03
Working with Students and Mathematics	Teachers (t)	0.016 (0.009)	13.5	0.023 (0.014)	12.5
	Raters (r)	0.001 (0.001)	1.0	0.007 (0.006)	4.2
	Lessons: Teachers (l:t)	0.047 (0.010)	39.8	0.073 (0.015)	38.4
	Teachers x Raters (tr)	-0.00 (0.002)*	0.0	0.007 (0.005)	3.9
	Residual (l:tr,e)	0.054 (0.006)	45.7	0.078 (0.008)	41.0
	Total		100.0		100.0
	Relative Error Variance		0.01		0.02
Errors and Imprecision	Teachers (t)	0.000 (0.000)	0.6	-0.001 (0.000)*	0.0
	Raters (r)	0.000 (0.000)	1.3	0.000 (0.000)	0.5
	Lessons: Teachers (l:t)	0.001 (0.000)	21.9	0.003 (0.001)	16.7
	Teachers x Raters (tr)	0.000 (0.000)	6.0	0.001 (0.001)	8.4
	Residual (l:tr,e)	0.002 (0.002)	70.1	0.013 (0.001)	74.5
	Total		100.0		100.0
	Relative Error Variance		0.01		0.002

Domain	Source of Variation	Average of Items in Each MQI Domain		Holistic Item	
		Estimated Variance Component (SE)	% of Total Variance	Estimated Variance Component (SE)	% of Total Variance
Common Core Aligned Practices	Teachers (t)	0.009 (0.005)	12.5	0.033 (0.016)	13.3
	Raters (r)	0.004 (0.003)	5.7	0.023 (0.017)	9.3
	Lessons: Teachers (l:t)	0.022 (0.005)	31.0	0.040 (0.015)	16.5
	Teachers x Raters (tr)	0.002 (0.002)	3.6	0.001 (0.007)	0.6
	Residual (l:tr,e)	0.034 (0.003)	47.2	0.149 (0.016)	60.3
	Total		100.0		100.0
	Relative Error Variance		0.01		0.02
Whole Lesson	Teachers (t)	0.087 (0.032)	38.0	0.158 (0.067)	26.1
	Raters (r)	0.001 (0.001)	0.6	0.003 (0.005)	0.6
	Lessons: Teachers (l:t)	0.053 (0.013)	23.2	0.175 (0.041)	28.9
	Teachers x Raters (tr)	-0.001 (0.004)*	0.2	0.023 (0.017)	3.9
	Residual (l:tr,e)	0.087 (0.009)	38.0	0.245 (0.027)	40.5
	Total		100.0		100.0
	Relative Error Variance		0.02		0.06
Ambitious Mathematics Instruction [a]	Teachers (t)	0.016 (0.006)	26.6	0.036 (0.016)	24.2
	Raters (r)	0.001 (0.001)	1.3	-0.001 (0.001)	0.0
	Lessons: Teachers (l:t)	0.022 (0.004)	37.9	0.055 (0.011)	37.1
	Teachers x Raters (tr)	0.000 (0.001)	0.0	0.001 (0.003)	1.1
	Residual (l:tr,e)	0.020 (0.002)	34.2	0.056 0.006)	37.7
	Total		100.0		100.0
	Relative Error Variance		0.02		0.02

Note. [a] Aggregate of Richness of Mathematics, Working with Students and Mathematics, and Common Core Aligned Practices. *Negative estimates were set to zero. Many ways exist to deal with these values but in practice the same result is likely (Brennan, 2001, p. 85). SE = standard error of the variance component estimate

Figure 1

Reliability Estimates (G) [on the Y Axis] for Combinations of Number of Raters by Number of Lessons Rated [on the X axis] for: Richness of Mathematics (Panel A), Working with Students and Mathematics (Panel B), Errors and Imprecision (Panel C), Common Core Aligned Practices (Panel D), Ambitious Mathematics Instruction (Panel E), and Whole Lesson (Panel F).


Panel A
raters 3
raters 4
raters 5
raters 6
Reliability
Lessons
Panel B
raters 3
raters 4
raters 5
raters 6
Reliability
Lessons
Panel C
raters 3
raters 4
raters 5
raters 6
Reliability
Lessons
Panel D
raters 3
raters 4
raters 5
raters 6
Reliability
Lessons
Panel E
raters 3
raters 4
raters 5
raters 6
Reliability
Lessons
Panel F
raters 3
raters 4
raters 5
raters 6
Reliability
Lessons