Using Video Clips of Mathematics Classroom Instruction as Item Prompts to Measure Teachers' Knowledge of Teaching Mathematics

Nicole Kersting

Educational and Psychological Measurement published online 25 February 2008
DOI: 10.1177/0013164407313369

Educational and Psychological Measurement
Volume XX Number X
Month XXXX xx-xx

10.1177/0013164407313369
http://epm.sagepub.com
hosted at
http://online.sagepub.com

Using Video Clips of Mathematics Classroom Instruction as Item Prompts to Measure Teachers' Knowledge of Teaching Mathematics

Nicole Kersting
LessonLab Research Institute

Responding to the scarcity of suitable measures of teacher knowledge, this article reports on a novel assessment approach to measuring teacher knowledge of teaching mathematics. The new approach uses teachers' ability to analyze teaching as a proxy for their teaching knowledge. Video clips of classroom instruction, which respondents were asked to analyze in writing, were used as item prompts. Teacher responses were scored along four dimensions: mathematical content, student thinking, alternative teaching strategies, and overall quality of interpretation. A prototype assessment was developed and its reliability and validity were examined. Respondents' scores were found to be reliable. Positive, moderate correlations between teachers' scores on the video-analysis assessment, a criterion measure of mathematical content knowledge for teaching, and expert ratings provide initial evidence for the criterion-related validity of the video-analysis assessment. Results suggest that teachers' ability to analyze teaching might be reflective of their teaching knowledge.

Keywords: *teacher content knowledge; mathematics; assess; video clips; validity*

There is a growing interest in understanding what knowledge is required for effective teaching and how it can be assessed. Although progress has clearly been made in defining such knowledge (Ball & Bass, 2000; Shulman, 1986, 1987), few advancements have been used so far in developing new assessments of teachers (Darling-Hammond & Baratz-Snowden, 2005). Many currently available tools to

Author's Note: The author is grateful to Noreen Webb for her support throughout the life of this project and to the teachers and professors who participated in it. This research was completed in partial fulfillment of the requirements for the Doctor of Philosophy degree at the University of California, Los Angeles. Portions of this research were presented at the annual meeting of the American Educational Research Association, April 2006, in San Francisco, California, and at the biannual meeting of the European Association for Research on Learning and Instruction, August 2005, in Nikosia, Cyprus. Finally, the author is indebted to Karen Givvin, Ronald Gallimore, and Rossella Santagata, who patiently reviewed this manuscript and provided many helpful comments and edits. Please address correspondence to Nicole Kersting, LessonLab Research Institute, Santa Monica, CA 90405; e-mail: nicolek@lessonlab.com.

measure teacher knowledge have limitations. Some have shown little or no relationship with teaching practice and student learning, and for others criterion-related validity has not been formally investigated (Darling-Hammond & Baratz-Snowden, 2005; Haertel, 1991). Others still are too long to be practical outside their original assessment context. Many measures have little face validity. As the need for assessments is growing in the current climate of educational reform and heightened accountability (No Child Left Behind [NCLB], 2001), with many initiatives focused on teachers and their knowledge, suitable measures have remained scarce. This article reports on the development and validation of a video-based assessment approach to measure teachers' knowledge of teaching mathematics.

Traditionally, teacher knowledge has been assessed through multiple-choice tests that measure subject matter content knowledge, general pedagogical knowledge, or classroom management skills, providing little information on how well a respondent is able to teach a given subject to students. Such assessments have been favored in high-stake decision contexts (e.g., licensing) because the forced answer format ensures stable psychometric properties, easy test administration, and low scoring cost. Although these tests unarguably measure some aspect of teacher knowledge, there is growing agreement that they assess only "low-level or marginally relevant knowledge and skills [and] not the candidate's deep knowledge of subject matter and actual teaching skills" (Darling-Hammond & Baratz-Snowden, 2005, p. 61).

Assessments developed more recently are more congruent with current views on effective instruction and the knowledge it requires. For example, pedagogical content knowledge, which is now considered a core domain within teacher knowledge related to quality of teaching and student learning (Hill, Rowan, & Ball, 2005), has been included in the blueprints of several newer assessments (Hill, Schilling, & Ball, 2004; Porter, Youngs, & Odden, 2001). Pedagogical content knowledge refers to a teacher's knowledge of subject matter content in ways that are useful for teaching (Ball, 2000). Beyond mastery of the content itself, it requires knowledge of how to best represent the content to learners, including knowledge of common student misconceptions or errors (Shulman, 1986, 1987). However, other critical aspects of teacher knowledge, among them the knowledge that is required to make real-time instructional decisions to support students with different learning goals in the same classroom (Ball & Bass, 2000, Cochran-Smith & Lytle, 1999), have yet to be operationalized for assessment purposes. These aspects of teacher knowledge address the heterogeneity, complexity, and contextual nature of real classroom instruction.

Many of these newer assessments consist at least in part of constructed response items, which historically have produced lower reliability estimates than multiple-choice tests because of interrater inconsistencies. Although for many of these newer assessments better measurement properties (i.e., reliability and validity) have been reported for their scores (Porter et al., 2001), for others questions regarding their technical quality (i.e., criterion-related validity, bias, and adverse impact) have

remained (Bond, 1998; Jaeger, 1998; Klein, 1998; Pool, Ellett, Schiavone, & Carey-Lewis, 2001; Stone, 2002). Better measurement is often achieved by substantially increasing the number of items and/or assessment tasks. This has led to much longer completion times that in combination with more complex administration and high scoring costs associated with constructed response items, will limit their use in other assessment contexts. Taken together, all of these problems highlight the need for exploring new assessment approaches.

Research on expertise in cognitive psychology and education might provide new directions for developing measures that reflect current views on teacher knowledge. In a series of experimental studies, expert teachers were found to systematically perceive and interpret classroom events differently from novices. For example, when asked to view and interpret videotaped classroom instruction, experts provided more coherent and richer interpretations of the observed teaching than did novice teachers. Moreover, different from novices, expert teachers were able to identify key instructional moments and offer alternative approaches (Carter, Cushing, Sabers, Stein, & Berliner, 1988; Carter, Sabers, Cushing, Pinnegar, & Berliner, 1987; Sabers, Cushing, & Berliner, 1991). These findings not only confirm that experts organize, store, and access their knowledge in ways different from novices (Bransford, Brown, & Cocking, 1999), but that measuring teachers' ability to analyze teaching may represent a promising approach to assess their knowledge of teaching.

The goal of this study then was to explore a novel approach to measure knowledge of teaching mathematics that uses teachers' ability to analyze teaching, and to examine its reliability and criterion-related validity. Video clips of classroom instruction that teachers were asked to view and to respond to in writing served as stimuli (i.e., item prompts) to elicit their knowledge of teaching. Using video clips of classroom instruction provides the opportunity to preserve the complexity and contextual nature of classroom teaching and to operationalize these two important aspects of teacher knowledge for assessment purposes.

Method

Sample

Sixty-two volunteer mathematics teachers of diverse backgrounds and professional experiences participated in this study. Of those, 62 completed the video-analysis assessment, 56 completed a criterion measure, 58 completed a background survey, and expert ratings were obtained for 53 teachers.

Teachers in this sample were sufficiently diverse. Teachers varied with respect to credential status and licensing, undergraduate college degree, number of years of teaching experience, and participation in mentorship programs. Almost all teachers in the sample (88%) taught mathematics during the year of the study. About half

(48%) reported to be fully credentialed. Mathematics teaching experience ranged from no experience to up to 33 years, with a mean of 6.4 years. About one third of teachers (36%) responded to have majored in mathematics; another 9% reported a mathematics minor. The remaining participants held college degrees as diverse as African American studies, engineering, or psychology. Seven teachers (12%) responded that they had been part of a mentorship program either as a mentor or as a mentee.

Seen in a different way, teachers varied with respect to their professional training. Nineteen teachers were first-year students in a teacher credential program at a large state university, 17 were in their last year of the teacher credential program, another 17 teachers were enrolled in a master's program of mathematics education, and 9 teachers held regular teaching assignments in a large urban school district. Presumably, the teachers represented a considerable range of teacher knowledge and teaching expertise.

Data Collection and Procedures

Data were collected on several occasions and in different locations. Teachers who were enrolled in a university program completed the video-analysis assessment, a criterion measure, and a background questionnaire in a computer lab at the university site. The video-analysis assessment and the background questionnaire were completed online through a password-protected Web site; the criterion measure of mathematical content knowledge for teaching was administered as a paper-and-pencil instrument. For teachers with regular teaching assignments, these data were collected at their respective school sites. On all data collection occasions, teachers were given up to 3 hours to complete the tasks. Teachers were provided with on-site technical support to control for effects of respondents' computer skills on task completion.

Measures

Four measures were used in this study: (a) the video-analysis assessment, (b) a measure of mathematical content knowledge for teaching to serve as a criterion, (c) expert ratings of teachers' knowledge of teaching mathematics, and (d) a background survey. Each of these will be described in turn.

Video-analysis assessment. The video-analysis assessment that was developed for this study consisted of a set of 10 selected video clips of mathematics instruction that were available for viewing via an interactive platform over the Internet. All video clips were excerpted from public-use U.S. tapes released by the TIMSS 1999 Video Study (Hiebert et al., 2003). The TIMSS 1999 Video Study collected nationally representative video samples of eighth-grade mathematics classrooms to

describe and compare mathematics instruction in seven different countries. The video clips, which varied between 1 and 3 minutes in length, featured episodes of either teacher helping behavior and assistance during private work time (THB) or student mistakes during whole-class interaction (SMS). Both types of clips were selected because (a) they presented rich teaching and learning opportunities and thus interesting prompts for analysis; (b) they represented common occurrences during instruction; and (c) considering the time constraint applicable to assessments, they were fairly short and self-contained. The video clips varied with respect to the mathematical topic area (e.g., geometry: angle theorems; algebra: fractions, variables, expressions), the complexity of the mathematical content, and the complexity of the teacher–student(s) interactions.

Along with each video clip, complementary information about the lesson and the larger learning unit was provided on the Web site. In addition, teachers were given a resource sheet showing the mathematical problems featured in the selected video clips for reference. For each clip, the task was the same. Teachers were asked to view the clip and to discuss how the teacher and the student related to each other *and* to the mathematical content. Teachers were asked to write their analytic comments in designated free-format text fields, where they were saved for later scoring.

A scoring rubric consisting of four dimensions was developed and applied to teachers' responses to the video clips. The dimensions described qualitative aspects of teachers' responses. The first three dimensions indicated whether (score of 1) or not (score of 0) a teacher's response considered central elements of the teaching and learning process: the teacher (through the inclusion of alternative teaching strategies), the mathematical content (through an analysis of the mathematical content), and the student(s) (through an analysis of student thinking/understanding). Conceptually, alternative teaching practices and the analysis of student thinking/understanding dimensions were thought to tap into respondents' pedagogical content knowledge (Shulman, 1986, 1987). Analyses of the mathematical content were understood either as measuring pedagogical content knowledge, if the mathematical content was analyzed in connection with the teaching and/or learning process, or as measuring mathematical content knowledge, if it was analyzed unrelated to those dimensions.

The fourth dimension, level of interpretation, indicated the overall quality of analysis contained in the response: that is, how analytic points were connected and how complete the analysis was. This dimension consisted of three categories. The first category reflected responses that were purely descriptions of the observed teaching episode. The second category represented responses that contained some analytic inference. Responses in this category focused either on the analysis of a single aspect in the observed teaching episode or on different aspects but without connecting analytic points to form a coherent argument. The third category described responses that showed a very comprehensive, coherent, and integrated interpretation, connecting analytic points through cause–effect relationships. This dimension most closely reflected findings from the expert–novice studies that suggest that expert teachers

perceive and interpret classroom events differently from novice teachers (Berliner, 1989; Bransford et al., 1999; Carter et al., 1987; Carter et al., 1988). Together, the four dimensions form the measurement model underlying the video-analysis assessment. (A figure illustrating this measurement model is available on request.)

Similar to testlets, in which several items are attached to a single item stem, in the video-analysis assessment four scores are assigned to a single teacher response to a video clip. Although the scores function much like items, they are not items in the traditional sense. That is why the term *score* is retained throughout the remainder of this article. The language used in the description of statistical analyses and results was adjusted accordingly (i.e., item means are referred to as score mean values). To avoid confusion, total scores are referred to as such.

Criterion measure. The Mathematical Knowledge for Teaching (MKT) instrument developed by Deborah Ball and colleagues at the University of Michigan (Ball, Hill, Rowan, & Schilling, 2002) was used as the criterion in this study. The instrument consists of a bank of multiple-choice items that measure teachers' mathematical content knowledge for teaching. Different from traditional content knowledge items, which assess whether a teacher can solve a given mathematical problem correctly, these items measure teachers' knowledge of standard and nonstandard algorithms or solution methods and knowledge of common student misconceptions and errors. Although the MKT instrument does not measure the exact same construct as the video-analysis assessment, it measures a closely related one.

Several forms have been created from the MKT item bank, for which good psychometric properties have been reported. Depending on form reliability (internal consistency) estimates between .72 and .79 have been reported (Ball et al., 2002). Furthermore, face, construct, and criterion-related validity has been documented for scores on some of the existing forms. A study in which this instrument was used to investigate the relationship between teacher knowledge and student learning found that teachers' mathematical knowledge for teaching was a statistically significant predictor of student achievement gains after key student- and teacher-level covariates were controlled for (Hill et al., 2005). MKT items were also mapped to National Council of Teachers of Mathematics (NCTM) and California content standards, indicating adequate coverage for central topic areas in the respective grade curriculum (Siedel & Hill, 2003).

For this study, 32 multiple-choice items were selected from the MKT item bank to form the criterion measure (Learning Mathematics for Teaching, 2004). The selected items met teaching standards for Grades 6 through 8 and matched the mathematical content of the video clips as closely as possible. Items predominantly assessed teacher knowledge in the area of number concepts, variables, and expressions. Internal consistency for scores on the form used in this study was estimated at .8, indicating that teachers' mathematical knowledge of teaching (i.e., the construct of interest) was measured with an adequate degree of consistency (Henson, 2001).

Expert ratings. Three expert ratings were obtained for a subset of teachers: (a) a rating of teachers' content knowledge, (b) a rating of teachers' pedagogical knowledge as it pertains to teaching mathematics, and (c) a rating of teachers' overall teaching experience. All three ratings used a 4-point scale: *limited knowledge/experience*, *proficient knowledge/experience*, *advanced knowledge/experience*, *expert knowledge/experience*. The ratings were provided by three professors at a university where the participating teachers were enrolled either in a teaching credential program for mathematics or a master's program in mathematics education. Each professor only rated teachers who were students in their respective classes and who they were well acquainted with. Professors based their ratings of teacher knowledge and expertise on in-class interactions, homework assignments, and classroom observations. Each teacher was rated by a single professor.

Professors' raw score ratings were used to compute standardized ratings (z scores) to control for calibration problems among the different rating professors. Both kinds of scores were used in the statistical analyses.

Background survey. A background questionnaire consisting of 16 questions was developed. Questions asked about teachers' professional and academic qualification and experiences, among them commonly used indicators of teacher quality. For example, undergraduate college degree, credential status, and number of years of experience are often reflected in teacher compensation models in school districts, whereas empirical evidence for their validity vis-à-vis student learning has been mixed (Fabiano, 1999; Rowan, Chiang, & Miller, 1997). The survey also asked teachers to describe their teaching practices. On a 6-point scale (ranging from *never* to *every day*), teachers indicated how often they or their students engage in a set of specific teaching or learning activities (Ball et al., 2002). For example, teachers reported how often they would ask their students to perform tasks that required methods or ideas not already introduced to them.

Analysis Plan

First, interrater agreement was estimated for the video-analysis data. Then the degree of correlated error among scores pertaining to the same video clip response (independence of scores) was investigated using multivariate generalizability theory. Next, classical item analyses were conducted and the reliability of the video-analysis scores (internal consistency) was estimated. In addition, exploratory IRT analyses were conducted for fitting one- and two-parameter models for ordered polytomous responses to the video-analysis data and the model fit was investigated. Finally, teacher scores on the video-analysis instrument, the criterion measure, and commonly used indicators of teacher quality were compared to examine the criterion-related validity of the video-analysis assessment.

Results

Interrater Reliability

Interrater reliability was calculated as percentage agreement and Cohen's kappa for each coding dimension to determine the degree to which the coding categories were applied consistently to teachers' written responses to the video clips. Interrater agreement was estimated on the basis of 50 randomly selected teacher responses from a total of 620 responses (62 teachers times 10 video clips; i.e., about 8% of the total number of responses). Those responses were coded along all four dimensions by two coders. Judgments were compared by dimension. Percentage agreement (calculated as the number of agreements over the number of agreements and disagreements) ranged from .79 to .85 across the four coding dimensions, suggesting that codes were applied with a high degree of consistency. Corresponding kappas, which adjust for chance agreement, were slightly lower and ranged from .65 to .78. Kappa values between .6 and .8 are often interpreted as indicating good rater agreement (Altman, 1991, p. 404).

Score Independence

Multivariate generalizability theory was used to investigate the degree of correlated error among scores that pertained to the same teacher response. Correlated error is a common occurrence when a single stimulus (e.g., a video clip or a reading passage) is used to generate responses to which multiple scores are assigned (Brennan, Gao, & Colton, 1995). The degree of correlated error might affect the psychometric properties of those scores, which has implications for the statistical analyses. For example, it provides information about the independence of scores. That is, whether scores may be kept separate in the item analyses or should be combined into a single score per clip. Independence of scores (i.e., local independence) is also a key assumption underlying all item-response theory models (Embretson & Reise, 2000). Moreover, correlated error may also artificially inflate internal consistency estimates, a commonly reported reliability coefficient that is based on inter-item correlations. For this study, it was preferable to keep scores within clips separate in the item analyses to learn more about the functioning of the different coding dimensions within and across the different video clips.

Similar to univariate generalizability theory, which provides a framework for specifying sources of error (referred to as facets) and for estimating their respective variance components, its multivariate extension allows for estimating covariance components (for a comprehensive treatment of generalizability theory, see Shavelson & Webb, 1991). Specifically, the estimated covariance components can be used to evaluate the degree of correlated error between different dimension scores pertaining to the same response. A one-facet model that specified the video clip as a

source (facet) of error was used (for the specified model and details of this analysis, see Kersting, 2005). Error correlations based on estimated covariance components were small, ranging from .10 to .25, and suggested that scores generated from the same response may be kept separate. Thus, the item analyses presented below are based on 40 scores (four scores each for 10 video clips).

Classical Item Analyses

Before conducting classical item analyses, scores from the level of interpretation dimension, which consisted of three categories, were recoded as 0, 0.5, and 1 to ensure that all scores carried equal weight in the subsequent analyses. This resulted in a common scale from 0 to 1 across all four dimensions.

Classical item statistics, including item difficulties and discrimination (corrected item total correlations), were computed for the video-analysis data. Score mean values covered a large portion (from $M_{s1} = .07$ to $M_{s40} = .89$) of the possible range (0–1). Similar to p values (indicating item difficulties), low mean values indicated categories that were rarely assigned and thus were considered difficult; high mean values indicated that categories had been frequently assigned and thus were considered easier. Score mean values increased at small and roughly equal increments along the knowledge continuum, with slightly more mean values in the upper half of the distribution, creating a fairly balanced scale that discriminated well among respondents with different knowledge levels.

Score mean values for scoring dimensions such as mathematical content and alternative solution methods across the different clips were comparatively small, indicating that few teachers included an analysis of the mathematical content or alternative solution methods in their analytic responses. This suggested that those dimensions might be associated with more advanced knowledge. In contrast, the student thinking/understanding dimension produced higher score means across the different clips, suggesting that teachers tended to include an analysis of student thinking or understanding more frequently in their responses. Finally, for the level of interpretation dimension, score mean values of about .5 were reported across the different video clips, indicating that for the majority of clips teachers did some interpretative work.

With one exception, all corrected item total correlations ranged from .22 to .60, with a median value of .37 and a mean value of .39. This suggested that scores discriminated well among respondents. Only the Alternative Teaching Strategy score for Clip 2 showed a lower item total correlation ($r = .13$). In combination with a small mean value ($M = .07$), this indicated that very few teachers with high overall scores included alternative teaching strategies in their response to this particular clip, which flagged this score for further investigation.

The total score distribution for this sample was fairly normal. Observed total scores covered a large portion (9–32.5 points) of the total possible score range (0–40 points),

with a mean value of 23.51 and a standard deviation of 7.94. No floor or ceiling effects were observed.

Internal Consistency

Reliability of scores was estimated for the entire scale (10 video clips) and by clip type (6 clips of THB and 4 clips of SMS) using Cronbach's alpha as a measure of internal consistency. All estimates were above the psychometric benchmark value of .8, which suggests that the construct of interest has been measured with a high degree of consistency (Henson, 2001). The internal consistency estimate for scores on the entire scale was .90; estimates for both clip types were .85. Confidence intervals calculated according to methods described by Fan and Thompson (2001) indicated that the internal consistency estimates remained above .8 (entire scale: .87–.93; THB and SMS: .81–.89). Similar estimates were produced in earlier pilot studies. Overall, the presented results suggest good technical quality of the video-analysis assessment.

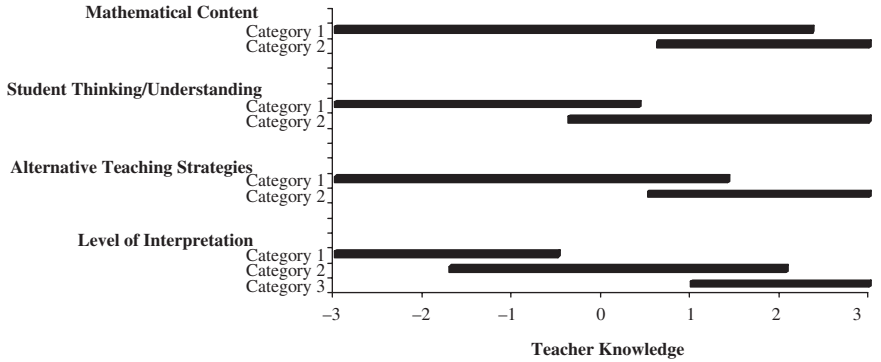
Item Response Theory Analyses

Item response theory (IRT) analyses were conducted to further investigate the psychometric properties of the video-analysis instrument. IRT analyses provide additional diagnostic information beyond what classical item statistics offer, which is particularly useful in the instrument development process. By using a single scale to represent person and item parameters, IRT scaling relates items or response categories directly to specific knowledge levels. That is, IRT estimates the probability to answer a given item correctly or a response to be classified in a specific category by providing the corresponding knowledge level. For instance, for the level of interpretation dimension, which consists of three categories, the model estimates two category transitions, one that indicates the knowledge level required to move from the lowest to the middle category and a second transition that indicates the knowledge level required to move from the middle category to the highest category. In contrast, for polytomous response data, the mean item score indicates the difficulty of the average category but does not distinguish categories from each other with respect to different knowledge levels.

However, IRT modeling requires much larger sample sizes than classical item analyses do to produce stable parameter estimates. Although additional teacher responses from a pilot study were included in the IRT calibration, increasing the sample size from 62 to 114, the IRT analyses are to be considered exploratory and the results are tentative.

Prior to fitting one- and two-parameter models for ordered polytomous responses to the video-analysis data, basic assumptions underlying IRT were checked. A check of the data indicated no serious violations of the unidimensionality and local independence

Figure 1
Relationship Between Teacher Knowledge of Teaching Mathematics and the Scoring Dimensions of the Video-Analysis Assessment



assumptions. The monotonicity assumption was checked by computing nonparametric item–response curves based on rest scores as described by Embretson and Reise (2000). In addition to the Alternative Teaching Strategy for Clip 2, which had been identified in the classical item analyses as potentially misfitting, five other scores, apparently at random, showed serious violations of the monotonicity assumption. Given the limited sample size, this might have been a result of a very low cell count for these categories for the particular clips. These individual scores were excluded from the subsequent statistical analyses. Thirty-four scores were retained.

Two polytomous IRT models, both of which assume ordered response categories, were fit to the video-analysis data: the partial credit model (PCM), which belongs to the family of Rasch models (Masters & Wright, 1996), and the less restrictive graded response model (GRM), which is an extension of the two-parameter binary model to the polytomous case (Samejima, 1969, 1996). A model comparison using the likelihood ratio test indicated that the GRM presented a better data–model fit (the difference in deviance was 1170.4010 on 34 degrees of freedom), and a subsequent residual analysis showed only very small residuals (<0.03) between observed and model predicted, suggesting a reasonably good data–model fit. Results of the IRT calibration of 34 scores are summarized graphically in Figure 1.

Figure 1 shows how the categories of the different scoring dimensions map onto the teacher knowledge continuum as estimated by the model. Teacher knowledge is represented through a logit scale ranging from -3 to $+3$, where -3 represents no or very limited knowledge, 0 represents average knowledge, and $+3$ represents expert knowledge. The horizontal bars represent the categories for each dimension. Their

position along the teacher knowledge continuum indicates the respective knowledge levels to which they correspond. Overlap of the horizontal category bars within a given dimension illustrates the estimated transitions between adjacent categories for the different clips, for example, the change from Category 0 (no analysis of mathematical content) to Category 1 (includes analysis of the mathematical content). The overlapping segments then show the dispersion of category transitions across the 10 clips. That is, the starting point of the bar of the higher category corresponds to the lowest estimated category transition among all clips; the end point of the bar of the lower category marks the transition for the clip with the highest transition estimate. The transitions for the remaining video clips fall in between. In summary, Figure 1 shows for each dimension the knowledge level that is associated with a specific response category.

Several findings are noteworthy in Figure 1. First, estimated locations for the three category locations for the level of interpretation dimension correspond well to the three broad knowledge levels: limited knowledge, average knowledge, and advanced knowledge. As shown in Figure 1, teachers in this sample who produced purely descriptive responses are characterized as having limited knowledge (Category 1); teachers who produced analytic responses that showed some interpretation corresponded to having average knowledge (Category 2); and teachers who produced very comprehensive analytic responses, including cause–effect relationships, were classified as holding advanced knowledge (Category 3).

Second, Figure 1 appears to confirm theoretical assumptions about the respective location estimates for the different dimensions. For example, concern for student thinking/understanding required at least average knowledge, providing some empirical evidence that novice teachers tend to focus primarily on the teacher and the teaching without considering the student explicitly part of that process. Furthermore, as hypothesized, analyses of the mathematical content and the inclusion of alternative teaching methods were limited to respondents with above-average knowledge.

Figure 1 also shows that when overlaying the category transition segments across dimensions, which are used to discriminate among respondents, the video-analysis assessment provides adequate discrimination coverage along the knowledge continuum. Category transition segments cover the knowledge continuum from about -1.7 to $+2.4$ on the 7-point logit scale. Similar to the results from the classical item analyses, a majority of category transition estimates fell in the upper half of the logarithmic scale, which suggests that the video-analysis assessment discriminates better between respondents with and above average knowledge. Follow-up studies need to explore the applicability of additional scoring dimensions or additional categories in existing dimensions to increase the number of category transitions on the lower half of the knowledge continuum and improve the assessment's discriminatory power for those knowledge levels.

Overall, both analytic approaches produced fairly comparable results despite the limited sample size used in the IRT calibration. Respondents' total scores and their

IRT-based trait-level estimates were highly correlated, $r = .97$. Although these results provide some indication of the robustness of the IRT analyses, a much larger sample would be required to produce stable item and level estimates.

Criterion-Related Validity

Teachers' total scores on the video-analysis instrument were compared to their total scores on the Mathematical Knowledge for Teaching criterion instrument, the expert ratings, and commonly used indicators of teacher quality to examine the criterion-related validity of the video-analysis assessment. A positive, statistically significant relationship of medium size ($r = .53$; $p < .01$) between the video-analysis total scores and total scores on the criterion measure was found. This indicates that the video-analysis assessment measures aspects of teacher mathematical content knowledge for teaching. In short, respondents with more mathematical content knowledge for teaching tended to provide more sophisticated interpretations of the video clips than teachers with lesser knowledge. The moderate size of the correlations underscores that the measured constructs of both instruments were similar but not identical. The criterion measure was designed to measure teachers' mathematical content knowledge for teaching, whereas the video-analysis assessment was designed to measure teachers' knowledge of teaching mathematics in concrete teaching situations, emphasizing the contextual and situational nature of teaching.

Among the expert ratings, only the rating of teachers' pedagogical knowledge was statistically significant and positively related to the video-analysis scores and the mathematical content knowledge for teaching scores, $r = .33$, $p < .05$ and $r = .30$, $p < .05$, respectively. Although smaller in size, the correlations provide cross-validation of the different measures and further evidence for the criterion-related validity of the video-analysis assessment.

An analysis investigating the relationship between the criterion measures and each of the video-analysis assessment's coding dimensions separately showed that the level of interpretation dimension alone ($r = .53$; $p < .01$) explained all of the shared variance between the video-analysis assessment based on all coding dimensions and the criterion measure of mathematical content knowledge for teaching ($r = .53$; $p < .01$). Although the mathematical content and student thinking/understanding dimensions also produced statistically significant correlations when entered in the model on their own ($r = .36$, $p < .05$ and $r = .37$, $p < .01$, respectively), their significant contributions disappeared once the level of interpretation dimension was added to the model. These results suggest that coding teachers' responses on the level of interpretation dimension alone would be sufficient to measure teachers' knowledge of teaching mathematics. This finding is noteworthy because it could significantly reduce the time and cost associated with the scoring of the video-analysis assessment, which affects the scalability of the measure.

Finally, no or very small correlations were observed between the scores of the video-analysis assessment and commonly used indicators of teacher expertise.

Discussion

This study set out to explore a novel, video-based assessment approach to measuring teacher knowledge of teaching mathematics by empirically investigating the reliability and validity of a prototype assessment. Overall, promising results were reported for the reliability and criterion-related validity of the video-analysis assessment developed for this study. Teachers' scores were sufficiently reliable (>.8) and close to estimates obtained in pilot studies, providing growing evidence for reliability. By meeting psychometric standards, which has been a challenge for assessments using constructed response formats, the video-analysis assessment appears to be a suitable instrument for measuring individual differences in teacher knowledge of teaching mathematics. A next step in the instrument development process would be a multiple-time-point administration to explore whether the instrument is sufficiently sensitive to measure individual change in teacher knowledge over time.

Furthermore, teachers' scores on the video-analysis assessment were statistically significantly and positively related to a criterion measure of mathematical content knowledge for teaching, which has been linked to student learning ($r = .53$), and to expert ratings of teachers' pedagogical knowledge for teaching mathematics ($r = .33$). This suggests that teachers used their pedagogical and mathematical content knowledge for teaching when analyzing classroom instruction, thereby providing initial evidence for the criterion-related validity of the video-analysis assessment. The strength of the association supports the preliminary conclusion that the video-analysis assessment measures a relevant construct within the teacher knowledge domain and has potential to be related to teaching and student learning.

Interestingly, the level of interpretation dimension appeared to function as an umbrella dimension. That is, this dimension alone produced the same amount of shared variance with the criterion measure as all four dimensions combined. Once the level of interpretation dimension was added to the regression model, the remaining three dimensions became nonsignificant. One possible interpretation of this result might be that this dimension described the sophistication of teachers' analyses and whether interpretative comments were connected, which incorporates to some degree information described by the other three dimensions (mathematical content, student thinking/understanding, and alternative teaching strategies).

From an assessment point of view, this is an important finding because it suggests that coding teachers' responses on this single dimension produces valid scores that discriminate well among respondents' knowledge of teaching mathematics. This could potentially greatly reduce the time and cost associated with the scoring

process. Moreover, it might simplify the application of computerized text analysis for the scoring of teachers' video-analysis assessment responses, further improving scoring efficiency and thus increasing the assessment's scalability.

The validity investigation indicated no statistically significant relationship between commonly used indicators of teacher expertise (e.g., number of years of experience, certification status, and college degree) and either scores on the video-analysis assessment or the criterion measure of mathematical content knowledge for teaching. This may not be surprising because existing research has shown mixed results. However, the current finding is noteworthy because the No Child Left Behind (NCLB, 2001) legislation defines teacher quality by some of those same indicators. Specifically, the NCLB defined quality teachers as teachers with full certification, a bachelor's degree, and demonstrated competence in the subject matter and teaching (NCLB, 2001). Assuming that high-quality teachers have greater knowledge and students who learn more than do teachers who do not meet those criteria, it appears important for this instrument development effort and educational policy to further investigate the relationship between those indicators, teacher knowledge, and student learning.

Clearly, the results presented here are promising yet preliminary. There is still much to be done. Most importantly, studies with larger and more representative samples are needed that extend the validity investigation to also include measures of teaching and student learning. In some ways, the use of video analysis as a measure of knowledge for teaching has more face validity than paper-and-pencil assessments. Regardless how much knowledge teachers have, if it is going to improve their teaching and their students' learning, it has to be accessed and used in the classroom. Analyzing classroom events might be seen as one of the basic skills of teaching. This new assessment approach might be a more direct way of measuring the critical knowledge teachers bring to their craft.

References

Altman, D. G. (1991). *Practical statistics for medical research*. London: Chapman & Hall.

Ball, D. L. (2000). Bridging practices: Intertwining content and pedagogy in teaching and learning to teach. *Journal of Teacher Education*, *51*, 241-247.

Ball, D. L., & Bass, H. (2000). Interweaving content and pedagogy in teaching and learning to teach: Knowing and using mathematics. In J. Boaler (Ed.), *Multiple perspectives on the teaching and learning of mathematics* (pp. 83-104). Westport, CT: Ablex.

Ball, D. L., Hill, H. C., Rowan, B., & Schilling, S. (2002). *Measuring teachers' content knowledge for teaching: Elementary mathematics release items.* Ann Arbor, MI: Study of Instructional Improvement.

Berliner, D. C. (1989). Implications of studies of expertise in pedagogy for teacher education and evaluation. In *New directions for teacher assessment: Proceedings of the 1988 ETC invitational conference*. Princeton, NJ: Educational Testing Service.

Bond, L. (1998). Disparate impact and teacher qualification. *Journal of Personnel Evaluation in Education*, *12*, 211-220.

Bransford, J. D., Brown, A., & Cocking, R. (Eds.). (1999). *How people learn. Brain, mind, experience, and school* (pp. 29-50). Washington, DC: National Academy Press.

Brennan, R. L., Gao, X., & Colton, D. A. (1995). Generalizability analyses of work keys listening and writing tests. *Educational and Psychological Measurement*, *55*, 157-176.

Carter, K., Cushing, K., Sabers, D., Stein, P., & Berliner, D. (1988). Expert-novice differences in perceiving and processing visual classroom stimuli. *Journal of Teacher Education*, *39*, 25-31.

Carter, K., Sabers, D., Cushing, K., Pinnegar, S., & Berliner, D. (1987). Processing and using information about students: A study of expert, novice, and postulant teachers. *Teaching and Teacher Education*, *3*, 147-157.

Cochran-Smith, M., & Lytle, S. (1999). Relationships of knowledge and practice: Teacher learning in communities. *Review of Research in Education*, *24*, 249-305.

Darling-Hammond, L., & Baratz-Snowden, J. (Eds.). (2005). *A good teacher in every classroom. Preparing the highly qualified teachers our children deserve*. San Francisco: Jossey-Bass.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.

Fabiano, L. (1999). *Measuring teacher qualifications* (NCES-WP-1999-04). Washington, DC: U.S. Department of Education, National Center for Educational Statistics.

Fan, X., & Thompson, B. (2001). Confidence intervals about score reliability coefficients, please: An EPM guidelines editorial. *Educational and Psychological Measurement*, *61*, 517-531.

Haertel, E. H. (1991). New forms of teacher assessment. *Review of Research in Education*, *17*, 3-27.

Henson, R. K. (2001). Understanding internal consistency reliability estimates: A conceptual primer on coefficient alpha. *Measurement and Evaluation in Counseling and Development*, *34*, 177-189.

Hiebert, J., Gallimore, R., Garnier, H., Givvin, K. B., Hollingsworth, H., Jacobs, J., et al. (2003). *Teaching mathematics in seven countries: Results from the TIMSS 1999 Video Study* (NCES 2003-013). Washington, DC: U.S. Department of Education, National Center for Educational Statistics.

Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal*, *42*, 371-406.

Hill, H., Schilling, S., & Ball, D. (2004). Developing measures of teachers' mathematics knowledge for teaching. *Elementary School Journal*, *105*, 11-30.

Jaeger, R. M. (1998). Evaluating the psychometric qualities of the National Board for Professional Teaching Standards' assessments: A methodological accounting. *Journal of Personnel Evaluation in Education*, *12*, 189-210.

Kersting, N. (2005). *Measuring teachers' knowledge of teaching mathematics: Instrument development and validation*. Unpublished doctoral dissertation, University of California, Los Angeles.

Klein, S. P. (1998). Standards for teacher tests. *Journal of Personnel Evaluation in Education*, *12*, 123-139.

Learning Mathematics for Teaching. (2004). Mathematical knowledge for teaching measures: Number concepts and operations [Instrument]. Ann Arbor, MI: Author.

Masters, G. N., & Wright, B. D. (1996). The partial credit model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory*. New York: Springer.

No Child Left Behind of 2001. (2001, January). Public Law No. 107–110, 107th Congress. Retrieved from http://www.ed.gov/legislation/ESEA02/.

Pool, J. E., Ellett, C. D., Schiavone, S., & Carey-Lewis, C. (2001). How valid are the National Board of Professional Teaching Standards assessments for predicting the quality of actual classroom teaching and learning? Results of six mini case studies. *Journal of Personnel Evaluation in Education*, *15*, 31-48.

Porter, A. C., Youngs, P., & Odden, A. (2001). Advances in teacher assessments and their uses. In V. Richardson (Ed.), *Handbook of research on teaching* (4th ed., pp. 259-297). Washington, DC: American Educational Research Association.

Rowan, B., Chiang, F., & Miller, R. J. (1997). Using research on employees' performance to study the effects of teachers on student achievement. *Sociology of Education*, *70*, 256-284.

Sabers, D. S., Cushing, K. S., & Berliner, D. C. (1991). Differences among teachers in a task characterized by simultaneity, multidimensionality, and immediacy. *American Educational Research Journal*, *28*, 63-88.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Pyschometrika Monograph*, No. 17.

Samejima, F. (1996). The graded response model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory*. New York: Springer.

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

Shulman, L. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher*, *15*, 4-14.

Shulman, L. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review*, *57*, 1-22.

Siedel, H., & Hill, H. (2003). *Content validity: Mapping SII/LMT mathematics items onto NCTM and California Standards.* Unpublished manuscript.

Stone, J. E. (2002). *The value-added achievement gains of NBPTS-certified teachers in Tennessee: A brief report*. Retrieved October 24, 2004, from www.education-consumers.com/briefs/stoneNBPTS.shtm