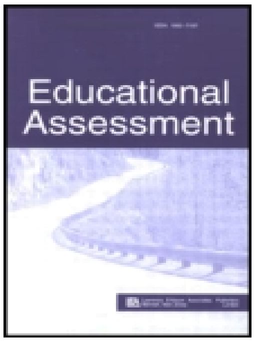

Routledge
Taylor & Francis Group

Educational Assessment

ISSN: 1062-7197 (Print) 1532-6977 (Online) Journal homepage: http://www.tandfonline.com/loi/heda20

Validating Arguments for Observational Instruments: Attending to Multiple Sources of Variation

Heather C. Hill , Charalambos Y. Charalambous , David Blazar , Daniel McGinn , Matthew A. Kraft , Mary Beisiegel , Andrea Humez , Erica Litke & Kathleen Lynch

To cite this article: Heather C. Hill , Charalambos Y. Charalambous , David Blazar , Daniel McGinn , Matthew A. Kraft , Mary Beisiegel , Andrea Humez , Erica Litke & Kathleen Lynch (2012) Validating Arguments for Observational Instruments: Attending to Multiple Sources of Variation, Educational Assessment, 17:2-3, 88-106, DOI: 10.1080/10627197.2012.715019

To link to this article: http://dx.doi.org/10.1080/10627197.2012.715019

Published online: 20 Sep 2012.

Submit your article to this journal

Article views: 319

View related articles

Citing articles: 8 View citing articles

Full Terms & Conditions of access and use can be found at
http://www.tandfonline.com/action/journalInformation?journalCode=heda20

Download by: [University Of Maryland] **Date:** 15 August 2017, At: 12:53

Educational Assessment, 17:88–106, 2012
Copyright © Taylor & Francis Group, LLC
ISSN: 1062-7197 print/1532-6977 online
DOI: 10.1080/10627197.2012.715019

Downloaded by [University Of Maryland] at 12:53 15 August 2017

Validating Arguments for Observational Instruments: Attending to Multiple Sources of Variation

Heather C. Hill
Harvard University

Charalambos Y. Charalambous
University of Cyprus

David Blazar, Daniel McGinn, and Matthew A. Kraft
Harvard University

Mary Beisiegel
Oregon State University

Andrea Humez
Boston College

Erica Litke and Kathleen Lynch
Harvard University

Measurement scholars have recently constructed validity arguments in support of a variety of educational assessments, including classroom observation instruments. In this article, we note that users must examine the robustness of validity arguments to variation in the implementation of these instruments. We illustrate how such an analysis might be used to assess a validity argument constructed for the Mathematical Quality of Instruction instrument, focusing in particular on the effects of varying the rater pool, subject matter content, observation procedure, and district context. Variation in the subject matter content of lessons did not affect rater agreement with master scores, but the evaluation of other portions of the validity argument varied according to the composition of the rater pool, observation procedure, and district context. These results demonstrate the need for conducting such analyses, especially for classroom observation instruments that are subject to multiple sources of variation.

Correspondence should be sent to Heather C. Hill, Graduate School of Education, Harvard University, Appian Way, MA 02138. E-mail: heather_hill@gse.harvard.edu

Downloaded by [University Of Maryland] at 12:53 15 August 2017

The past 5 years have seen the increased use of classroom observation instruments to assess and improve teaching quality. Race to the Top has spurred numerous states to adopt or adapt standardized instruments, such as Framework for Teaching or the Teacher Advancement Program, as part of newly redesigned teacher evaluation systems (Sawchuck, 2009). Research studies have demonstrated that observation instruments, when used for formative teacher evaluation and tailored feedback, can successfully improve teaching quality (Allen, Pianta, Gregory, Mikami, & Lun, 2011; McCollum, Hemmeter, & Hsieh, in press). In addition, a number of content-specific instruments now enable researchers to evaluate the effects of policies and professional development programs in a given subject area (Grossman et al., 2010; Matsumura, Garnier, Slater, & Boston, 2008; Newton, 2010).

The growing use of classroom observation instruments has led to a corresponding increase in attention to validity, that is, the degree to which scores represent the underlying construct they seek to measure. To investigate validity, some researchers have focused on correlating teacher performance on these instruments with student test score gains (Grossman et al., 2010; Kane & Staiger, 2012; Kane, Taylor, Tyler, & Wooten, 2010) assuming that if these correlations are high, then a potential mechanism for promoting student learning—providing better instruction—is empirically supported. Others have begun to deploy the validity argument approach advocated by Kane (2001, 2004, 2006), which involves developing a detailed interpretation of scores and comprehensively assessing the assumptions contained in that interpretation. For example, Bell, Gitomer, McCaffrey, Hamre, and Pianta (this issue) specify and investigate an extensive set of assumptions relating teachers' scores on the Classroom Assessment Scoring System to the quality of their teaching. The validity argument approach has also received wide attention outside of classroom observation research, constituting the main focus of the most recent *Educational Measurement* handbook chapter on validity (Kane, 2006). Examples of these types of validity argument applications can be found in several fields, including educational assessment broadly (e.g., Bangert, 2009; Hill, Kapitula, & Umland, 2011; Schilling & Hill, 2007), college placement exams (e.g., Chapelle, Enright, & Jamieson, 2007), and medicine (e.g., Hawkins, Margolis, Dunning, & Norcini, 2010; McGaghie, Cohen, & Wayne, 2011).

In this article, we contend that users of classroom observation instruments must pay particular attention to whether and how much interpretations of scores are affected by variation in implementation conditions. We make this claim because, unlike most standardized assessments, classroom observation instruments function amid a range of mutable factors: different rater pools, divergent scoring designs, and diverse content areas and teacher populations. As the *Standards for Educational and Psychological Testing* (American Educational Research Association/American Psychological Association/National Council on Measurement in Education [AERA/APA/NCME], 1999) strongly suggest, the evidence supporting validity arguments may be sensitive to the populations and contexts surrounding test implementation (see Standards 1.2 and 1.4; pp. 17–18); if so, the interpretations of scores intended by researchers and practitioners may require amendment or be invalid in specific contexts.

To demonstrate the need for incorporating variability due to implementation into the assessment of validity arguments, we examine three sample validity assumptions from a mathematics observation instrument, the Mathematical Quality of Instruction (MQI). We focus in particular on assumptions regarding rater agreement with master scores, score generalizability, and correlation with criterion variables. Our results suggest that variation in implementation conditions, in fact, can significantly affect the evaluation of validity arguments.

Downloaded by [University Of Maryland] at 12:53 15 August 2017

BACKGROUND

An argument-based approach to validity has been effectively articulated by Kane (2001, 2004, 2006), who, drawing on earlier work in the area (Cronbach, 1988), has popularized a method for determining the meaning of scores from assessments. In this approach, test developers specify an interpretation for and planned use of scores and then subsequently establish an interpretive argument, a network of linked assumptions that must be true to support the proposed score interpretation and test use. Test developers then investigate these assumptions by gathering empirical evidence using a variety of methods. Although Kane (2001) argued that each interpretive argument will be unique, his own work (2006) and that of his adherents (e.g., Bell et al., this issue; McGaghie et al., 2011) tend to identify four areas requiring investigation:

- Assumptions involving *scoring*, typically focused around whether items are being used consistently and yield accurate and desired information.
- Assumptions involving *generalization*, typically focused around whether the sample of tasks and/or observations adequately represents the universe of potential observations.
- Assumptions involving *extrapolation*, typically focused around whether the assessment represents the constructs intended and the measure as a whole aligns with external indicators of examinee success in the domain(s) of interest.
- Assumptions involving *decisions*, typically focused around whether consequences based on scores from the instrument are appropriate.

Kane (2001) noted that validating an interpretive argument is an iterative process, one in which original interpretations may be revisited and revised in light of evidence collected during the empirical inquiry. He also argued that in strong interpretive arguments, test developers selectively identify the most problematic assumptions for inquiry.

These most problematic assumptions, as instantiated in interpretive arguments developed over the last decade, reveal differences in validity arguments between in situ assessments—that is, assessments designed to capture performance in examinees' natural environments—and other assessments of performance such as selected-response examinations. In particular, in situ assessments require validation techniques that help account for variance related to raters and measurement occasions. Bell et al. (this issue) and Hawkins et al. (2010) have reported results from in-depth examinations of scoring and generalizability studies for in situ assessments; both help to determine the effects of the sampling of observations and raters on scores. Analogous examinations were not observed in studies of selected-response assessments, where validity argument assumptions focused instead on the cognitive processes of respondents and the structure of the data (e.g., Chapelle et al., 2007; Schilling & Hill, 2007, on Test of English as a Foreign Language selected-response tasks).

Results from validation studies for in situ assessments also suggest that scoring assumptions may prove particularly questionable. Evidence described in Bell et al. (this issue), Hawkins et al. (2010), and Llosa (2008) shows only limited support for the assumption that raters use specific observation items consistently and accurately. By contrast, evaluations of constructed response assessments—for instance, when raters score a common task given to examinees—tended to find that raters were likely to use items consistently (e.g., Chapelle et al., 2007; Lane, Liu, Ankenmann, & Stone, 1996). One reason for this difference may be that variation owing to the

Downloaded by [University Of Maryland] at 12:53 15 August 2017

content of lessons makes the assignment of scores for in situ assessments higher inference than, for example, the scoring of a common mathematics task, leading to disagreements between raters.[1]

Problematic scoring assumptions would be challenging enough for developers of in situ assessments. However, we argue that there are still more sources of variance that should be investigated in the context of validity arguments, particularly in the case of classroom observation instruments. We base this claim on evidence that classroom observation instruments are implemented in widely diverse educational settings with differing data collection and scoring designs. For example, districts experimenting with enhanced teacher evaluation systems have implemented different arrangements regarding raters, from using a pool of qualified evaluators (e.g., Cincinnati; see Taylor & Tyler, 2011) to more traditional arrangements involving only principals as raters, which implies lower selectivity and certification standards (e.g., Tennessee; see Hill & Herlihy, 2011). Key district policies and practices, such as curriculum materials and student assessments, also vary, and that variation may affect the evaluation of validity assumptions. Researchers may wish to modify instruments, sample instruction from within lessons, or use an instrument across diverse subjects (e.g., physical education and English) or diverse content areas within the same subject matter (e.g., arithmetic and geometry), thus adding additional sources of variance beyond that stemming from raters and occasions.

Given the wide range of implementation designs and contexts, an important issue emerges: whether and how much variation in these contexts affects the evaluation of validity arguments. We contend that answering this question should be a key part of evaluating the interpretation of scores and corresponding assumptions derived from classroom observational instruments.

METHODOLOGY

The overall approach in this article is to specify assumptions for a validity argument and then, for a selected set of assumptions, examine whether hypothetical differences in implementation of an instrument—in this case, the MQI—lead to different evaluations of those assumptions. Our goal is not to assess the validity argument for a particular instrument, let alone to do so comprehensively. Nor do we have readily accessible data that would allow us to examine wide differences in implementation contexts. Instead, we aim to illustrate the need for such investigations by demonstrating that even with modest variation in implementation contexts—in this case, within the same or closely related studies—the evaluation of validity assumptions changes. In preparation for these analyses, in this section we describe the instrument, our interpretation of scores, and our approach to these analyses.

Instrument and Intended Uses

The MQI instrument is designed to provide information about the quality of teachers' enactment of mathematics instruction and, as designed, includes four major dimensions: the Richness of

[1] This is also the case in portfolio assessments where the content of the portfolio entries can vary within and across classrooms (see, e.g., Koretz, Stecher, Klein, & McCaffrey, 1994).

the Mathematics (Richness), Working with Students and Mathematics (Working with Students), Teacher Errors and Imprecision (Errors), and Student Participation in Meaning-Making and Reasoning (SPMMR).[2] For all dimensions except Errors, higher scores indicate better performance; for Errors, higher scores indicate more problematic instruction. Each dimension consists of between two and five items that identify specific behaviors, as well as a "holistic" item that records each segment of a lesson's overall quality on that dimension. The instrument is currently intended for use with videotaped lessons of elementary and middle school classroom mathematics instruction (Hill et al., 2008; Hill, Kapitula, & Umland, 2011).

Interpretation of MQI Scores

Specifying an interpretation of scores is the first step in constructing a validity argument. Briefly, we intend a teacher's scores to represent the quality of his or her instruction, in that the scores capture characteristics of the mathematics teaching during that teacher's lessons. This includes the extent to which the teacher presents mathematical ideas and procedures correctly and with meaning, elicits and develops students' own mathematical thinking and reasoning, and engages students in cognitively demanding activities such as questioning and reasoning. To support this interpretation of scores, several conditions would have to be met: raters would have to score specific items accurately, consistently, and without bias; scores from specific items, segments, and lessons should generalize to reliably represent teachers' overarching mathematical quality of instruction within the constraint of feasible data collection plans; teachers' scores should also correlate with the outcomes of teaching (i.e., student learning) as well as with the use of resources theorized to produce those scores, including teacher knowledge, preparation for teaching, school resources, and curriculum materials (see Cohen, Raudenbush, & Ball, 2003; Kennedy, 2010, for a related discussion). These assumptions are organized and summarized in Table 1 under the four basic assumptions included in Kane's (2006) framework. For the purposes of this article, we select three assumptions (i.e., Assumptions 1.1, 2.1, and 3.3) and use them to investigate whether differences in the implementation of the MQI might lead to differences in the evaluation of validity assumptions. We continue by briefly outlining the data and the analyses used to this end.

Data

The results that follow derive from analyses of data from three studies. The first consists of a Generalizability study (G-study) conducted for the MQI; we use this data to demonstrate that commonly implemented changes to scoring designs may affect the generalizability of scores (Assumption 2.1). The second and third studies both examine mathematics instruction in relationship to criterion variables, such as teacher knowledge and student outcomes. These studies, which have nearly the same scoring design and share a common rater pool, provide information on the effect of the quality of the rating pool on a scoring assumption (Assumption 1.1) and on the robustness of an extrapolation assumption (Assumption 3.3) to changes in district context. For these latter two analyses, we combine data from the latter two studies.

[2] A fifth major dimension, Classroom Work is Connected to Mathematics, consists of a single item and is not discussed here.

Downloaded by [University Of Maryland] at 12:53 15 August 2017

Downloaded by [University Of Maryland] at 12:53 15 August 2017

TABLE 1
Validity Argument and Assumptions

1. Scoring assumptions
    1.1. Raters' understanding and use of the items are accurate, in that they coincide with instrument developers' understanding and use of the items.
    1.2. Raters can consistently apply the items when scoring instruction.
    1.3. Raters use the items without bias in that the same instructional behaviors and quality enacted by diverse teachers would be regarded similarly.
2. Generalizability assumption
    2.1. Sufficient variance lies at the teacher level (as opposed to the rater, lesson, or measurement level) so that, under reasonable budget constraints, reliable estimates for the mathematical quality of teachers' instruction can be achieved.
3. Extrapolation assumptions
    3.1. MQI items are grouped into four distinct factors that represent the four theoretical dimensions of the MQI. These factors then form a second-order factor that represents the overall mathematical quality of instruction.
    3.2. Teachers' scores represent their mathematical quality of instruction, as opposed to classroom climate, pedagogical expertise, or their students' characteristics.
    3.3. MQI scores are related to teachers' internal resources and practices, including their MKT, preparation, and habits, and to school and curricular resources.
    3.4. Higher MQI scores are positively related to student gains.
    3.5. Items represent important aspects of the mathematical quality of instruction occurring in classrooms.
4. Decision assumptions
    4.1. Feedback and advice to teachers based on MQI scores appropriately reflects key teacher weaknesses and strengths.
    4.2. Conclusions reached in the context of large-scale studies using the MQI are valid, in that the instrument performs at a minimum level of reliability and accuracy.

Note. MQI = Mathematical Quality of Instruction; MKT = Mathematical Knowledge for Teaching.

Next we elaborate upon the data and data collection processes for each study, then describe the analyses to be performed; a summary of these studies is also presented in Table 2.

Generalizability Study

The first type of data comes from a G-study intended to determine an efficient scoring design for the MQI.

TABLE 2
Study Details

	Grade Level	*Districts N*	*Teachers N*	*Lessons per Teacher (Design)*	*Raters N*
Generalizability study	6–8	1	8	3	10
NCTE	4–5	4	244	3	33
LMT:MSIS	4–5	1	38	6	33

Note. NCTE = National Center for Teacher Effectiveness; LMT:MSIS = Learning Mathematics for Teaching Math Solutions Impact Study.

Sample of teachers and lessons. From a pool of 24 middle school mathematics teachers in one district, we sampled eight with different levels of mathematical knowledge for teaching (see Hill et al., 2011, for more details on the larger study). Six videotaped lessons were available per teacher; from those, we sampled three lessons per teacher that were approximately equal in length (i.e., each of the sampled lessons contained between six to eight 7.5-min segments). Because of the small sample employed in this G-study, the results presented in the next section should be considered exploratory.

Sample of raters. Ten mathematics education graduate students and former teachers were recruited via e-mail to colleagues in mathematics education departments. Raters attended a 2-day intensive training on the instrument. At the end of training, raters took a certification exam, in which they were asked to score 16 segments from lessons taught by four different teachers. Based on these results, one rater whose scores did not meet the certification threshold was excluded from the analysis presented below.

Lesson scoring. Each rater watched and scored the 24 lessons just described (three lessons per each of eight teachers). Following our coding protocol, raters "skimmed" each lesson once, then during the second watch, assigned scores for each MQI item for every 7.5-min segment of the lesson. The raters did so by using a 3-point scale (low, medium, high).

Large-Scale Studies

The second type of data derives from nearly identical data sets under construction by the National Center for Teacher Effectiveness (NCTE; $n = 244$ teachers) and Learning Mathematics for Teaching Math Solutions Impact Study (LMT:MSIS; $n = 38$ teachers) projects, efforts to study mathematics instruction and mathematics-focused professional development, respectively.

Sample of teachers and lessons. Both studies took place with samples of fourth- and fifth-grade teachers; recruitment for both studies proceeded in a similar manner. First, we presented study information to teachers in 93 schools from five districts; schools were chosen largely on district nominations and size, as NCTE required a minimum of two teachers at each of the sampled grades. Among eligible teachers within these schools, 55% ultimately agreed to participate in the study. Teachers participating in each project consented to have a sample of their mathematics lessons videotaped (three and six per teacher, for NCTE and LMT:MSIS, respectively). Teachers were allowed to select the dates for videotaping in advance; we asked only that we see a typical lesson and that teachers exclude days on which students would be taking tests or exams. Data collection days—or pairs of days, in the case of LMT:MSIS—were separated by at least 2 weeks to ensure variation in the content taught. Consequently, the lessons used in the analyses presented next varied in terms of length, content, time of day, and the part of the school year in which they were filmed. We videotaped lessons using a three-camera rig supplied by thereNow; video quality was high, and transcripts of lessons revealed that nearly all teacher and student talk was captured. Once captured, video was streamed for raters.

Downloaded by [University Of Maryland] at 12:53 15 August 2017

Downloaded by [University Of Maryland] at 12:53 15 August 2017

Sample of raters. To generate a rater pool for scoring video, we recruited individuals in a similar way to how we recruited for the G-study, including posting notices on mathematics education listservs and sending e-mails to colleagues working in mathematics education departments. We created online training modules based on the live training delivered to G-study raters; these modules contained practice videos and provided automated feedback about submitted scores. To become an "active" MQI rater for these two projects, we required both initial certification and participation in ongoing calibration seminars. These efforts yielded two datasets—one each for certification and calibration—that allow us to examine our scoring assumption. In total, we certified 38 out of 43 applicants (88%); 33 of those certified went on to work for the projects and participate in ongoing calibration sessions.

At the same time NCTE and LMT:MSIS were recruiting and training raters, a number of other projects were using the MQI training and certification website for similar purposes. Because we did not recruit these individuals, we can provide less information about their background and characteristics. However, from discussions with other projects, we know that at least some had a mathematics or mathematics education background. For one of the analyses presented next, we use certification results from 96 such individuals, 45% of whom passed certification.

Lesson scoring. As in the G-study, to score a videotaped lesson raters "skimmed" the lesson once in its entirety and then watched it a second time, scoring 7.5-min segments on each of the 17 items outlined in Table 3. Raters scored these segment-level items using a 1 (*low*) to 3 (*high*) scale. Raters also assigned each video a lesson-level score of 1 (*low*) to 5 (*high*) to represent an overall assessment of the mathematical quality of instruction (Overall MQI).

Other data. Both NCTE and LMT:MSIS also included surveys measuring teachers' Mathematical Knowledge for Teaching (MKT; Ball, Thames, & Phelps, 2008) and other background characteristics. Mathematical knowledge for teaching was measured by a set of 32 items for NCTE and 60 and 57 items on two different forms for LMT:MSIS. These items were scored using item response theory, with item response theory reliabilities for these forms at 0.87, 0.90, and 0.87, respectively. NCTE's survey also contained a number of other Likert-type items, including a set of four items measuring teachers' habits around preparing for mathematics teaching (Cronbach's $\alpha = 0.79$) and nine items measuring teachers' perceptions of the school resources available to support their teaching (Cronbach's $\alpha = 0.80$).

Analyses

Assumption 1.1 involves an examination of rater accuracy, which we define here as the extent to which raters' judgments match "true" scores, as instantiated in master scores for the same segments.[3] To test the robustness of Assumption 1.1 to variation in the rater pool, we compared the performance of project and nonproject raters on 16 certification lesson segments. For each group, rater responses were compared to master scores generated by the MQI developers. Rater accuracy is reported for the certification test as a whole (16 segments × 17 items), and raters'

[3] This use of the term "accuracy" is more restrictive than is often found in the measurement literature. We use it here to avoid the more cumbersome "rater agreement with master scores" below.

Downloaded by [University Of Maryland] at 12:53 15 August 2017

TABLE 3
Mathematical Quality of Instruction Dimensions and Items

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:

- *Linking and Connections*: Linking and connecting mathematical representations, ideas, and procedures.
- *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.

In addition, *holistic richness* captures raters' perception of the teacher's general performance on this dimension during the segment.

Working with students and mathematics: This dimension captures whether teachers can understand and respond to students' mathematically substantive productions (utterances or written work) or mathematical errors. The dimension consists of the following items:

- *Remediation of student errors and difficulties*: With this item, we mean to mark instances of remediation in which student misconceptions and difficulties with the content are *substantially* addressed.
- *Responding to student mathematical productions in instruction:* Teacher responds to student productions during instruction in mathematically appropriate ways such as identifying mathematical insights in specific student questions, comments, or work; building instruction on student ideas or methods.

In addition, *holistic working with students and mathematics* captures raters' perception of the teacher's general performance on this dimension during the segment.

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*, such as solving problems incorrectly; defining terms incorrectly; forgetting a key condition in a definition; equating two non-identical 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.

In addition, *holistic errors and imprecision* captures raters' perception of the teacher's general performance on this dimension during the segment.

Student participation in meaning-making and reasoning: This dimension captures evidence of students' involvement in cognitively activating classroom work. Attention here focuses on student participation in activities such as:

- *Providing explanations*.
- *Posing mathematically motivated questions or offering mathematical claims or counterclaims*.
- *Engaging in reasoning and cognitively demanding activities*, such as drawing connections among different representations, concepts, or solution methods; identifying and explaining patterns.

In addition, *holistic student participation* captures raters' perception of the teacher's general performance on this dimension.

Downloaded by [University Of Maryland] at 12:53 15 August 2017

performance is represented with a simple percent correct, indicating the percentage of item-segment combinations that raters scored correctly under the criteria of exact agreement with a master score.

To further examine this accuracy assumption (Assumption 1.1), we also assessed the degree to which rater accuracy is affected by content area within mathematics. To do so, we generated rater scores from 29 calibration seminars by comparing submitted scores to master scores on the same clips. Of the 29 total calibration sessions, active raters must have participated in at least three to be included in this analysis; on average, raters participated in roughly 20 sessions and scored two clips per session. Raters' scores were computed as an average percent correct during weekly calibration (calculated from 2 segments × 17 items). To examine the role of content area on raters' accuracy, we report these scores separately for two different content domains (arithmetic/algebra vs. measurement/geometry).

To test the robustness of Assumption 2.1, about the generalizability of teacher scores, we first determined the variance components attributable to teachers, lessons, and raters using a G-study with a two-facet nested design (lessons nested within teachers and crossed with raters). Using this variance decomposition, we then conducted Decision-studies (D-studies) that enabled an examination of how the generalizability coefficient changed when varying the number of raters and lessons observed per teacher (see Brennan, 2001, and Shavelson & Webb, 1991, for more information). To examine whether variations in the scoring design affected the generalizability of findings, we also compared the results of the D-studies in which the entire data set was used to the results of D-studies in which the data set was restricted to scores from the first 30 min per lesson.

To assess the robustness of Assumption 3.3, about the relationship of MQI scores to other indicators of teacher and contextual resources, we estimated correlations between teachers' MQI scores and other teacher and school characteristics. We did so first for the entire sample from both studies, then divided the data set by district/curriculum materials to demonstrate how the local context may affect evidence regarding extrapolation assumptions. For this analysis, we used Fisher's (1921) z transformation to calculate the two-tailed significance levels for comparisons between Pearson correlations.

RESULTS

Next we evaluate these three sample assumptions relating to the interpretation of MQI scores. For each assumption, we briefly present main results and then investigate the robustness of findings to variability in instrument implementation context.

Robustness of Accuracy Scoring Assumption to Rater Selection and Lesson Content

As previously noted, scoring assumptions for in situ assessments often receive only limited support, perhaps on account of the variability in the content of observations. Furthermore, we hypothesize that the evaluation of scoring assumptions may be affected by variation in contextual factors, including the available pool of raters, training and certification of these raters, and variation associated with operational scoring, such as the content of the lesson

Downloaded by [University Of Maryland] at 12:53 15 August 2017

observed. For instance, although research projects might avoid hiring (or actively remove) poor raters, organizations intending to use an instrument for coaching or evaluation must often work with existing employees—individuals who may or may not meet certification criteria. If there is a relationship between selectivity and rater quality, this might affect the evaluation of our first validity assumption (Assumption 1.1). The same would be true if rater accuracy is a function of the content area included in the lesson; MQI scoring rules may be easier to apply in some content areas than others, leading to different evaluations of Assumption 1.1, depending upon that content.

To shed light on these issues, we first compare rater scores on 16 certification test segments to master scores generated by the MQI developers. On this test, raters hired by our projects scored 73% of items correctly. This average percentage correct score is due in part to the fact that we set a cut score for hiring,[4] ensuring some degree of scoring accuracy. Although this does not meet the standard 80% bar set by many, agreement levels are much higher than those expected by random score assignment, suggesting that, to some extent, the raters' understanding and use of the items coincided with that of the developers.

More importantly, our analysis suggests that rater accuracy is likely to vary depending upon the available pool of raters and the ability of the organization implementing an instrument to select raters based on performance. We can examine the scores from a pool of raters not selectively recruited by our projects ($n = 96$) as a proxy for the situation in which a district must take all potential raters. Despite still containing many individuals with mathematics teaching backgrounds, this group scored only 66% of items correctly on the certification assessment. This suggests that, as previously hypothesized, less selectivity in rater certification means weakened support for the accuracy-related scoring assumption, that raters will understand and use items in ways similar to instrument developers.

We also examine whether raters' accuracy varies by the content of the lessons that they observe. One reason to do this is that MQI raters often note the difficulty of applying the instrument to lessons on geometry and measurement; these two domains are different from arithmetic and algebra in that they focus more on definition and mathematical proof, two areas not covered as well by the MQI. Thus, this may mimic a situation in which instruments are applied to widely disparate content areas. In this case, results suggest that raters are equally accurate in these two mathematical content domains. In weekly calibration, raters scored on average 67% of the 47 arithmetic/algebra item-segment combinations correctly and 66% of the 10 measurement/geometry item-segment combinations correctly, a negligible difference.[5] Despite finding little difference, we argue that similar tests are necessary for instruments that span wider ranges of content or even generic instruments used to code lessons in different subject matter areas. Unless convincing empirical evidence suggests that lesson content—or more broadly, subject matter content—does not affect rater accuracy, ignoring this source of variation might lead to inappropriate evaluations of the validity assumptions related to raters.

[4] The hiring cut score is based on a deviance metric, where we calculate raters' average absolute deviations from the master score. We do not use this metric here, as it provides the same picture of hiring and calibration practices as the percent correct score.

[5] There were many more item-segment combinations for arithmetic/algebra than for measurement/geometry due to the smaller number of measurement/geometry lessons in these studies.

Downloaded by [University Of Maryland] at 12:53 15 August 2017

Robustness of Generalizability Assumption to Scoring Design

Assumption 2.1 suggests that reliable teacher scores can be achieved under reasonable budget constraints. For in situ assessments, several factors influence the reliabilities of such scores including rater variability, the observations sampled from practice, and other random or systematic sources of error. G-studies are designed to estimate these sources of variance and examine their influence on the replicability of scores. Table 4 demonstrates the results from our G-study analysis. A more detailed analysis of part of the results presented in Table 4 appears elsewhere (Hill, Charalambous, & Kraft, 2012); we briefly note that the variance decomposition differs per MQI dimension, with important consequences for score generalizability. For example, the variance component for raters in the case of SPMMR is about seven times as large as that for Working with Students. As can be seen in Table 5, these differences are consequential in the D-studies, where we explore how varying the number of raters and lessons per teacher affects the generalizability coefficient (ρ). These results suggest that reliable estimates of at least $\rho = 0.70$ can be obtained for most dimensions with three lessons per teacher scored by two raters each; the only exception is Working with Students, for which it would be necessary to either employ four raters per lesson or employ three raters and increase the number of observed lessons to four per teacher. Thus, for users who can implement the training and use procedures similar to those followed in our G-study, we recommend a two-rater, three- or four-lesson scoring design.

Yet, given existing variation in the implementation of observational instruments for both research and practice, we expect that this advice will not be consistently followed; in fact, we can use the results from the D-study to examine the effect of departing from this advice on the evaluation of this validity assumption. Examining the "whole lesson" scoring condition in Table 5, we easily can see that a different design would yield markedly different reliabilities for most dimensions. For instance, for Errors, using only a single rater to observe only a single lesson—as is often the case with a principal observing a teacher only once—would yield a reliability estimate less than half that obtained when having three lessons per teacher scored by three raters each. Such differences are not trivial, especially given the variety in

TABLE 4
Variance Decomposition for the Four Dimensions of the Mathematical Quality of Instruction (Average Across Items)

	Richness	*Errors and Imprecision*	*Student Participation in Meaning-Making and Reasoning*	*Working With Students and Mathematics*
Teachers (t)	42.52	31.88	32.78	27.56
Lessons: teacher (l:t)	10.52	8.81	7.22	10.25
Raters (r)	6.17	13.04	28.58	4.56
Teachers*Raters (t*r)	7.83	6.45	0.00	12.87
Residual ((l:t)*r, e)	32.97	39.82	31.43	44.77
Total	100.00	100.00	100.00	100.00

Note. Cells represent the percentage of variance explained by different facets in a G-study. Source is Generalizability study.

Downloaded by [University Of Maryland] at 12:53 15 August 2017

TABLE 5
Comparison of the Reliability Estimates (ρ) for Whole Lesson versus the First Thirty Minutes of a Lesson

	Richness		*Errors and Imprecision*		*Student Participation in Meaning-Making and Reasoning*		*Working With Students and Mathematics*	
	Whole Lesson	*30 Min*	*Whole Lesson*	*30 Min*	*Whole Lesson*	*30 Min*	*Whole Lesson*	*30 Min*
One lesson								
1 Rater	0.45	0.50	0.37	0.34	0.46	0.32	0.29	0.32
2 Raters	0.58	0.59	0.50	0.46	0.59	0.41	0.41	0.42
3 Raters	0.64	0.63	0.57	0.53	0.65	0.45	0.48	0.47
4 Raters	0.67	0.65	0.61	0.57	0.68	0.48	0.53	0.49
Two lessons								
1 Rater	0.59	0.65	0.51	0.49	0.63	0.49	0.41	0.47
2 Raters	0.71	0.73	0.64	0.62	0.74	0.58	0.55	0.58
3 Raters	0.76	0.77	0.71	0.68	0.79	0.62	0.62	0.63
4 Raters	0.79	0.78	0.74	0.71	0.81	0.65	0.66	0.66
Three lessons								
1 Rater	0.66	0.73	0.58	0.57	0.72	0.59	0.47	0.56
2 Raters	0.77	0.80	0.71	0.70	0.81	0.68	0.61	0.66
3 Raters	0.81	0.83	0.77	0.75	0.85	0.71	0.68	0.71
4 Raters	0.84	0.84	0.80	0.78	0.87	0.73	0.73	0.74
Four lessons								
1 Rater	0.69	0.77	0.63	0.63	0.77	0.66	0.51	0.61
2 Raters	0.80	0.83	0.75	0.74	0.85	0.74	0.65	0.72
3 Raters	0.84	0.86	0.81	0.79	0.88	0.77	0.72	0.76
4 Raters	0.86	0.87	0.83	0.82	0.90	0.78	0.76	0.78

Note. Source is Generalizability study.

designs prevalent in both the research and practice worlds. Researchers are bound by budget and other constraints in selecting a scoring design; a survey conducted by Hill and Herlihy (2011) suggests that states have adopted highly variable designs as well.

We also examined the effects of watching only the first 30 min of a lesson rather than the entire lesson, something that principals might do when pressed for time or researchers might do if pressed for resources. By excluding segments after the 30-min mark from our G- and D-studies, we can simulate reliabilities in this real-life condition. As shown in Table 5, the estimated reliabilities for Richness, Working with Students, and Errors remain largely unchanged. By contrast, watching only the first 30 min of each lesson yields notably lower estimated reliabilities for SPMMR. For example, consider the reliability obtained for the two-lesson, two-rater combination ($\rho = 0.74$); to obtain this level of reliability when scoring only the first 30 min of each lesson, two additional lessons would be needed. Consequently, even this slight change in implementation requires doubling the number of lessons to be scored, a nontrivial change given the costs associated with scoring (approximately $40 per lesson) in our preferred design.

Downloaded by [University Of Maryland] at 12:53 15 August 2017

Overall, we see that variability in the implementation of the scoring design may have significant consequences for the evaluation of the generalizability assumption examined here. We imagine that other forms of variability—for instance, scoring lessons live versus on videotape or the scheduling of observations (e.g., observing lessons taught the same week vs. spread across a school year)—may have similar effects.

Robustness of Extrapolation Assumption to Local Contexts

Most views of teaching consider what occurs in classrooms to be the result of several factors, including teachers' own knowledge, skills, beliefs, preparation, and planning, as well as the curriculum materials from which they work and district/school resources and policies (see Bell et al., this issue; Cohen et al., 2003; Kennedy, 2010; Rowland, Turner, Thwaites, & Huckstep, 2009). This often leads those constructing validity arguments to develop assumptions that specify these relationships. For illustrative purposes, in this section we focus on Assumption 3.3 (see Table 1), which states that MQI scores are related to teachers' resources and practices (e.g., MKT, preparation for teaching, school resources). If this assumption holds, then one would expect to see a relationship between observational scores and the resources available for teaching. However, these correlations between MQI scores, teacher MKT, and other factors may also be sensitive to district contexts and curriculum use, potentially changing the evaluation of this assumption depending upon location. Curriculum materials that are highly supportive of teachers' use of multiple methods and explanations, for instance, may moderate the relationship between MKT and MQI; teachers may have less need to construct multiple methods or explanations by drawing on their own knowledge, relying instead on that supplied by their textbook. Similarly, a school's general resources may matter little for the mathematical quality of instruction except in districts where teachers report very low levels of such resources. In this case, resources may matter for the worst-off teachers in the sample.

To explore these issues, we use data from both our projects to examine the extent to which contextual variability poses problems for this extrapolation assumption. This analysis is benefitted by the fact that videos were scored by the same pool of raters using the same protocol; raters were, in fact, blind to district and project during scoring. For our first analysis, we correlated average scores representing the four MQI dimensions (Richness, Errors, Working with Students, and SPMMR) with teachers' scores on the MKT. When pooling data from these two projects, the correlation between the MKT and MQI is, in absolute values, between 0.26 (SPMMR) and 0.44 (Overall MQI; see Table 6). This would indicate moderate correspondence between MQI scores and MKT, a major resource thought to shape instruction (Ball et al., 2008). However, Table 7 shows that these correlations vary by district/curriculum. Because district is collinear with curriculum in these data sets, we cannot determine the unique effect of each; however, we can see that for Richness, for example, correlations between MQI and MKT are between 0.13 and 0.46. Similar variability in the correlations is seen for Working with Students and SPMMR. Without extensive analyses of curricula and district environments, we cannot explain why correlations with MKT dimensions varied so widely. However, we note that the variability in correlation is substantively large, ranging from small to moderate. In addition, examining the significance of pairwise comparisons of correlations using Fisher's z transformation and a loose criterion for significance ($p < .10$) in recognition of the small sample size of curriculum/district combinations, there is some evidence of heterogeneity in three of

Downloaded by [University Of Maryland] at 12:53 15 August 2017

TABLE 6
Correlations Between MKT and MQI Dimensions

MQI Dimension	*MKT*[a]
Richness	0.34***
Working with Students	0.31***
Errors and Imprecision	−0.37***
SPMMR	0.26***
Overall MQI	0.44***

Note. Source is National Center for Teacher Effectiveness and Learning Mathematics for Teaching Math Solutions Impact Study data. MKT = Mathematical Knowledge for Teaching; MQI = Mathematical Quality of Instruction; SPMMR = Student Participation in Meaning-Making and Reasoning.

[a]$n = 291$.

*$p < .05$. **$p < .01$. ***$p < .001$.

these contrasts (e.g., MKT and Richness for Investigations vs. Everyday Mathematics). This finding supports our hypothesis that, for the MQI, validity assumptions regarding extrapolation may be sensitive to context.

Other items on the NCTE survey also provide evidence that the relationship between MQI and expected contributors to the mathematical quality of instruction may be context specific. Analyzed as a whole, the data show no significant correlation between overall MQI and teacher reports of spending time preparing for teaching mathematics ($r = -0.01$), nor between MQI and

TABLE 7
MKT and MQI Correlations by Curriculum/District

	Richness	*Working With Students*	*Errors*	*SPMMR*	*Overall MQI*
Investigations (n = 104; 2 districts)					
MKT	0.46***	0.42***	−0.40***	0.36***	0.50***
Resources	0.01	0.00	0.18	0.08	−0.01
Everyday Mathematics (n = 32; 1 district)					
MKT	0.25	0.20	−0.48**	0.26	0.28
Resources	0.03	0.13	−0.10	−0.07	0.20
Harcourt materials (n = 90; 1 district)					
MKT	0.26*	0.22*	−0.29**	0.11	0.44***
Resources	0.12	−0.04	0.01	0.04	0.07
Math Expressions (n = 38; 1 district)					
MKT	0.13	0.25	−0.48**	0.34*	0.40*
Resources	n/a	n/a	n/a	n/a	n/a

Note. Source is National Center for Teacher Effectiveness and Learning Mathematics for Teaching Math Solutions Impact Study data. MKT = Mathematical Knowledge for Teaching; MQI = Mathematical Quality of Instruction; SPMMR = Student Participation in Meaning-Making and Reasoning.

*$p < .05$. **$p < .01$. ***$p < .001$.

Downloaded by [University Of Maryland] at 12:53 15 August 2017

general school resources for supporting teaching (e.g., time for professional growth, materials, and respect for teachers; $r = .00$), as reported on teacher surveys. Yet, for the Everyday Mathematics district, the latter relationship appears more substantial ($r = .20$), although not statistically significant. It is worth reporting that scores on the resource measure for the district using Everyday Mathematics were, on average, roughly 1 full standard deviation below those for other districts, suggesting that resources may become important in the production of MQI only when there are fewer in a given location.

This variability in correlations between the MQI and resources suggests that the evaluation of this extrapolation assumption may be quite sensitive to district contexts or curriculum materials. This idea is further supported by the fact that in two pilot studies ($n = 10$ elementary teachers and 25 middle school teachers), we found correlations between MQI and MKT dimensions of between 0.30 to 0.80 and 0.30 to 0.70, respectively (Hill et al., 2008; Hill, Umland, Litke, & Kapitula, 2012), substantially higher than those previously reported here. This suggests that the evaluation of such validity assumptions across multiple implementation contexts is critical.

CONCLUSION

The results of our evaluations of the MQI validity assumptions in light of hypothesized or actual variation in contextual factors demonstrate the importance of assessing the robustness of a validity argument in the specific contexts in which the instrument will be used. In these analyses, we found that less rigorous rater selection depressed rater accuracy and weakened support for the associated scoring assumption. Support for the generalizability assumption varied by the scoring design, including whether the rater viewed the first 30 min of the lesson or the entire lesson. Likewise, we found that an extrapolation assumption may be sensitive to either the curriculum materials used or the district policies and resources available to teachers. In fact, with the exception of the mathematical content observed, we found that contextual factors exerted a significant influence on the evaluation of the assumptions contained in our validity argument.

These results are consistent with our findings reported elsewhere. In an investigation focused on score generalizability and scoring design rather than validity (Hill, Charalambous, et al., 2012), in which we compared scores assigned after a first and second watch of the video, we saw that during the second watch raters were more likely to capture and score an error and slightly less likely to record Richness behaviors. This suggests that developers might investigate the impact of data collection mode (live or videotaped) on the evaluation of validity assumptions. The relationship between scores on the MQI and student outcomes also appeared to vary across studies and different student assessments (Hill et al., 2011; Kane & Staiger, 2012).

As previously noted, we found these context-specific differences in evaluation of the validity assumptions not in the context of a widely diverse set of studies but in studies that were for the most part conducted by instrument developers, using common methods and foci. That, in turn, suggests that even modest variations in implementation contexts may affect the evaluation of validity assumptions. This is a key point for researchers and evaluators planning to use a classroom observation instrument: Inferences based on scores from the instrument may be more or less valid depending upon choices made during implementation (e.g., qualifications of the rater pool, scoring design) and upon the contexts in which the instrument is used. Along with the recommendations made in the *Standards for Educational and Psychological Testing* (AERA/APA/NCME, 1999, see Standard 1.4, p. 18), this, in turn, suggests that although it is not common practice to reassess the validity of scores arising from classroom observational instruments, such reassessments may yield potentially important information, particularly in cases where scores will be used to make high-stakes decisions. It also implies that, when reporting on validity, researchers and evaluators should document carefully the conditions under which a classroom observation instrument is implemented. Doing so will enable the identification of different contextual factors that influence the characteristics of reported scores; amassing evidence on the influence of these factors will support the measurement community in determining aspects of implementation that are most likely to affect score validity, and hence, are important to specify.

Given this, and given the practical barriers to each state or district conducting an investigation like the one outlined here, we think it necessary for the measurement community to develop a method that will systematically assess the degree to which variability in implementation contexts affects the evaluation of assumptions derived from score interpretations. The first step in this method could entail identifying elements most likely to affect the score interpretations. Previously (Hill, Charalambous, et al., 2012), we have argued for viewing classroom observation measures not as instruments but as *systems*, which we defined as a collection of elements that together produce scores representing individual teachers' instructional quality. Here, in a consideration of validity more broadly, we argue that the idea of classroom observation systems can again provide a useful heuristic for locating different elements the modification of which might have an impact on validity. These elements include the observational instrument itself, the set of raters conducting the observations, rater training and certification, the scoring design (e.g., number of observations, raters per observation), and the scoring conditions (e.g., scoring live lessons vs. videotaped lessons, scoring part of the lesson or the entire lesson). These seem prime candidates for inclusion in studies seeking to understand how variability in local contexts affects score interpretations.

The second step in developing a method for determining the sensitivity of validity evidence to contexts could involve modifying one aspect in this system while keeping others constant, then isolating its influence on validity. In this way, developers could determine how changing rater qualifications and certification standards affects multiple aspects of the validity argument. This second step might also take the form of some sort of sensitivity analysis, where the goal is not only to explore whether such alterations produce differences in support for validity assumptions (as shown in the present article) but also, and more important, to determine the *level of robustness* of the validity assumptions to these modifications. Following Rosenbaum's (2002) work, this sensitivity analysis could aim to determine the degree of variation in the component of interest that would lead to invalidating, or qualitatively altering, a validity assumption. For instance, such analysis could help determine minimum rater qualifications and certification standards so that acceptable rater accuracy is still obtained. These sensitivity analyses could be based on empirical modifications of the component of interest (e.g., recruiting raters that differ in their knowledge) or on thought experiments, as is the case with the D-studies previously discussed.

Whatever the method, the rapid pace with which education systems are adopting observational instruments into their teacher development and evaluation systems, and attaching stakes to the scores from these instruments, places a sense of urgency on constructing more robust validity

Downloaded by [University Of Maryland] at 12:53 15 August 2017

arguments. Such observational instruments do have the potential to improve practice (Taylor & Tyler, 2011). However, if these instruments are implemented poorly and thus discredited—either internally among schooling organizations or externally, through media attention or legal action—that promise will diminish. Analyses of the extent of score validity under different conditions can help guide researchers, administrators, and policymakers in selecting the best approach to the process of implementing classroom observational systems.

ACKNOWLEDGMENTS

This research was supported by the William T. Grant and Spencer Foundations. Data collection was supported by the National Center for Teacher Effectiveness through Institute of Education Sciences grant #R305C090023, the Measures of Effective Teaching project, and National Science Foundation grant #0918383. We thank Rip Correnti, Joan Herman, and Felipe Martinez for assistance with an earlier draft of the paper, and Nina Cohodes for help assembling the final product.

Downloaded by [University Of Maryland] at 12:53 15 August 2017

REFERENCES

Allen, J. P., Pianta, R. C., Gregory, A., Mikami, A., & Lun, J. (2011). An interaction-based approach to enhancing secondary school instruction and student achievement. *Science*, *333*, 1034–1037.

American Educational Research Association/American Psychological Association/National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: AERA.

Ball, D. L., Thames, M. H., & Phelps, G. (2008). Content knowledge for teaching: What makes it special? *Journal of Teacher Education*, *59*, 389–407.

Bangert, A. W. (2009). Building a validity argument for the community of inquiry survey instrument. *The Internet and Higher Education*, *12*, 104–111.

Bell, C., Gitomer, D., McCaffrey, D., Hamre, B., Pianta, R., & Qi, Y. (2012/this issue). An argument approach to observation protocol validity. *Educational Assessment, 17*, 62–87.

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

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.). (2007). *Building a validity argument for the Test of English as a Foreign Language*™. New York, NY: Routledge.

Cohen, D. K., Raudenbush, S., & Ball, D. L. (2003). Resources, instruction and research. *Educational Evaluation and Policy Analysis*, *25*, 119–142.

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

Fisher, R. A. (1921). On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron*, *1*(4), 3–32.

Grossman, P., Loeb, S., Cohen, J., Hammerness, K., Wyckoff, J., Boyd, D., & Lankford, H. (2010, May). *Measure for measure: The relationship between measures of instructional practice in middle school English Language Arts and teachers' value-added scores.* (NBER Working Paper 16015). Cambridge, MA: National Bureau of Economic Research.

Hawkins, R. E., Margolis, M. J., Dunning, S. J., & Norcini, J. J. (2010). Constructing a validity argument for the mini-clinical evaluation exercise: A review of the research. *Academic Medicine*, *85*, 1453–1461.

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

Hill, H. C., Charalambous, C. Y., & Kraft, M. (2012). When rater reliability is not enough: Observational systems and a case for the G-study. *Educational Researcher*, *41*(2), 56–64.

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.

Hill, H. C., Umland, K. U., Litke, E., & Kapitula, L. (2012). Teacher quality and quality teaching: Examining the relationship of a teacher assessment to practice. *American Journal of Education, 118*, 489–519.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, *38*, 319–342.

Kane, M. T. (2004). Certification testing as an illustration of argument-based validation. *Measurement: Interdisciplinary Research and Perspectives*, *2*, 135–170.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education/Praeger.

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains.* Seattle, WA: Bill & Melinda Gates Foundation. Retrieved from http://www.metproject.org/reports.php

Kane, T. J., Taylor, E. S., Tyler, J., & Wooten, A. (2010, March). *Identifying effective classroom practices using student achievement data* (NBER Working Paper 15803). Cambridge, MA: National Bureau of Economic Research.

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

Koretz, D., Stecher, B., Klein, S., & McCaffrey, D. (1994). The Vermont Portfolio Assessment Program: Findings and implications. *Educational Measurement: Issues and Practice, 13*(3), 5–16.

Lane, S., Liu, M., Ankenmann, R. D., & Stone, C. A. (1996). Generalizability and validity of a mathematics performance assessment. *Journal of Educational Measurement*, *3*, 71–92.

Llosa, L. (2008). Building and supporting a validity argument for a standards-based classroom assessment of English proficiency based on teacher judgments. *Educational Measurement: Issues and Practice*, *27*(3), 32–42.

Matsumura, L. C., Garnier, H. E., Slater, S. C., & Boston, M. D. (2008). Toward measuring instructional interactions "at-scale." *Educational Assessment*, *13*, 267–300.

McCollum, J. A., Hemmeter, M. L., & Hsieh, W. (in press). Coaching teachers for emergent literacy instruction using performance-based feedback. *Topics in Early Childhood Education.*

McGaghie, W. C., Cohen, E. R., & Wayne, D. B. (2011). Are United States medical licensing exam step 1 and 2 scores valid measures for postgraduate medical residency selection decisions? *Academic Medicine*, *86*(1), 48–52.

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.

Rosenbaum, P. R. (2002). *Observational studies* (2nd ed.). New York, NY: Springer-Verlag.

Rowland, T., Turner, F., Thwaites, A., & Huckstep, P. (2009). *Developing primary mathematics teaching: Reflecting on practice with the Knowledge Quartet*. London, UK: Sage.

Sawchuck, S. (2009). New teacher-evaluation systems face obstacles: Stimulus funds require districts to revamp teacher yardsticks. *Education Week*. Retrieved from http://www.edweek.org/ew/articles/2009/12/11/15evaluate.h29.html

Schilling, S. G., & Hill, H. C. (2007). Assessing measures of Mathematical Knowledge for Teaching: A validity argument approach. *Measurement: Interdisciplinary Research and Perspectives*, *5*, 70–80.

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

Taylor, E. S., & Tyler, J. H. (2011, March). *The effect of evaluation on performance: Evidence from longitudinal student achievement data of mid-career teachers* (NBER Working Paper 16877). Cambridge, MA: National Bureau of Economic Research.

Downloaded by [University Of Maryland] at 12:53 15 August 2017