When Rater Reliability Is Not Enough : Teacher Observation Systems and a Case for the Generalizability Study

Heather C. Hill, Charalambos Y. Charalambous and Matthew A. Kraft

EDUCATIONAL RESEARCHER 2012 41: 56

DOI: 10.3102/0013189X12437203

When Rater Reliability Is Not Enough: Teacher Observation Systems and a Case for the Generalizability Study

Heather C. Hill[1], Charalambos Y. Charalambous[2], and Matthew A. Kraft[1]

In recent years, interest has grown in using classroom observation as a means to several ends, including teacher development, teacher evaluation, and impact evaluation of classroom-based interventions. Although education practitioners and researchers have developed numerous observational instruments for these purposes, many developers fail to specify important criteria regarding instrument use. In this article, the authors argue that for classroom observation to succeed in its aims, improved observational systems must be developed. These systems should include not only observational instruments but also scoring designs capable of producing reliable and cost-efficient scores and processes for rater recruitment, training, and certification. To illustrate how such a system might be developed and improved, the authors provide an empirical example that applies generalizability theory to data from a mathematics observational instrument.

Keywords: classroom research; measurements; observational research; policy; policy analysis; teacher assessment

Accumulated research evidence over the past two decades has shown that teachers matter for student learning (Nye, Konstantopoulos, & Hedges, 2004; Rockoff, 2004; Rowan, Correnti, & Miller, 2002; Teddlie & Reynolds, 2000). In fact, teacher effects typically explain a higher percentage of variance in student achievement than do school- and system-level factors (Scheerens & Bosker, 1997). For instance, Gordon, Kane, and Staiger (2006) showed that, net of student demographic characteristics and baseline scores, the average difference for students assigned a top-quartile versus bottom-quartile teacher is 10 percentile points. Examining teacher effects over a period of four consecutive years, Kyriakides and Creemers (2008) suggested that, cumulatively, teacher effects can explain up to 34% of the variance in student achievement. Given these results, it is not surprising that interest in measuring teacher quality has grown. In particular, reformers have proposed classroom observation as a means to several ends, including teacher development, teacher evaluation, and impact evaluation of classroom-based interventions.

However, current rhetoric tends to characterize measures of teacher quality, including classroom observations, as if only "true" teacher quality affects teachers' ratings. This is despite the fact that researchers have widely documented the multiple sources of variance in observational scores due to the sampling of lessons, differences among raters, and even the characteristics of the observational instrument itself. In an era that will undoubtedly see major expansion in the number and use of observational instruments, practitioners and researchers alike need to more carefully examine the sources of variation in observational scores and to consider their implications for how these ratings are used.

To this end, we highlight important issues in measuring teaching—and by extension, teacher—quality. We argue that major instrument developers—including states, researchers, and other nongovernmental entities—must go beyond simply writing instruments; they must create observational *systems* in which quality observational instruments, well-trained raters, and robust scoring designs are combined to produce reliable teacher scores. Key decisions during instrument development include determining which items to remove or modify and identifying an optimal number of items to measure intended constructs. Key decisions pertaining to raters include consideration of their initial qualifications, training, and certification. Key decisions in developing a scoring design include determining the intended use of scores and the most cost-effective combination of lessons, and raters per lesson, to arrive at the desired score reliability.

We argue that generalizability theory (Brennan, 2001; Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Marcoulides, 1989; Shavelson & Webb, 1991) can assist in the design of cost-efficient systems that produce reliable scores. Generalizability theory provides a comprehensive framework for making judgments about the multiple elements of observational systems, something that often-reported interrater agreement measures cannot do. Additionally, generalizability theory can provide empirical evidence regarding the optimal number of raters and lessons required to produce desired reliabilities, rather than

[1]Harvard Graduate School of Education, Cambridge, MA
[2]University of Cyprus, Nicosia, Cyprus

Educational Researcher, Vol. 41 No. 2, pp. 56–64
DOI: 10.3102/0013189X12437203

grounding such choices in "common practice." To support this claim, we provide an illustration by describing a generalizability study conducted for the Mathematical Quality of Instruction (MQI), an instrument for measuring mathematics instruction.

Existing and Planned Practice in Instructional Observation

The United States boasts a long history of efforts to measure teacher and teaching quality, including, in various eras, examinations of subject matter knowledge, teaching portfolios, and value-added scores. One constant, however, has been observation-based evaluations of teacher effectiveness, typically conducted by a principal or another administrator. In recent years, researchers and policy makers have pushed for reforms to these observation-based evaluations. The publication of the New Teacher Project's report describing weak existing district teacher evaluation practices (Weisberg, Sexton, Mulhern, & Keeling, 2009) and competition among states to win Race to the Top (RTTT) funds comprised two sources of pressure toward reform. Another source relates to the Gates Foundation–funded Measures of Effective Teaching project, which has advocated for the use of multiple and rigorous measures in teacher evaluation (*Measures of Effective Teaching*, n.d.). These different sources of pressure have resulted in a large number of states—including New York, Florida, Maryland, Georgia, and Tennessee—making significant revisions to their practices around teacher observation.

Given this new policy environment, we emphasize in this article the importance of observational *systems*, which we define as a collection of elements that together produce scores representing individual teachers' instructional quality. These elements include the observational instrument itself, the set of raters recruited or available to conduct the observations, rater training and certification, and the scoring design used. A scoring design consists of specifications regarding the number and length of observations to be collected per teacher, the number of raters per observation, and certification or other rater requirements.

Although these latter elements might be seen as merely logistical details to be negotiated with stakeholders once an observational instrument has been adopted, decisions regarding raters and scoring designs have important consequences for the reliability of teachers' scores. This is because in addition to actual teaching quality, teacher scores are influenced by other aspects of the instructional environment, including the curriculum and content covered during observed lessons, the students assigned to the class, the degree to which raters (observers) agree about what they see, and random variation (see, e.g., Kennedy, 2010). Although several of these sources of variation can be addressed in the design of an observational system, a review of extant and planned observational systems suggests that many are being overlooked.

Consider, for instance, raters. An informal online poll of state officials engaged in teacher evaluation reform (National Center for Teacher Effectiveness, 2011) suggests that many states intend to rely exclusively on principals to conduct observations. This reliance on principals as raters makes it nearly impossible to exclude individuals who are using the observational instrument in unintended or inconsistent ways; it also preserves existing practices in which only one rater evaluates each teacher.[1] In addition, we suspect that the use of principals as raters may affirm the institutionalization of a process that many teachers describe as ad hoc and unsystematic (Johnson, 1990; Peterson, 1987, 2000) and also may cause rater quality and score reliability to continue to remain unexamined or unreported, as is currently the norm.

Further, we observe that when rater quality is examined, many practitioners (and researchers) report only rater agreement levels with one another—that is, interrater reliability (see, e.g., Heneman & Milanowski, 2003, about Cincinnati; Sartain, Stoelinga, & Brown, 2009, about Chicago). Yet even strong rater agreement, typically expressed in statements such as "the raters' scores match 80% of the time or more," does not assure the consistency of teacher scores and may even mask problems with the data. This occurs because rater agreement levels are influenced by the number of points on a rating scale, the frequency of target behaviors in classroom teaching, and chance agreement. More important, however, rater agreement rates attend to only one source of variation—the rater—leaving unstudied other sources of variation (e.g., lessons) that affect the consistency of teacher scores (see Brennan, 2011, and Marcoulides, 1989, for longer discussions). This metric also fails to estimate interactions between raters, teachers, and lessons—whether, for instance, some raters may be harsher with certain groups of teachers or whether the prevalence of raters' scores varying by specific lesson content suggests the use of multiple raters per lesson. As a result, despite their common use, rater agreement rates do not provide a comprehensive picture of the reliability of scores generated from observational systems.

Similar issues exist around decisions about the number of observations appropriate for producing teacher scores. Current practices appear limited to one or two observations per teacher per year (Weisburg et al., 2009), and evidence suggests that states have made variable decisions about the number of observations required in their newly designed systems. For example, Tennessee intends to require four observations per year for tenured teachers (National Center for Teacher Effectiveness, 2011), whereas new legislation in Louisiana requires only one per year (Louisiana Act 54, 2010); in neither case is there evidence that states generated these numbers via scientific study. If either the content of the lesson observed or day-specific random variation (e.g., students are distracted by an upcoming sporting event) exert a strong influence on teacher scores, one, two, or even four lessons may not be enough to arrive at the level of reliability needed to inform high-stakes decisions. We also know from anecdotal evidence that principals may be pressed for time and thus "sample," in a sense, a half hour from a lesson before moving onto their next responsibility. If the reliability of teacher scores is not affected by this sampling, then, at least from a measurement perspective, it is a smart strategy. However, if principals are systematically missing important aspects of instruction because of this time-saving approach, their ratings might not consistently capture the overall quality of instruction occurring in the classroom.

Two other issues around the design of observational systems are important to highlight. The first pertains to questions regarding the design and/or adoption of an observational instrument. The number of items[2] on an instrument—and by extension, the cognitive load for raters—is one example of such a design question. For instance, the Framework for Teaching, a commonly used observational instrument, has 76 indicators grouped

beneath 22 actual items for observers to track (Danielson Group, 2011). Other instruments, such the Tennessee observation framework (Tennessee Department of Education, 2011; 14 indicators grouped beneath 6 items) and the Boston teacher evaluation instrument (Boston Public Schools, 2010; 44 indicators grouped beneath 6 items), contain fewer. Despite such variability, we found no studies of how the quantity of items on an instrument might affect raters' performance and consequently the characteristics of the resulting teacher scores.

Second, although RTTT regulations suggest that overall evaluation scores should inform performance rewards, promotion, retention, and the tenure process, policy makers are not provided with details about how these scores should be constructed. These details, however, have important consequences for scoring designs and estimated score reliability. If absolute decisions are made regarding teacher efficacy—for example, comparing untenured teachers against a criterion of effectiveness—a scoring design with more raters per lesson and/or observations per teacher will be required. However, if the intent is to compare teachers with one another—for instance, to eliminate the bottom 10% of performers—less intensive data collection and scoring will typically be required (see Shavelson & Webb, 1991, pp. 84–87, for an explanation on which variance components are incorporated into the calculation of the absolute-decision and the relative-decision reliability indices).

In sum, based on existing and planned practice, we suspect that instrument developers have paid little attention to how decisions about the design of their observational systems will affect the reliability and validity of the resulting teacher scores. In this article, we focus in particular on the reliability of such scores, leaving a discussion of validity (e.g., rater accuracy, the impact of instrument or scoring designs on correlations with "true" teaching quality or student learning) for another venue. For reliability, there exists a well-established framework, generalizability theory, which allows us to examine the multiple influences on score reliability within a single analysis (Brennan, 2001; Cronbach et al., 1972; Shavelson & Webb, 1991). Generalizability studies, or G-studies, decompose variability in teacher scores into different components (e.g., teachers, lessons, and raters), their interactions, and measurement error. This partitioning of variance can inform decisions regarding the improvement of the instrument or raters' training on particular aspects of the instrument. Using information from G-studies, instrument developers can then conduct decision-type studies (D-studies) to identify the optimal data collection and scoring designs for a desired score reliability.

An Example: Developing the MQI Observational System

Study Design

To illustrate how this might occur, we describe a G-study and a series of D-studies conducted for the MQI instrument. The MQI is designed to provide information about teachers' enactment of mathematics instruction. Three of its major dimensions are the richness of the mathematics (Richness), teacher errors and imprecision (Errors and Imprecision), and student participation in mathematical meaning making and reasoning (SPMMR; see the appendix). The MQI was designed to provide both a multidimensional and a balanced view of mathematics instruction and is currently intended for use with videotaped lessons of classroom mathematics instruction (Hill et al., 2008; Learning Mathematics for Teaching Project, 2011). Although the MQI is designed as a measure of *teaching* quality rather than *teacher* quality, we view teaching quality as a critical element in teacher quality; good teachers typically teach well.

Because the MQI was intended to yield estimates of instructional quality for individual teachers, we faced several questions regarding instrument and scoring design. Most important, we needed to determine whether the instrument could meet target score reliabilities under specific designs; we also needed to estimate the most cost-effective means for collecting and scoring lessons. These challenges are similar to those any state or other major instrument developer might face as it moves from an initial phase of testing to wide-scale use in teacher evaluation. We also wanted to estimate the characteristics of the instrument under "real" district conditions, including using one rater per teacher, capturing only one or two lessons per year, and allowing for the possibility that a given rater will watch only part of a lesson. Finally, we sought to examine how using scores to make absolute versus relative decisions affects estimated score reliability and, consequently, recommended scoring design requirements. Although we demonstrate these issues by using a small study designed to improve a mathematics instrument, we expect that our analyses might inspire states and other instrument developers to undertake more rigorous studies of their higher stakes instruments.

A first step in our process was to sample teachers, lessons, and raters. From a pool of 24 middle school teachers participating in a related study, we sampled 8 who represented different levels of mathematical knowledge of teaching (see Hill, Umland, & Kapitula, 2011). Because in previous studies (e.g., Hill et al., 2008) we found teacher knowledge to be positively related to the quality of teacher instruction, we expected notable variations in teachers' MQI scores—an optimal feature of G-study analysis. From the six videotaped lessons available per teacher, we sampled three that contained between six and eight 7.5-minute segments. The segment length was set based on prior raters' feedback that segments greater than 7.5 minutes were difficult to score. Sampling lessons with six to eight segments increased comparability.

To form a rater pool, we recruited 10 graduate students and former teachers who each attended a two-day training session on the instrument. At the conclusion of the training, raters took a certification examination, which asked them to rate 16 segments from four different lessons taught by four different teachers; the segments were purposefully selected to represent a wide range of instructional quality. To determine rater inclusion, we set a cut score of 0.20 average absolute deviations from the master score; this corresponded to rater scores that were off by 1 point from the master score 40% of the time or off by 2 points 20% of the time. This criterion excluded 1 rater, and thus the findings reported below are based on the scores of the remaining 9 raters. We note that had we been a state or district that required principals to serve as raters by design, we could neither set a certification cut score nor exclude any rater.

Each of the nine individuals assigned scores of low, medium, or high (1, 2, or 3) for every item for each segment within our

Table 1

Variance Decomposition for the Richness Dimension of the Mathematical Quality of Instruction Instrument

	Individual Items					Overall Richness	
	(I)	(II)	(III)	(IV)	(V)	(VI)	(VII)
Source of Variation	Representations	Multiple Solution Procedures/ Solutions	Explanations	Developing Generalizations	Mathematical Language	Average of Items (I)–(V)	Holistic
Teachers (t)	0.97	34.61	22.01	0.00	41.55	42.52	45.70
Lessons:teachers (l:t)	20.24	28.00	12.01	23.94	16.99	10.52	2.76
Raters (r)	9.14	2.61	21.48	7.61	4.99	6.17	13.96
Teachers × Raters (t × r)	0.00	0.00	8.52	13.12	3.64	7.83	3.27
Residual [(l:t) × r, e]	69.65	34.78	35.99	55.33	32.83	32.97	34.31
Total	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Note. Cells represent the percentage of variance explained by different facets in a generalizability study.

Table 2

Variance Decomposition for the Errors and Imprecision Dimension of the Mathematical Quality of Instruction Instrument

	Individual Items			Overall Errors and Imprecisions	
	(I)	(II)	(III)	(IV)	(V)
	Major Errors	Notation and Language	Lack of Clarity	Average of Items (I)–(III)	Holistic
Teachers (t)	13.11	31.23	21.12	31.88	36.04
Lessons:teachers (l:t)	5.90	7.22	11.71	8.81	3.20
Raters (r)	2.59	10.72	9.53	13.04	13.51
Teachers × Raters (t × r)	14.35	5.62	1.61	6.45	5.26
Residual [(l:t) × r, e]	64.04	45.21	56.03	39.82	41.99
Total	100.00	100.00	100.00	100.00	100.00

Note. Cells represent the percentage of variance explained by different facets in a generalizability study.

sample of 24 lessons (eight teachers with three lessons each). To analyze the data, we first aggregated the segment scores to the lesson level, based on the view that most mathematics classes feature purposeful differences in instructional methods as the teacher interacts with students through different phases of the lesson. Although one segment may intentionally feature high mathematical richness, for instance, another may intentionally feature the reverse as students practice a familiar procedure for skill proficiency.

Using these data, we then conducted a G-study to determine the variance components attributable to teachers, lessons, and raters; their two-way interactions; and the combination of the three-way interaction and the measurement error. We calculated these variance components for each of the individual MQI items and each MQI dimension. In the latter case, we did so in two ways. First, we averaged the items that pertained to a given dimension (Column VI in Table 1 and Column IV in Tables 2 and 3; e.g., for "average richness," items included representations, multiple solution procedures/solutions, explanations, developing generalizations, and mathematical language). We averaged across items because of our views about these dimensions[3] and because an exploratory factor analysis showed the items clustered in three groups corresponding to the MQI dimensions. Second, we partitioned the variance for a "holistic" item for each dimension; these holistic items allowed raters to use item-specific information but also to exercise more judgment in assigning a score (see the appendix and Tables 1–3 for details). In fact, a question for this G-study was whether the holistic score could achieve a similar level of reliability as the average of the more specific items that comprise a dimension.

Given these decisions, we chose to analyze our data using a G-study design where lessons were nested within teachers and crossed with raters.[4] We considered lessons to be nested within teachers (as opposed to being crossed) because, unlike typical designs in which participants are administered exactly the same items on an examination, our participants were not expected to teach exactly the same lessons (see Erlich & Borich, 1979, p. 12, for a similar discussion). Because we averaged across items, items were not considered a facet in our analysis; we acknowledge, however, that alternative specifications of the model (e.g., items as fixed facets) are possible.

Before presenting our findings, we pause to note that the data used in our analysis were not collected with the intent of shaping the instrument's use in high-stakes teacher evaluation but rather were intended to help us refine the MQI for research and teacher development purposes. Thus the analysis described below does not provide a template for inquiries into the properties of teachers' scores for use in formal evaluations; to do so, states or

Table 3
Variance Decomposition for the Student Participation in Meaning Making and Reasoning (SPMMR) Dimension of the Mathematical Quality of Instruction Instrument

	Individual Items			Overall SPMMR	
	(I)	(II)	(III)	(IV)	(V)
	Student Explanations	Student Questioning and Reasoning	Enacted Task Demand	Average of Items (I)–(III)	Holistic
Teachers (t)	17.77	14.16	21.96	32.78	27.11
Lessons:teachers (l:t)	39.81	11.74	6.09	7.22	2.09
Raters (r)	10.71	33.10	24.19	28.58	27.12
Teachers × Raters (t × r)	2.26	0.05	1.23	0.00	2.48
Residual [(l:t) × r, e]	29.45	40.94	46.52	31.43	41.19
Total	100.00	100.00	100.00	100.00	100.00

Note. Cells represent the percentage of variance explained by different facets in a generalizability study.

other instrument developers would undoubtedly want to work with a larger sample of teachers and conduct random sampling of teachers, lessons, and raters from the grades and districts affected by the observation instrument. Nevertheless, the findings presented below can be thought of as an illustration of the affordances of G-Theory in uncovering issues pertaining to the design and improvement of observational systems.

Generalizability Study Results

We began our analysis by decomposing the variance in teacher scores on specific items of the MQI (see Tables 1–3); this information can help inform the refinement of an observational instrument. Specifically for Richness, two items—representations and developing generalizations—exhibited negligible teacher-level variation; instead, a notable portion of the variance appeared to lie between lessons within teachers (i.e., teachers would feature these elements in one lesson but not in another) or appeared as measurement error. Based on raters' reports of its difficulty to score, we chose to drop the item corresponding to representations from the instrument and to rewrite the developing generalizations item. Table 1 also shows that the estimated variance attributable to raters for the mathematical explanations item was large relative to other items, implying differences in raters' interpretation and use of mathematical explanations. This issue also surfaced in follow-up interviews, where raters did appear to have differing interpretations of the item. Training for this item was changed to address these inconsistencies.

Tables 1–3 also display the two more global estimates described above: the average of dimension-specific items (e.g., Column VI, an average of the five items of Richness in Table 1) and the holistic judgment made about the segment for each dimension. As shown, the variance components associated with teachers were higher for the dimension averages compared with those of the specific items within the dimension. Further, Tables 1–3 demonstrate that the average and holistic scores yielded roughly comparable teacher-level variances, especially for Richness. Although scores on specific items necessarily informed raters' overall judgments, this finding appears to be promising, for if raters can arrive at an overall judgment using a shorter instrument, this would result in a substantial reduction in scoring time and thus costs. We note, however, that a more rigorous test of this would be to compare dimension-average scores with holistic scores obtained *without* asking the raters to first score each individual item of a dimension, something we are currently exploring in a new G-study.

As suggested above, results from our G-study provided information that the interrater agreement estimates alone did not. For instance, representations and mathematical language had 69% and 55% agreement rates, respectively—not a tremendous difference and well below the conventional 80% threshold. Yet these two items performed remarkably differently in the G-study. Although a relatively small portion of the variance (i.e., less than 10%) was attributed to raters in both cases, the portion of variance attributed to teachers was notably different: less than 1% for representations, and a little more than 40% for mathematical language. As noted, representations was dropped, whereas mathematical language was the strongest specific item within the dimension. Similarly, major errors and developing generalizations had the first- and second-highest level of interrater agreement (85% and 83%, respectively), yet the G-study indicated that very little of the variance in scores was attributable to teachers for these items. These high agreement rates likely occurred because of the scarcity of such behaviors during instruction—in other words, most of the matches were probably due to raters indicating that the element did not occur. These examples highlight the fact that developers who use only interrater reliability to investigate the integrity of their instrument miss critical information and, in fact, may mask limitations of the instrument. To see the bigger picture, a G-study must be conducted.

Design Study Results

Next, we used the information generated by the G-study to help delineate two important aspects of the scoring design: the number of raters needed per lesson and the number of lessons per teacher required to achieve acceptable reliability. To do so, we entered the estimates for the dimension-average scores (Column VI in Table 1 and Column IV in Tables 2 and 3) into a series of D-studies. In what follows, we start by reporting estimates of relative reliability (ρ values, which correspond to the generalizability coefficient, or G-coefficient) because districts often make

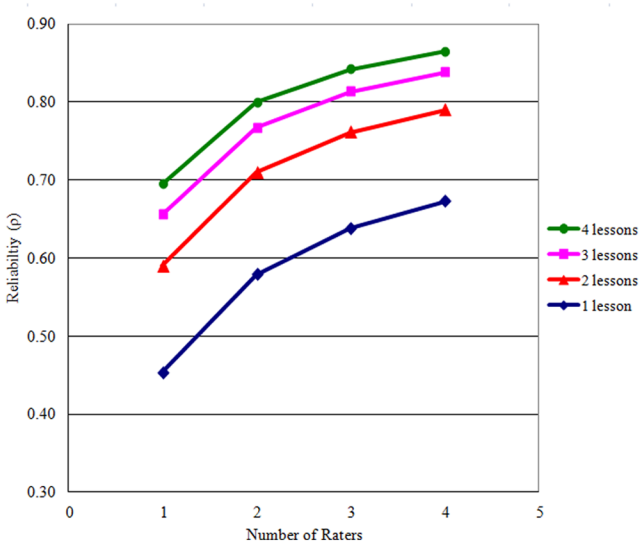

FIGURE 1. *Richness: the reliability of different combinations of raters and lessons.*

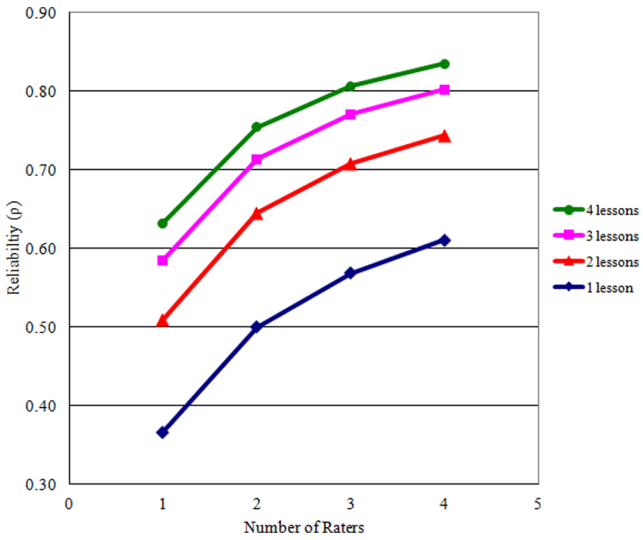

FIGURE 2. *Errors and Imprecision: the reliability of different combinations of raters and lessons.*

or plan to make *relative* rather than absolute decisions about teachers, such as laying off the lowest 5% of teachers due to budget shortfalls or rewarding the top 5% with merit pay. We then move to a discussion of how estimating absolute reliability (φ, the index of dependability coefficient) would affect our design. Figures 1–3 display the results of this analysis for each of the three dimensions. We assume for the sake of illustration that the variance components from our videotaped observations are similar to those that would be obtained through live observations, although this remains an important empirical question for further research.

Assuming a scoring design of quarterly observations of teachers by an MQI trained and certified principal (one rater, four lessons), we estimated the reliability of teacher scores to be

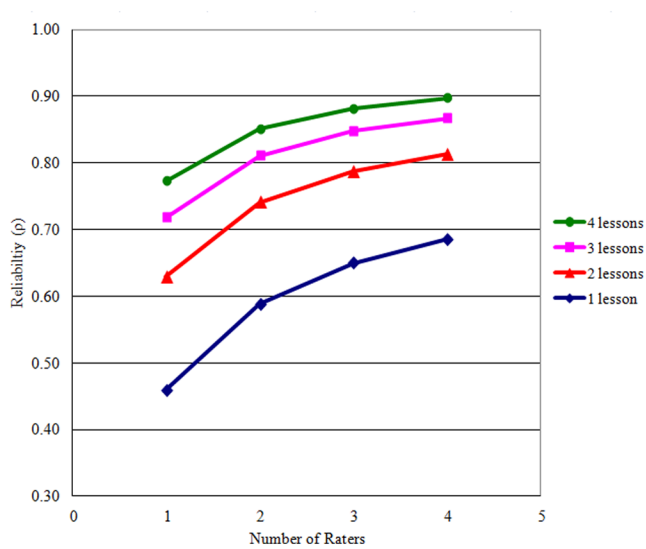

FIGURE 3. *Student Participation in Mathematical Meaning Making and Reasoning: the reliability of different combinations of raters and lessons.*

0.69, 0.63, and 0.77 for Richness, Errors and Imprecision, and SPMMR, respectively. For a more typical scoring design of two observations per year by a principal, our data indicate that the MQI would return estimated reliabilities of 0.59, 0.51, and 0.63. This finding suggests that using the MQI instrument with typical scoring designs would produce scores that are not sufficiently reliable to support the decisions desired by current policy proposals. Based on the similarity of our results to others in the field (e.g., Newton, 2010), we suspect the same will be true of other observational systems that employ scoring designs with a single rater and few observations.

Using Figures 1–3, we next sought to identify an optimal scoring design by examining how much the estimated reliability of teacher scores improves relative to the "costs" of adding lessons and/or raters. We assumed a fixed unit cost per lesson collected and scored. Figures 1–3 demonstrate the diminishing marginal returns to reliability for both lessons (smaller gaps between curves) and raters (decreases in slope) across each dimension. In particular, adding a fourth lesson appears to increase the estimated reliability of teacher scores by only a marginal amount. The figures also show that adding a second rater to each lesson markedly improves estimated teacher score reliability. Thus, we identify the three lesson–two rater combination as an optimal combination for research purposes, with estimated teacher score reliabilities for Richness, Errors and Imprecision, and SPMMR of 0.77, 0.71, and 0.81, respectively. We note that under other logistics and budget assumptions, a different combination may be optimal and that D-studies are designed to provide precisely this information.

We then examined the effect on score reliability of watching only the first 30 minutes (four segments) of a lesson rather than the entire period, something that principals might do when pressed for time or researchers might do if pressed for resources. As shown in Table 4, the estimated reliabilities for Richness and for Errors and Imprecision remain largely unchanged. By

Table 4

Comparison of the Reliability Estimates (ρ) for Different Combinations of Raters and Lessons for the Whole Lesson and the First 30 Minutes of a Lesson

Number of Lessons and Raters	Richness		Errors and Imprecision		Student Participation in Meaning Making and Reasoning	
	Whole Lesson	30 Minutes	Whole Lesson	30 Minutes	Whole Lesson	30 Minutes
One lesson						
1 rater	0.45	0.50	0.37	0.34	0.46	0.32
2 raters	0.58	0.59	0.50	0.46	0.59	0.41
3 raters	0.64	0.63	0.57	0.53	0.65	0.45
4 raters	0.67	0.65	0.61	0.57	0.68	0.48
Two lessons						
1 rater	0.59	0.65	0.51	0.49	0.63	0.49
2 raters	0.71	0.73	0.64	0.62	0.74	0.58
3 raters	0.76	0.77	0.71	0.68	0.79	0.62
4 raters	0.79	0.78	0.74	0.71	0.81	0.65
Three lessons						
1 rater	0.66	0.73	0.58	0.57	0.72	0.59
2 raters	0.77	0.80	0.71	0.70	0.81	0.68
3 raters	0.81	0.83	0.77	0.75	0.85	0.71
4 raters	0.84	0.84	0.80	0.78	0.87	0.73
Four lessons						
1 rater	0.69	0.77	0.63	0.63	0.77	0.66
2 raters	0.80	0.83	0.75	0.74	0.85	0.74
3 raters	0.84	0.86	0.81	0.79	0.88	0.77
4 raters	0.86	0.87	0.83	0.82	0.90	0.78

contrast, watching only the first 30 minutes of the lesson yields notably lower estimated reliabilities for SPMMR. Thus, the effect of this form of sampling appears to vary by dimension, at least for our instrument.

As noted above, it is also possible within the D-study framework to determine the impact of different assumptions about the use of teacher scores. Above, we used the G-coefficient (ρ), which corresponds to relative decisions (e.g., rewarding the top 5% of teachers). If states and districts intend to use scores to make absolute decisions—for example, to hold teachers to certain tenure criteria—the dependability coefficient (ϕ) should be used. For the same combination of lessons and raters, the absolute-reliability estimates (ϕ) are lower than the relative-reliability estimates reported above. For example, the three lessons–two raters combination yields estimated absolute reliabilities of 0.73, 0.62, and 0.60 for Richness, Errors and Imprecision, and SPMMR, respectively. To obtain estimates equal to those reported above for relative decisions when the three lesson-two rater combination is used (ρ = 0.77, 0.71, and 0.81), at the minimum an additional rater would be required for the first two dimensions; at least three more raters per lesson and three more lessons per teacher would be needed for the latter dimension. This implies that when absolute decisions need to be made, more lessons and/or raters are required to achieve similar reliabilities to those obtained for relative decisions. Given that findings from other major instruments are likely to be similar, we suspect that the use of these instruments would best be limited to informing relative decisions in many locations.

Finally, G-studies and D-studies can provide additional information about proposed uses of the instrument. For instance, in future analyses we will conduct G-studies that compare the performance of the MQI for live and taped mathematics lessons. One could also examine whether an instrument intended for use across all academic subjects performs equally well on reading, mathematics, and science, holding other elements of the scoring design constant. Instrument developers may also be interested in the differences in score characteristics between rater pools constructed solely of specially recruited, trained, and certified individuals versus a system in which all principals are trained and then allowed to conduct teacher evaluations. Finally, in practice, some states might wish to design a "triage" system in which the majority of teachers are measured at relatively low reliability but a few (e.g., those flagged during routine evaluations, those with low value-added scores, those who are novices) are measured with more observations and much higher target reliabilities. D-studies can help determine the design of such a system.

Conclusion

This article highlights issues related to the development and use of observational evaluation systems, particularly those that will contribute to high-stakes decisions about teachers. The empirical findings from this study are limited to the instrument in question; there is no optimal number of observations or raters that transcends specific instruments and rater populations, and we caution against extrapolating our results to other observational instruments and scoring designs. That said, we think this analysis holds several important lessons that are applicable across instrument types and state settings. First, we argue that, contrary to common practice, it is misleading to talk about the reliability of specific instruments; instead, reliability inheres in the joint

combination of instruments, rater training and certification systems, and specific scoring designs that constitute an observational system. Second, our analysis demonstrates empirically the hazard of using a common metric—80% interrater agreement—as a sole measure of the reliability of a classroom observation system. Some items below this threshold performed well in our G-study analysis, whereas other items that met this threshold performed poorly. Third, although reaching high rater agreement levels for such items is clearly preferable, it may not be feasible for some complex performance arenas within teaching, nor should it be used as the sole criterion for determining score reliability, as our findings clearly show. For all these reasons, we consider generalizability theory to be a significant asset in the development of observational systems.

We also demonstrated that once an instrument has been written, developers and users have many more components to which they must attend, such as identifying a data collection and scoring design and improving the instrument via better rater training and/or item design. Our communications with state education officials suggest that decisions about these observational system elements are more often informed by established norms or personnel limitations than by an understanding of how they impact the reliability of teacher scores (see Hill & Herlihy, 2011). Continued improvement and winnowing of productive and nonproductive items also are critical even after the start of official use of the instrument with teachers.

Finally, we note that developing reliable and litigation-proof observational systems takes time, expertise, and generous financial resources. Given that the United States is moving toward national standards for content and curriculum, and given that there is little reason to believe that the basics of good teaching vary greatly from Mississippi to New York, we argue for focusing national efforts on developing a set of carefully tested classroom observation systems. By doing so, we could have confidence that the scores and instructional feedback derived from observational systems will become trusted inputs in teacher development and teacher evaluation systems.

Appendix

Richness of the Mathematics: This dimension captures the depth of the mathematics offered to students. Rich mathematics focus either on the meaning of facts and procedures or on key mathematical practices. The dimension consists of the following items:

- *Representations:* A representation is typically a visual or verbal display of quantitative information that expresses the mathematical concept at hand using a medium other than numbers and symbols.
- *Explanations:* Giving mathematical meaning to ideas, procedures, steps, or solution methods.
- *Multiple procedures or solution methods:* Considering multiple solution methods or procedures for a single problem.
- *Developing generalizations:* Using specific examples to develop generalizations of mathematical facts or procedures.
- *Mathematical language:* Using dense and precise language fluently and consistently during the lesson.

Errors and Imprecision: This dimension is intended to capture teacher errors or imprecision of language and notation, uncorrected student errors, or the lack of clarity/precision in the teacher's presentation of the content. This dimension consists of the following items:

- *Major mathematical errors or serious mathematical oversights* (e.g., solving problems incorrectly, defining terms incorrectly, forgetting a key condition in a definition; equating two nonidentical mathematical terms).
- *Imprecision in language or notation*: imprecision in use of mathematical symbols (notation), use of technical mathematical language, and use of general language when discussing mathematical ideas.
- *Lack of clarity* in teachers' launching of tasks or presentation of the content.

Student Participation in Meaning Making and Reasoning: This dimension captures evidence of students' involvement in cognitively activating classroom work and the extent to which students participate in and contribute to meaning making and reasoning. Attention here focuses on student participation in activities such as:

- *Providing explanations* (e.g., students explain mathematical methods or ideas).
- *Posing mathematically motivated questions or offering mathematical claims or counterclaims* (e.g., asking why a mathematical procedure works).
- *Engaging in reasoning and cognitively demanding activities* (e.g., drawing connections among different representations, concepts, or solution methods; identifying and explaining patterns).

NOTES

The research reported in this paper was supported by the W. T. Grant and Spencer Foundations (#200900175), the Institute for Educational Sciences (R305C090023), the National Science Foundation (0335411 & 0918383), and the Bill and Melinda Gates Foundation. We would like to thank three anonymous reviewers for constructive comments on earlier versions of this manuscript.

[1]An important issue is that using principals to conduct teacher observations means that there typically is no overlap between Principal A and Principal B in teachers observed. Without being able to compare principals' ratings for the same teachers, it is impossible to isolate the rating tendencies of a principal (e.g., an overall tendency toward harshness) from a true measure of teaching quality. This means that any comparisons of teachers based on principal ratings only hold *within schools*—that is, two teachers with the same scores in different schools cannot be assumed to provide equivalent-quality instruction because their scores cannot directly be compared.

[2]We characterize an *item* as a prompt for a rater to assign a score; *indicators* are typically a list of activities that fall under each item. Although each instrument uses its own language to describe indicators and items, we standardize them here for ease of use.

[3]We argue that in order to take into account the way diverse lesson content and curriculum materials influence scores, averaging across items is necessary. By way of illustration, a lesson that requires using manipulatives to give meaning to a mathematical operation (e.g., using colored chips to help students understand integer subtraction) lends

itself better to engaging in activities such as employing representations and providing explanations; in contrast, a lesson involving a rich mathematical problem that admits different solution approaches allows for more work around considering and discussing multiple solution paths. Although the specific teacher behaviors in these two lessons might be different, both lessons feature elements of rich instruction.

[4]We calculated the variance components shown in Tables 1–3 using ANOVA in SPSS (for a fully crossed design) and then employed Brennan's (2001, Chapter 1) formulas to calculate the respective components for a partially nested design. For the meaning of the different variance components involved in this design, see Shavelson and Webb (1991, pp. 52–54) and substitute "occasions" with "lessons."

REFERENCES

Boston Public Schools. (2010). Evaluation form for teachers. Retrieved from http://www.bostonpublicschools.org

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

Brennan, R. L. (2011). Generalizability theory and classical test theory. *Applied Measurement in Education, 24*, 1–21.

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

Danielson Group. (2011). *Framework for teaching: Components of professional practice*. Retrieved from http://charlottedanielson.com/theframeteach.htm

Erlich, O., & Borich, G. (1979). Occurrence and generalizability of scores on a classroom interaction instrument. *Journal of Educational Measurement, 16*(1), 11–18.

Gordon, R., Kane, T. J., & Staiger, D. O. (2006). *Identifying effective teachers using performance on the job* (Discussion Paper 2006-01). Washington, DC: Brookings Institution.

Heneman, H. G., & Milanowski, A. T. (2003). Continuing assessment of teacher reactions to a standards-based teacher evaluation system. *Journal of Personnel Evaluation in Education,17*, 173–195.

Hill, H. C., Blunk, M., Charalambous, C., Lewis, J., Phelps, G. C., Sleep, L., et al. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and Instruction, 26*, 430–511.

Hill, H. C., & Herlihy, C. (2011). Prioritizing teaching quality in a new system of teacher evaluation. *Education Outlook*. Retrieved from http://www.aei.org/outlook/101089

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

Johnson, S. M. (1990). *Teachers at work: Achieving success in our schools*. New York: Basic Books.

Kennedy, M. M. (2010). Attribution error and the quest for teaching quality. *Educational Researcher, 39*, 591–598.

Kyriakides, L., & Creemers, B. P. M. (2008). A longitudinal study on the stability over time of school and teacher effects on student learning outcomes. *Oxford Review of Education, 34*, 521–545.

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

Marcoulides, G. A. (1989). The application of generalizability analysis to observational studies. *Quality and Quantity, 23*, 115–127.

Measures of Effective Teaching. (n.d.). Retrieved from the Bill and Melinda Gates Foundation website: http://www.gatesfoundation.org/united-states/Pages/measures-of-effective-teaching-fact-sheet.aspx

National Center for Teacher Effectiveness. (2011). *Online poll of states engaged in reform of teacher evaluation systems*. Cambridge, MA: Authors.

Newton, X. (2010). Developing indicators of classroom practice to evaluate the impact of a district mathematics reform initiative: A generalizability analysis. *Studies in Educational Evaluation, 36*, 1–13.

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis, 26*, 237–257.

Peterson, K. D. (1987). Teacher evaluation with multiple and variable lines of evidence. *American Educational Research Journal, 24*, 311–317.

Peterson, K. D. (2000). *Teacher evaluation: A comprehensive guide to new directions and practices*. Thousand Oaks, CA: Corwin Press.

Rockoff, J. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review, 94*(2), 247–252.

Rowan, B., Correnti, R., & Miller, R. J. (2002). What large-scale, survey research tells us about teacher effects on student achievement: Insights from the "Prospects" study of elementary schools. *Teachers College Record, 104*(8), 1525–1567.

Sartain, L., Stoelinga, S. R., & Brown, E. (2009). *Evaluation of the excellent in teaching pilot: A report to the Joyce Foundation*. Chicago: Consortium on Chicago School Research.

Scheerens, J., & Bosker, R. J. (1997). *The foundations of educational effectiveness*. Oxford, UK: Pergamon.

Shavelson. R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Thousand Oaks, CA: Sage.

Teacher Evaluation Act, La. Rev. Stat. §§ 17:3881-3886, 17:3901-3905, 17:3997, 17:3891-3895 (2010).

Teddlie, C., & Reynolds, D. (2000). *The international handbook of school effectiveness research*. London: Falmer Press.

Tennessee Department of Education. (2011). *Framework for evaluation and professional growth*. Retrieved from http://www.tn.gov/education/frameval/

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

AUTHORS

HEATHER C. HILL is a professor at the Harvard Graduate School of Education, 6 Appian Way #445, Cambridge, MA 02476; *heather_hill@harvard.edu*. Her research focuses on teacher knowledge and teaching quality in mathematics.

CHARALAMBOS Y. CHARALAMBOUS is a lecturer of educational research and evaluation in the Department of Education, University of Cyprus, Theophanides Building, 11-13 Dramas Street, Nicosia, 1077, Cyprus; *cycharal@ucy.ac.cy*. His research focuses on instructional quality and its measurement, factors contributing to instructional quality, and its relation to student learning.

MATTHEW A. KRAFT is a doctoral student at the Harvard Graduate School of Education, Longfellow Hall, 13 Appian Way, Cambridge, MA 02138; *matthew_kraft@mail.harvard.edu*. His research focuses on human capital strategies to improve teacher quality and quantitative analyses of instructional innovations in schools.

Manuscript received January 2, 2011
Revisions received June 28, 2011; October 3, 2011; and December 20, 2011
Accepted December 29, 2011