J Math Teacher Educ (2011) 14:305–325
DOI 10.1007/s10857-011-9181-0

Prediction assessments: Using video-based predictions to assess prospective teachers' knowledge of students' mathematical thinking

Anderson Norton · Andrea McCloskey · Rick A. Hudson

Published online: 12 March 2011

Abstract In order to evaluate the effectiveness of an experimental elementary mathematics field experience course, we have designed a new assessment instrument. These video-based prediction assessments engage prospective teachers in a video analysis of a child solving mathematical tasks. The prospective teachers build a model of that child's mathematics and then use that model to predict how the child will respond to a subsequent task. In this paper, we share data concerning the evolution and effectiveness of the instrument. Results from implementation indicate moderate to high degrees of inter-rater reliability in using the rubric to assess prospective teachers' models and predictions. They also indicate strong correlation between participation in the experimental course and prospective teachers' performances on the video-based prediction assessments. Such findings suggest that prediction assessments effectively evaluate the pedagogical content knowledge that we are seeking to foster among the prospective teachers.

Keywords Teacher knowledge · Instrument development · Video · Preservice teacher education · Pedagogical content knowledge

Introduction

Since Shulman's (1986) seminal work on the topic, pedagogical content knowledge (PCK) has become a central focus of mathematics teacher education. Researchers have addressed PCK by designing various approaches to supporting teachers' professional development.

A. Norton
Department of Mathematics (0123), Virginia Tech, Blacksburg, VA 24061, USA
e-mail: norton3@vt.edu

A. McCloskey (✉)
Penn State, 268 Chambers Building, University Park, PA 16802, USA
e-mail: amccloskey@psu.edu

R. A. Hudson
University of Southern Indiana, 8600 University Blvd, SC 3261, Evansville, IN 47712, USA
e-mail: Rhudson@usi.edu

These approaches include engaging teachers in examining mathematical tasks (e.g., Arbaugh and Brown 2005), examining students' written work (e.g., Carpenter et al. 1988); analyzing video (e.g., Star and Strickland 2008; Stockero 2008), reflecting on whole-class instruction (e.g., Chamberlin 2005), and conducting clinical interviews (e.g., Crespo and Nicol 2003; Wright 2000). Yet, several questions about teachers' development of PCK remain open to further research; among them, how can teacher educators reliably assess growth in teachers' PCK? Our paper addresses this question, specifically as it applies to prospective elementary school teachers' understandings of students' mathematical thinking.

Shulman (1986) identified understanding students' mathematical thinking as an important aspect within his initial description of PCK:

> Pedagogical content knowledge also includes an understanding of what makes the learning of specific topics easy or difficult: the conceptions and preconceptions that students of different ages and backgrounds bring with them to the learning of those most frequently taught topics and lessons (p. 9).

Since then, several researchers have demonstrated that a focus on students' mathematical thinking supports positive teacher change (Fennema et al. 1996; Norton and McCloskey 2008; Steinberg et al. 2004). In fact, each of their projects found that when teachers investigate individual students' thinking outside of the classroom, they tend to ask more probing questions and spend more time listening to student explanations inside the classroom.

The project in which we developed our instrument—the Iterative Model Building (IMB) project—engaged prospective elementary school teachers (PSTs) in building models of students' mathematics during an early field experience. *Model building* refers to teachers' ongoing and iterative activities of providing reasonable explanations for students' actions (including verbalizations) as the children solve mathematical tasks. von Glasersfeld and Steffe (1991) defined models as follows: "A model 'simulates reality'; it is a conceptual construct that is treated as though it gave an accurate picture of the real world" (p. 95). Models of students' mathematics, then, simulate the mental actions students perform as they solve mathematical problems, and the power of a model is determined by its potential for explaining observed events and predicting future ones (von Glasersfeld and Steffe 1991). In the case of modeling students' mathematics, this might mean predicting students' future problem-solving behavior. Here, we introduce a new video-based assessment instrument based on that premise.

The purpose of this paper is to establish the utility, reliability, and validity of the video-based prediction assessment instrument as a measure of PSTs' abilities to model students' mathematical thinking. In the next section, we discuss the theoretical underpinnings of the instrument within related literature on students' mathematical thinking, video analysis, and teacher knowledge. We then situate the design and utility of the instrument within the IMB project. Finally, we share findings that establish the reliability and validity of the instrument.

Theoretical underpinnings

Focusing on students' mathematical thinking

Professional development programs that focus on students' mathematical thinking have produced results that consistently indicate the value of the approach for both students and

teachers. In fact, the phrase "learning from students" generally refers to such professional development programs (Crespo 2000; Norton and McCloskey 2008). Focusing on students' mathematical thinking also goes hand-in-hand with research on "listening to students" (Confrey 1994; Davis 1997; Schifter 1998; Wallach and Even 2005). Such dispositions—focusing on students' mathematical thinking and listening to students—form the basis for building models of students' mathematics (Confrey 1993).

Cognitively Guided Instruction (CGI) (Carpenter et al. 1988) engaged elementary school (inservice) teachers in a program that "focused directly on enabling teachers to understand their students' thinking" (Fennema et al. 1996, p. 405). The project remains the most influential professional development study of its kind, which assumes that "knowledge of children and their mathematics is crucial to teaching for understanding" (Ball 1994, p. 21). The researchers supported teachers' efforts to focus on students' thinking by providing cognitive models of student strategies along with video clips to illustrate students' actions in applying those strategies. They found that teachers learned from students in the sense that students' invented strategies and explanations informed teachers' instructional use of problem-solving activities. Instructional changes, in turn, opened up more opportunities for student strategies and explanations. The project also attempted to correlate teachers' predictions about students' problem-solving strategies with measures of student achievement.

To examine potential correlations, CGI used a video-based prediction instrument similar to the one we introduce here. The teachers were shown videos of their own students solving word problems and were asked to predict student success and strategies. The researchers then measured the correlation between teachers' success in predicting students' actual responses and student achievement—as indicated by student performance on tests of number facts and problem-solving ability. The researchers found that teachers' predictions of student success were significantly correlated with student achievement, but that "ability to predict the strategies that students would use was not" (Carpenter et al. 1988, p. 399). The researchers attributed the mixed findings to the observation that teachers traditionally make instructional decisions based on expectations of student success but not based on student strategies.

Several studies affirm the findings of CGI, while contributing to increasingly nuanced understandings of professional development programs that focus on students' mathematical thinking. For example, Vacc and Bright (1999) conducted a CGI-based professional development study with prospective elementary school teachers and found that focusing on students' mathematical thinking had a significant impact on the PSTs' beliefs about teaching. However, "their use of knowledge of children's mathematical thinking during instructional planning and teaching was limited" (p. 89). Still, other studies demonstrate that continued focus on students' mathematical thinking can affect instructional change over time (Arbaugh and Brown 2005; Chamberlin 2005; Steinberg et al. 2004; Wallach and Even 2005). Moreover, a follow-up study revealed that CGI teachers whose practice had changed by focusing on students' mathematical thinking continued to learn from their students in noticeable ways (Franke et al. 2001).

Confrey (1988) claimed that professional development programs must support not only "investigations of students' mathematical thinking"—one of two avenues to professional development identified by Schifter (1998, p. 55)—but also means of reproducing children's thinking. "Teaching constructively involves building models of student thinking" (Confrey 1993, p. 307). These models become the basis for constructivist teaching because they inform teachers of what students bring to bear on mathematical situations, which, in turn, inform instructional decisions such as the selection and design of mathematical tasks (Steffe 1991; Thompson et al. 2007). Clinical interviews (diSessa 2007; Piaget 1970) and

teaching experiments (Confrey and Lachance 2000; Steffe and Thompson 2000) provide opportunities for teachers and researchers to build models through intensive, iterative interactions with small groups of students (usually one or two students).

Video analysis in teacher education

Video analysis can play a critical role in supporting teacher learning through model building because it provides opportunity for teachers to review and reflect on the mental activity underlying students' actions. Video clips provide an ideal medium for assessing teacher knowledge because they are both replicable and authentic, in that they depict real students and teachers engaged in doing mathematics. These same characteristics make video a valuable instructional tool, because discussions about teaching can be framed around a single data source. Moreover, the same clip can be used for multiple purposes: as a context for assessing what prospective teachers know and as a context for group discussion and learning.

The use of videos in teacher education is not a new practice. For example, Annenberg Media's (1997) online library of whole-class mathematics lessons and Kathy Richardson's (1990) series of videotaped mathematics assessments with individual children have been used in American elementary mathematics methods courses to illustrate teaching and assessment practices consistent with inquiry-based teaching and to highlight children's thinking about mathematical topics. Mathematics educators continue to design innovative uses for video within prospective and inservice teacher education. The recent studies summarized below provided us with evidence that video can be used as an effective tool for supporting and assessing prospective teacher learning.

Van Es and Sherin (2002) designed a software tool, VAST (Video Analysis Support Tool), to promote teacher learning by supporting teachers as they analyze video from their own classrooms. The VAST scaffolds teachers as they notice and analyze those features of classroom interactions that the authors identify as being most important to "reform pedagogy." Through their work with VAST, the professional development providers taught a group of secondary mathematics and science prospective teachers to use the software tool. They found that over the course of a semester, the prospective teachers moved away from descriptive and evaluative noticings toward more analytic and interpretive noticings.

In the Integrating Mathematics and Pedagogy (IMAP) project, as reported by Philipp et al. (2007), some prospective elementary teachers were randomly assigned to a video-based field experience course designed to supplement a mathematics course. These prospective teachers watched, analyzed, and discussed videos that highlighted children's mathematical thinking. Members of the research team conducted some of the videotaped interviews, and on other video clips, the children were interviewed by fellow prospective teachers. After measuring changes in the prospective teachers' mathematical content knowledge and beliefs using instruments developed by the researchers, the authors concluded that the video-based field experiences had been effective in supporting the prospective teachers' willingness and ability to analyze children's mathematical thinking.

Stockero (2008) reported on the effectiveness of using a video-case curriculum in a middle-school mathematics methods course. The PSTs in this course viewed, analyzed, and discussed video clips of students solving mathematical tasks. Building on the framework of attributes of reflection developed by van Es and Sherin (2008), Stockero collected and analyzed data about the nature of the group discussions and reflections written by the prospective teachers during their accompanying field experience. After gaining experience in analyzing the video clips and their own teaching, the PSTs in Stockero's study improved the sophistication of their reflections.

Other projects have considered how video can be useful for program evaluation purposes. Star and Strickland (2008) report on an effort to document what prospective secondary mathematics teachers notice when they observe a classroom. This study began with the premise that "PSTs' ability to learn from observations of teaching (either live or in the video) is critically dependent on what is actually noticed (attended to)" (p. 121). It follows that if teacher preparation programs are going to continue the long-standing tradition of requiring prospective teachers to observe classrooms, then we should make an effort to document what it is that PSTs notice about classroom interactions, and whether those noticings can be improved through coursework. As a way to answer these questions, Star and Strickland designed instruments that make use of mathematics classroom clips collected in the TIMSS video study.

Star and Strickland (2008) administered the video-based assessments to prospective teachers both before and after they had completed a secondary mathematics methods courses and found that the prospective teachers were initially most likely to notice issues of classroom management and least attentive to the actual mathematical content of the lessons. However, by the end of the methods courses, the prospective teachers had improved in their ability to notice more substantive features of the mathematics lessons.

Assessing teacher knowledge

Assessing teacher knowledge has gained increasing levels of attention from politicians and teacher educators alike. Instruments are designed with the underlying assumption that the quality of instruction that a teacher is capable of providing is at least partially attributable to his or her knowledge. The connection between knowledge and teaching practice is not a simple one, and we have much work to do in understanding its complexity. In the meantime, licensing bodies want to ensure that teachers are qualified, and program providers want to evaluate the effectiveness of preparation or professional development efforts. In the United States, legislative pressures to identify teachers as "highly qualified" (e.g., No Child Left Behind 2001) have increased attempts to assess teacher quality using teacher knowledge as a proxy. Most states (currently 46) require a minimum score on at least one portion of the Praxis Series, an assessment series administered by the Educational Testing Service, for initial licensing (ETS 2006). For many elementary teacher candidates in the United States, this requires taking the Praxis II examination on Curriculum, Instruction, and Assessment.

The Praxis II test includes questions that could be considered indicators of mathematics pedagogical content knowledge, such as those asking for analyses of student work. However, this conceptualization of mathematics PCK is limited by the test's multiple-choice format and the fact that only twenty percent of the questions focus on mathematics. Hill et al. (2008) have reported partial success in measuring PCK using carefully designed multiple-choice items, but they also caution that measuring PCK is complicated and that multidimensionality of this domain of knowledge requires the use of similarly complex measurement techniques.

In addition to large-scale, policy-driven efforts, and constrained, multiple-choice formats, there are other research projects that seek to assess a more nuanced conception of PCK. Koirala et al. (2008) shared the performance assessment task that they "designed to assess secondary school mathematics prospective teachers' pedagogical content knowledge and skills" (p. 127). This task was not centered on video recordings but rather on students' written work, gathered from released items from their state's tenth-grade test. Prospective teachers taking the test were asked to analyze the student work, provide meaningful

feedback, and then use that information to design lessons. This assessment operationalizes PCK by including the interpretation of student work and the planning of appropriate lessons.

Kersting (2008) reported on a similar "novel approach to measuring knowledge of teaching mathematics" (p. 847). This project used video clips that served as item prompts, but unlike our project, Kersting's approach used clips of classroom instruction episodes. Analysis of early results using several techniques (e.g., using item-response theory and determining criterion-related validity by comparing the scores to other measures) demonstrated that this instrument does in fact measure an important dimension of knowledge useful for teaching mathematics.

The current study focuses on assessing mathematical PCK, but it is embedded in an ongoing, longitudinal study—the IMB project—aimed at deepening our collective understanding about how to prepare mathematics and science elementary teachers and to identify early indicators of PSTs who will eventually develop into high-quality teachers. Indeed, our prediction assessments were developed as part of the larger project, which seeks to add to the knowledge base in both of these domains: teacher preparation and teacher assessment. In order for the reader to develop a sense of the role prediction assessments play in the larger research project, we provide a brief description of that project.

Methods

The Iterative Model Building project

Innovative approaches to teacher preparation have received increased attention and research funding (Cochran-Smith and Zeichner 2005; Education Commission of the States 2003a, b; Kane et al. 2006). Iterative Model Building (IMB) is a NSF-funded project that restructures field experiences to support elementary mathematics and science education methods courses and investigates the effect of this restructuring. The experimental design of the IMB project allows for comparisons between two types of field experiences: the "traditional" field experience, in which PSTs teach mathematics and science lessons to small groups of children and reflect on their teaching individually; and the "experimental" field experience, in which PSTs conduct a variation of a teaching experiment (which we refer to as "formative assessment interviews") with pairs of children and then reflect on whole-class teaching through lesson study.

An increasing number of PSTs will participate in the IMB field experience as the project's implementation is scaled up. Data of various kinds will be collected, including beliefs surveys, written reflections, lesson plans, lesson observations, the prediction assessments discussed here, and complementary prediction assessments for science. Furthermore, these data will be collected at multiple points of the teachers' careers: while they are enrolled in their preservice mathematics and science methods and field experience courses, while they are student teaching, and during their first year as professional teachers. Prediction assessments are administered during the mathematics methods course, and we hypothesize that success on the prediction assessments may serve as an early indicator of the quality of the mathematics instruction provided by the teacher in subsequent years. The ongoing study will attempt to document whether results of PSTs' prediction assessments during the field experience semester correlate with later and more costly measures of

teacher quality, such as lesson observations during the teachers' student teaching semester and first year of teaching.

Formative assessment interviews and lesson study

In the experimental approach, PSTs engage in a series of interviews with elementary students over a 6-week period. These interviews, which we refer to as formative assessment interviews (FAIs), typically last 20–30 min and are conducted at an elementary school during the PSTs' weekly field experience course that accompanies their mathematics methods courses. Like clinical interviews, FAIs involve intensive interactions with a pair of children in order to investigate their ways of thinking about a particular topic. Beyond clinical interviews, we expect PSTs to use successive interviews to refine their models of students' thinking. On the other hand, the FAIs do not reach the level of recursion that Steffe and Thompson (2000) have ascribed to teaching experiments, but do provide the prolonged engagement with a pair of students that is common to teaching experiment methodology. In other words, teaching experiments informed our design of FAIs, but we do not expect PSTs to engage in the high level of recursive hypothesis testing involved in genuine teaching experiments.

During the FAIs, a PST poses tasks to two students in an attempt to elicit their thinking. A second PST observes the interactions and takes observation notes about the students' responses. After the interview, the pair of PSTs meets to discuss the children's responses, changes in the children's thinking, and the implications for whole-class lessons. During the following week, one pair of PSTs describes their model of student thinking and shares a video clip of the prior week's FAI with four other PSTs. The group collectively reflects on the model, discusses its efficacy, and when necessary, modifies the proposed model during this group model building session.

The second innovation of the IMB approach, lesson study, has been used as a tool for reflection about lesson development and teaching (Lewis 2000). During each week of the 6-week mathematics field experience, and based on findings of prior FAIs and model building sessions, a pair of PSTs prepares and co-teaches a lesson for an entire class of elementary students. Following the lesson, the pair of PSTs meets with other observers to discuss student thinking during the lesson and to revise the lesson using a lesson study protocol. This lesson study team includes the two PSTs who co-taught the lesson, four other PSTs, the host classroom teacher, and a university supervisor. Based on feedback provided in the lesson study meeting, the pair of PSTs revises their initial lesson plan to incorporate the suggestions made by their peers.

Initial design of the videos and rubric

We designed prediction assessments to capture one aspect of PCK: teachers' ability to understand, model, and predict student thinking. In our study, two administrations of prediction assessments were held during the PSTs' mathematics methods course. Course instructors administered one prediction assessment during the first week of the semester and a different prediction assessment at the end of the semester. Thus, two prediction assessment instruments with two different videos and tasks were created. All PSTs in the mathematics methods course were expected to complete the prediction assessments, whether they were enrolled in the "traditional" field experience or the "experimental" field experience.

During the prediction assessment, the PSTs watch a video clip of an elementary student working on a mathematical task that requires the student to find and justify his or her answer. On the video, an interviewer (first author) poses tasks and follow-up questions to probe the student's thinking, using manipulatives as a medium. However, before the clip begins, the PSTs are asked to solve a similar mathematical task. PSTs record their solutions on one side of the prediction assessment response sheet and then turn it over to the backside on which they record their models and predictions. Once they have recorded their solutions and turned over their sheets, they are shown the video clip.

After PSTs watch the student solve tasks for about 8 min, the video is paused, and the PSTs write an individual reflective response about how they believe the student was reasoning. This is their model of the student's thinking. The PSTs are given about 8 min to record their models and then they watch another short video clip in which a new, related task is posed to the same student—the same task that the PSTs had solved. The video is paused a second time, and the PSTs are asked to write how they predict the student will respond. The PSTs then turn in their prediction assessment forms and view the student's actual response, which often becomes fodder for classroom discussion. Following the administration of the prediction assessment, identifying information is removed from the response forms.

Initially, a rubric was developed to measure two components: *model* and *prediction*. The *model* component measures PSTs' ability to model the student's thinking and the PSTs' use of evidence to construct a model. The *prediction* component measures two dimensions of the PSTs' response: accuracy and detail. Thus, we are interested in whether the PST can arrive at an accurate prediction of how the student will respond, with enough detail to envision the student's particular actions.

Refining the instrument

We conducted two pilot studies of the instruments, during the spring 2008 and fall 2008 semesters. As a result of these studies, we made a number of changes in an attempt to refine the video instruments and rubric used to assess PST responses. After each administration of the prediction assessments, the first and third authors independently scored the responses. They then reconciled differences in scores. As a result of the reconciliation during the pilot studies, wording of the rubric was refined and new components were formed. For example, the initial *model* component only measured the quality of the PSTs' models, not whether the PST had explicitly used the model to form their prediction. So, we included a new component on the rubric to measure the PSTs' *use of model*.

In another analysis of the PSTs' responses, we found evidence that some PSTs did not seem to know how to correctly complete the mathematical task in the video, suggesting a weakness in their own content knowledge. We hypothesized that a lack of related content knowledge might limit PSTs' abilities to construct models and predict student thinking. Thus, we incorporated an item to measure the PSTs' content knowledge of the subject by having them complete the mathematical task before they begin watching the video. We adapted the instrument and rubric to include this *content knowledge* component. We share the final version of the rubric in "Appendix".

Participants

A total of 45 subjects participated in the study, but due to missing data, three participants' data were not included in the final analyses. All subjects were enrolled in a teacher

education program at a large university in the United States, and all data presented here were collected during the spring semester of 2009. Most of the subjects were white, female, and traditionally–aged college students (ages 18–24). All of the subjects were majoring in elementary or special education and had completed coursework in at least three mathematics content courses. At the time of the study, they were concurrently enrolled in a block of classes that included a mathematics methods course, a science methods course, and a field experience course focusing on mathematics and science content.

There were two approaches used for the field experience course: the Iterative Model Building (IMB), or "experimental," approach and a traditional approach. Approximately half of the subjects participated in each approach. The traditional field experience approach is similar in some aspects to the IMB approach. For example, PSTs in both approaches were in the field experience for 2.75 h per week for a 6-week mathematics experience, followed by a 6-week science experience. The PSTs in the traditional approach interviewed students one time and presented weekly lessons to small groups of children. However, these PSTs did not engage in ongoing FAIs, nor did they collectively reflect in a lesson study meeting. The lessons in the traditional approach were created for small groups of elementary students, and the PSTs wrote individual reflections about their lesson. Prior to the start of the semester, the PSTs individually registered for one of the two sections of the field experience course, without knowing which course would be associated with the IMB approach or the traditional approach.

Context of the videos

The pre-assessment

The pre-assessment video featured a fifth-grade student named "Maddie" who worked with a partner on a computer program called *TIMA: Sticks* (Olive 2000). The program creates a virtual environment that serves as a medium for students to carry out actions on virtual fractions sticks. For example, students can draw fractions sticks, partition them into a specified number of equal pieces, copy those pieces, and measure various sticks relative to a specified "unit stick" (the designated whole stick). Throughout the video, the interviewer posed tasks for Maddie to solve within the computer environment. The PSTs watch about 8 min of the video from which to build their models of Maddie's thinking.

At one point, Maddie had constructed a 7/7-stick but called it "one-seventh," so the interviewer decided to ask Maddie and her partner to produce fractions sticks, 1/7, 2/7, 3/7, and so on, as illustrated in Fig. 1.

Maddie and her partner continued producing and measuring sticks until they produced 7/7, before they measured the stick, Maddie exclaimed, "it's going to say a whole!" Once

Fig. 1 Maddie's work in TIMA: Sticks

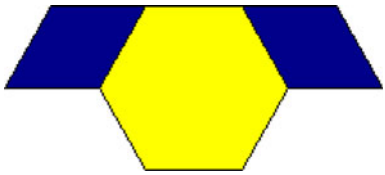

Fig. 2 The doggy problem

Maddie measured the stick, the interviewer asked why she had called it 1/7 when it did not look like the 1/7 stick (illustrated in Fig. 1): "Do you think they should both have that same name, one-seventh?" Maddie responded by saying, "actually, no I don't because this little piece [the left-most seventh of the 7/7-stick] is one-seventh, so I think it's [the 7/7-stick] one whole."

Here, the video was paused so that PSTs could record their models of Maddie's thinking. The video was resumed with Maddie wondering aloud about what stick would come after 7/7. The interviewer encouraged Maddie to make the next stick [an 8/7 stick] and asked her what measure the stick will have. The video was paused again while PSTs recorded their predictions: "What will Maddie call the stick she just made?" When the video resumed, Maddie conjectured that the stick would measure 8/8.

The post-assessment

The post-assessment video featured a fourth-grade student named "Lane" who was solving fractions tasks with pattern blocks. In the first video clip of the assessment, the interviewer asked Lane to consider a figure of a hexagon and a trapezoid as 9/7. The interviewer asked Lane to determine the whole. After considerable work and several proposed answers that he later deemed incorrect, Lane successfully found the answer of one hexagon and one triangle. Later, the interviewer showed Lane a second figure consisting of a hexagon and two rhombi (illustrated in Fig. 2) and asked Lane to produce a representation of the whole, given that the figure shown represented 10/6 of the whole. Lane referred to the figure as a "doggy." He successfully solved the task by placing 10 triangles on top of the figure—two on each rhombus and six on the hexagon. He continued by counting the triangles and then removing four of them, saying, "ten minus four equals six."

After watching Lane complete such tasks, the PSTs were prompted to build a model of Lane's reasoning and articulate it in writing. Next, the video was resumed, and the PSTs watched the interviewer pose a follow-up task to Lane, asking him to use the same figure to produce the whole, if the "doggy" now represented 10/7 of the whole. The PSTs were then asked to use their model to predict how Lane would solve this task and to articulate their predictions in writing. Once the PSTs completed their predictions, the video resumed, showing how Lane resolved the task. In particular, Lane laid ten triangles on the figure again and, after counting, removed three of them. Once the interviewer probed, Lane explained "ten minus seven is three."

Sample responses from two PSTs

Here, we include sample responses from two of the PSTs, "Bree" and "Tabitha," on the post-assessment. Both raters scored Bree's response at the highest assessment levels for each of the four components: *content knowledge*, *modeling*, *use of model*, and *prediction*. The raters reconciled to score Tabitha's response a 0 on all four components. These

contrasting responses are provided so that the reader might understand how the raters used the rubric to score PST responses.

For the content knowledge component, Bree drew a hexagon divided into six equilateral triangles, with an additional equilateral triangle appended—a correct response indicating some degree of content knowledge related to the task Lane was posed. For the model component, Bree responded as follows:

> Lane is thinking about improper fractions by seeing how far away they are from one whole. He uses pattern blocks to divide up the blocks into equal sized pieces. For example, he places 6 green triangles inside of/on top of the yellow hexagon to show that 1 hexagon equals a whole and 6 triangles equal a whole. He then sees how many more triangles there are in the blocks that exceed the whole.

This response demonstrates that Bree used *evidence* to explain *how* Lane was reasoning, which satisfies the two criteria for a top score of 2 on the model component. Bree explained that Lane reasoned by determining how many extra pieces, beyond the whole, were in the improper fraction.

For the prediction component, she responded as follows:

> Lane will use the same strategy as above. He will lay out the green triangles into the same shape and size as the yellow hexagon and blue rhombi. Then, Lane will count seven of the triangles to find the whole because he knows 7/7 equals 1 whole. Lane will say 10–7 equals 3 and there are 3 triangles left so this (7 triangles) makes a whole.

The prediction was almost perfectly *accurate* down to each *detail*, satisfying the two criteria for the highest score (4). Bree even correctly predicted that Lane would use a subtractive strategy. Her response also made clear that she used her model in forming the prediction, so the response also earned the highest score for the use of model component.

In contrast to Bree, Tabitha provided low-scoring responses to the prompts about Lane's reasoning. In response to the content knowledge question, Tabitha wrote the fraction "10/7", but did not draw a representation of the whole. For the model component, Tabitha made comments about two of the problems that Lane had solved previously:

> 9/7 can you make the whole?…Lane uses strategies by using the blocks to visually try to make sense of the improper fraction in his mind. For example he says "6 goes into the big hexagon and one more makes 7."
>
> ---
>
> 10/6
> Doggie is 10/6, you need 4, 6 goes into that. Lane really isn't understanding improper fractions.

Tabitha's model did not clearly articulate how Lane was thinking. Her final sentence leaves considerable ambiguity about Lane's thinking and suggests that she had focused primarily on what Lane was able to do correctly or incorrectly and did not reflect the fact that Lane was eventually able to reason correctly about the 10/6 problem. Although she did quote one of Lane's utterances, Tabitha did not use this quote as evidence for a substantive model. Thus, the model component was scored 0.

Tabitha was also unable to successfully predict Lane's response to the final task involving the doggy problem for 10/7. She responded, "I predict that Lane will not know what the whole is." Although this response is related to fractions, it did not specify any

detail about what kinds of actions Lane may have taken in response to the new task. Furthermore, Tabitha's prediction was *inaccurate*, since Lane was able to produce the whole. Thus, Tabitha's prediction was scored 0. Finally, because Tabitha did not use her model to support her prediction, Tabitha scored a 0 on the use of model component.

Methods for establishing reliability and validity of the instrument

After implementing the instrument as described earlier, we used participant responses to establish its reliability and validity. Specifically, we measured inter-rater reliability and construct validity. In addition, our documented efforts in refining the instrument—especially in developing the final version of the rubric (Appendix)—provide a degree of face validity. Finally, we measured associations among the four components that we assessed. We describe our approach to each of these efforts here and then, report on the results of the measures in the next section.

Inter-rater reliability

To establish reliability of using the rubric, each of two raters independently scored all PST responses on both the pre- and post-assessments. PSTs' names and course information had been replaced with codes and randomized so that neither rater knew which responses came from which group. The standard measure of inter-rater reliability is Cohen's kappa, but this measure fails to account for relative differences when the data are ordinal in nature. Thus, for each of the four components, we calculated a weighted kappa, which accounts for ordinal data. Finally, the raters reconciled disparate scores by re-examining PST responses, discussing how each rater scored the response, and negotiating a consensus view on the appropriate score based on the prediction assessment rubric.

Construct validity

In order to document the construct validity of the prediction assessment instrument, we examined the performance of the PSTs who participated in the IMB approach compared with those who participated in the traditional approach. The PSTs enrolled in the field experience using the IMB approach participated in workshops on model building and had opportunities to engage in the process of model building throughout the course of a semester. Thus, the instrument is valid if differences between the two groups appear on the post-assessment (but not on the pre-assessment). This approach to establishing construct validity amounts to turning the instrument on its head: Whereas, in the future, the instrument will be used to distinguish differences between traditional and IMB groups, detecting assumed differences between the two groups now can provide an indication that such use of the instrument is indeed valid. We tested for such differences using chi-square contingency tables.

Face validity

The instrument is intended to measure whether teachers are able to attend to student thinking and to build models of students' thinking. Through the process of creating and refining the videos, response forms, and rubric, we have attempted to establish face validity for the instrument. In particular, the rubric focuses explicitly on criteria associated with good modeling practices (von Glasersfeld and Steffe 1991): The *model* component focuses

on whether the PSTs rely on evidence and whether they attempt to make inferences about how the student is thinking (rather than simply describing what the students seem to know or not know); the *prediction* component focuses on accuracy and detail, as described earlier; and the *use of model* component measures degrees of evidence that the PST used their model to form a prediction. The *content knowledge* component simply measures correctness as an indicator of mathematical knowledge.

Component associations

By considering the association between the four components, we can examine whether they correlate in expected ways. We hypothesized that a lack of content knowledge would limit one's ability to model and predict student thinking. This hypothesis is supported by several previous studies, which indicate the content knowledge is necessary (but not sufficient) for "effective teaching" (e.g., Hill et al. 2004). Similarly, we expected to see positive associations among the various components of modeling: *model*, *prediction*, and *use of model*. Gamma and Somer's d scores are commonly used to examine associations in ordinal data (Goodman and Kruskal 1954). We calculated Somer's d scores for associations between the content knowledge component and the other three components, because Somer's d scores measure the kind of directional associations we hypothesized. We calculated Gamma scores for each pairwise (non-directional) association among the other three components.

Results

Inter-rater reliability

Here, we report on weighted kappa scores for each component of the pre- and post-assessments and use descriptors by Landis and Koch (1977) to interpret those results. Reliability measures for the *content knowledge* component were substantial (0.788, 0.829), and where disagreements occurred, they were typically accounted for by scorer errors rather than real disagreements. The *modeling* component resulted in moderate to substantial reliability (0.518, 0.661). For the *prediction* component, reliability varied from perfect on the pre-assessment (1.000) to fair on the post-assessment (0.395). The *use of model* component resulted in moderate reliability (0.523, 0.474).

It is important to point out that the variability in scoring may be accounted for by the necessarily subjective nature of the rubric. We say "necessarily" because, just as PSTs' models relied on making inferences about student thinking, scorer assessments relied on making inferences about PSTs' unobservable modeling processes. For example, in attempting to assess whether a PST had used their model to make a prediction, it was often difficult to know the role that model building may have played in the prediction. Thus, the scorers must infer from PSTs' responses whether they had used student thinking to help them make the prediction. Such high levels of inference can lead to disparity in scores. We also note that because of differences in the tasks posed in the interviews, the pre-assessment involved predicting a simple verbal response from the student, whereas the post-assessment (as described in the sample responses provided earlier) involved predicting the students' actions. This at least partly explains why the weighted kappa score for the prediction component of the former was perfect.

Table 1 Percentages of PSTs' prediction assessment scores for each component by field experience

Score	Control pre ($n = 19$) (%)	IMB pre ($n = 23$) (%)	Control post ($n = 19$) (%)	IMB post ($n = 23$) (%)
Content knowledge				
0	10.5	4.3	10.5	17.4
1	89.5	95.7	89.5	82.6
Modeling				
0	5.3	26.1	5.3	8.7
1	57.9	39.1	78.9	43.5
2	36.8	34.8	15.8	47.8
Prediction				
0	0.0	0.0	0.0	13.0
1	0.0	0.0	5.3	4.3
2	73.7	65.2	52.6	21.7
3	5.3	8.7	36.8	26.1
4	21.1	26.1	5.3	34.8
Use of model				
0	15.8	52.2	21.1	17.4
1	42.1	21.7	63.2	21.7
2	42.1	26.1	15.8	60.9

Construct validity

We report results of the two groups' performances on the prediction assessment instrument in Table 1. We report results from the chi-square contingency tables in Table 2. For the pre-assessment, there were no significant differences between the two groups for the *content knowledge* ($\chi^2 = 0.599$, $p = 0.439$), *model* ($\chi^2 = 3.489$, $p = 0.175$), and *prediction* ($\chi^2 = 0.390$, $p = 0.823$) components. For the *use of model* component on the pre-assessment, differences did exist between the two groups ($\chi^2 = 6.052$, $p = 0.049$). On the post-assessment, there was no statistical significance between the two groups on *content knowledge* ($\chi^2 = 0.400$, $p = 0.527$). There were significant differences on *model* ($\chi^2 = 5.574$, $p = 0.062$), *prediction* ($\chi^2 = 9.897$, $p = 0.042$), and *use of model* ($\chi^2 = 9.707$, $p = 0.008$).

The statistical tests cited earlier suggest that there was little difference between the two groups at the beginning of the study. The only significant difference was on the use of model component, on which the control group was superior to the IMB group: 42.1% of the control group scored at the highest level on this component, compared with only 26.1% of the IMB group; additionally, 52.2% of the IMB group scored at the lowest level, compared with 15.8% of the control group. After participating in the experimental approach, a higher percentage of IMB PSTs scored at the top level of classification on the post-assessment on modeling (47.8%, compared with 15.8% for the control), prediction (34.8%, compared with 5.3% for the control), and use of model (60.9%, compared with 15.8% for the control). These findings suggest that the prediction assessment instrument has the potential to measure differences in the model building and predictive capabilities that we would expect to see after explicit instruction designed to strengthen these teaching skills.

Table 2 Results of chi-squared comparisons of IMB group ($n = 23$) and control group ($n = 19$)

Scale	Pre	Post
Content knowledge	0.599 ($p = 0.439$)	0.400 ($p = 0.527$)
Modeling	3.489 ($p = 0.175$)	5.574** ($p = 0.062$)
Prediction	0.390 ($p = 0.823$)	9.897** ($p = 0.042$)
Use of model	6.052* ($p = 0.049$)	9.707** ($p = 0.008$)

* Control group was significantly higher at $p < 0.10$

** IMB group was significantly higher at $p < 0.10$

Table 3 Interactions between components on pre-assessment ($n = 42$)

Scale	Modeling	Prediction	Use of model
Content knowledge	−0.162 ($p = 0.702$)	0.333* ($p = 0.075$)	−0.017 ($p = 0.958$)
Modeling		−0.300 ($p = 0.204$)	0.646** ($p = 0.001$)
Prediction			0.200 ($p = 0.429$)

* Association significant at $p < 0.10$

** Association significant at $p < 0.01$

Table 4 Interactions between components on post-assessment ($n = 42$)

Scale	Modeling	Prediction	Use of model
Content knowledge	0.694** ($p = 0.006$)	0.602* ($p = 0.063$)	0.556* ($p = 0.052$)
Modeling		0.667** ($p = 0.001$)	0.895** ($p = 0.000$)
Prediction			0.822** ($p = 0.000$)

* Association significant at $p < 0.10$

** Association significant at $p < 0.01$

Component associations

Interaction scores for the pre- and post-assessments are shown in Tables 3 and 4, respectively. There were only two significant correlations on the pre-assessment: a directional association between content knowledge and prediction; and a non-directional association between modeling and use of model. The former association supports the hypothesis that successful predictions of students' actions rely, in part, on PSTs' own mathematical knowledge. The latter association supports the notion that PSTs who construct weak models have difficulties in using them to make predictions. On the post-assessment, all interactions were significant.

The components showed relatively little correlation on the pre-assessment as compared with the post-assessment. Two factors may have contributed to this difference. First, the pre-assessment was given during the first week of the mathematics methods and field experience classes. Thus, the PSTs had had little experience in thinking about children's mathematical reasoning. By the end of the semester, the PSTs had had opportunities to build their PCK and to connect that knowledge to experiences with children and classrooms. Secondly, the pre-assessment involved a prediction that was relatively straightforward, asking PSTs to predict what the student would say in response to the question posed in the video. Indeed,

both content knowledge and prediction components had more PSTs score lower on the post-assessment. Thus, the post-assessment may have been more difficult than the pre-assessment, and the wider range of scores on the post-assessment would increase likelihood of association between the components, as measured by Somer's d and Gamma scores.

Conclusions

The goal of this study was to establish the utility, reliability, and validity of a video-based instrument for assessing a particularly important aspect of PCK. Previous studies have established the critical role that teachers' knowledge of students' mathematics plays in teaching for understanding (Arbaugh and Brown 2005; Ball 1994; Confrey 1993; Franke et al. 2001). Of course, recognition of this role has already compelled several researchers to develop teacher knowledge assessments centered on understanding students' mathematics.

In this section, we position our instrument's utility relative to those of previous studies. We then summarize findings from measures that support the instrument's reliability and validity as a tool for assessing PSTs' knowledge of students' mathematics. Finally, we consider implications related to associations among the measured components. We intend to use the instrument in the future to distinguish modeling abilities between PSTs participating in innovative professional development programs, such as IMB, and those enrolled in traditional programs.

Utility of the instrument

Our instrument incorporated two major aspects previously used in assessing PCK, namely the use of interviews with children (Crespo and Nicol 2003; diSessa 2007; Wright 2000) and analyzing video (Star and Strickland 2008; Stockero 2008). Rationales for both aspects have been vigorously argued and supported by previous research findings. For example, diSessa (2007) defended the use of clinical interviews against arguments that they occur outside of naturally occurring situations by arguing that valuable results have been produced regarding understanding students' thinking (e.g., the vast contributions of Piaget). Moreover, CGI researchers (Franke et al. 2001) have integrated both aspects—interviews with children and analyzing video—in a demonstrably effective manner that is similar to our own. Thus, we focus on comparisons between their video prediction instrument and ours.

Our instrument differs from that of CGI in several key regards. First, the CGI researchers worked with inservice teachers who were already familiar with the students in the videos and thus, had no need to watch video clips of the students' problem-solving behavior before forming predictions. Second, the teachers in CGI predicted student success and choice of strategies; they learned about common student strategies from the researchers ahead of time and attempted to match students with strategies from that existing set. In contrast, our instrument required PSTs to build models of how students reason and to predict specifically what the students would do in response to a mathematical task. These new models may be based on existing models built from previous experiences with other students in FAIs or classroom interactions, but not from previous knowledge of the particular students in the video, nor from an explicit list of student strategies provided by instructors.

Our rubric also delineates four components related to the process of model building. Collectively, these components correspond to previously identified aspects of good model building: evidence-based inferences explaining how the student is reasoning and usefulness

in forming accurate and detailed predictions (von Glasersfeld and Steffe 1991). The instrument inherits desirable traits from other video-based assessments: It is replicable and therefore sharable; it is authentic in that the videos illustrate students' genuine problem-solving activities; and, after the assessment, the videos can generate valuable classroom discussions around the PSTs' models and predictions. Such discussions might support the kinds of improvements in noticing noted by other researchers (Star and Strickland 2008; Van Es and Sherin 2002), namely helping PSTs advance from descriptive and evaluative noticings toward more analytic and interpretive ones.

Reliability and validity

Arguments provided above and in the methods section support the face validity of the instrument. We emphasize one of those arguments here: The power of any model — regardless of whether it occurs in the natural sciences or in attempts to simulate students' mathematical thinking—is determined by its effectiveness at explaining previously observed activity and predicting future activity. This view was a central tenet for the design of the prediction assessment instrument. Accuracy of predictions, in particular, provided an objective means for assessing the quality of PSTs' models. Other evaluations of reliability and validity are supported by statistical measures.

Weighted kappa scores ranged from 0.395 to 1.000; these scores indicated fair inter-rater reliability for one of the components, moderate reliability for three components, substantial reliability for three other components, and perfect reliability for another component. Interestingly, the lowest and highest scores both occurred for the prediction component, which is largely due to the nature of the student actions in the video. In the pre-assessment, the student provided a simple verbal response; in the post-assessment, the PSTs needed to predict the student's actions, for which accuracy and detail become more subjective descriptors. Overall, the results suggest the rubric provided a sufficient guide for reliable assessments.

The construct validity of our assessments is supported by significant differences in favor of the IMB PSTs over the PSTs in the control group. Because the field experiences of IMB PSTs focused on model building, we should expect valid assessments of model building to show no group differences on the pre-assessment and significant group differences on the post-assessment. This is precisely what we found, with the exception of the *use of model* component on the pre-assessment, which favored the control group. This exception only makes more profound the finding that IMB PSTs significantly outperformed the control group on that component in the post-assessment. The IMB PSTs outperformed the control group on the *prediction* and *model* components in the post-assessment as well. Only differences on the *content knowledge* component remained clearly insignificant, but this component measured a presumed precondition for modeling rather than modeling itself, and nearly all PSTs solved the task correctly on both the pre-assessment and the post-assessment.

Component associations

Previous research would lead us to expect a strong directional association between the content knowledge component and the other three components. In particular, Hill et al. (2004) found strong content knowledge to be a necessary, but not sufficient, prerequisite

for what they call "effective teaching." Indeed, this necessary but not sufficient condition is precisely what gave rise to the notion of PCK, which is knowledge for subject-specific teaching. On the post-assessment, we did find significance in the expected directional association between the content knowledge component and each of the other three components. For example, in the case of Tabitha described earlier, her failure to understand the mathematics of the task likely limited her ability to assess how Lane was thinking about improper fractions and led her to incorrectly predict that he would respond to the problem incorrectly. These results affirm findings from previous studies about the role of content knowledge, and they justify our inclusion of a content knowledge component in the instrument.

Also as expected, each of the other pairwise associations reached significance on the post-assessment. These associations indicate strong connections between the various aspects of modeling that we assessed. On the other hand, only two associations reached significance on the pre-assessment. We have explained this difference by noting that the pre-assessment involved a relatively straightforward prediction on the student's verbal response, whereas the post-assessment required PSTs to predict specific actions. In fact, a different post-assessment video was used in our pilot study, and the post-assessment video described here was developed as another iterative improvement to the instrument, in part to address that issue. Including the prediction of more nuanced student responses to tasks (rather than simple verbal responses) requires PSTs to rely on fine-grained models of students' thinking, and greater associations between the components will become apparent. We also note that stronger associations on the post-assessment might indicate greater maturity in the PSTs' modeling abilities; as the PSTs develop richer models of students' thinking, we would expect stronger inter-dependencies between the model, prediction, and use of model components.

Limitations and implementation

In closing, we share final thoughts about implementation of the instrument, by way of reflecting on potential for further refinements. For example, in future implementations, we intend to use two videos in the pre-assessment, with half of each group (IMB and control) responding to one video and half responding to the other video. Then, in the post-assessment, the PSTs will view the other video so that each will respond to a novel situation. We also intend to replace the pre-assessment video with one that poses questions and tasks that allow PSTs to make more nuanced and detailed predictions about students' actions. These changes will enable us to more accurately assess PSTs' growth in model building. They will also help us determine whether differences in performance on the pre- and post-assessments (especially the lack of component associations that we noticed on the pre-assessment) can be attributed to an initial lack of modeling ability among all PSTs.

In using prediction assessments with cohorts of PSTs, we will also benefit from larger sample sizes. Although we achieved many significant results at the 0.10 level with a relatively small sample size, larger sample sizes will increase the power of statistical tests. Moreover, accompanying qualitative analyses should help us to better understand the nature of group differences and component associations.

Finally, we invite teacher educators to use our instrument in similar studies or as a pedagogical tool for teaching and assessing. All materials, including videos, user's guides, and response sheets, will soon be available at the Web site: http://imb.crlt.indiana.edu/. In addition to its research value, we hope others find value in the instrument's utility in generating classroom discussion about students' mathematical thinking.

Acknowledgments The research reported in this paper was supported by a DR-K12 grant from the National Science Foundation (NSF), under grant number DRL-0732143. The authors wish to thank all the members of the IMB research team, as well as Dionne Cross, who collaborated with us in collecting data.

Appendix: prediction assessment rubric

Content knowledge:
0: Incorrectly solved the problem
1: Correctly solved the problem
Model:
0: Does not use evidence (descriptions of student actions or statements) to describe what or how the student is thinking
1(a): Uses evidence to support an explanation of what the student knows or thinks, but not *how* the student is thinking
1(b): Explains how the student is thinking but does not provide explicit evidence to support this explanation.
2: Uses evidence to support a reasonable explanation for *how* the student is thinking
Prediction (accuracy and detail):
0: Makes no prediction relevant to the situation
1: Makes an inaccurate prediction with some detail relevant to the situation, but not enough to unambiguously envision what the student might have done or said in response to the task/question
2(a): Makes an inaccurate prediction, but with enough relevant detail to envision what the student might have done or said in response to the task/question
2(b): Makes a prediction that might be correct, but remains too vague to determine
3(a): Makes an accurate prediction with some detail relevant to the situation, but not enough to unambiguously envision what the student would do or say in response to the task/question
3(b): Makes multiple predictions, one of which accurately describes what the student would do or say
4: Makes an accurate prediction with sufficient detail to envision what the student would do or say in response to the task/question
Use of model:
0: There is no evidence (or there is counter-evidence) that the PST teacher used an explanation of the students' thinking (model) to form any of the predictions
1: There is some evidence that the PST used a model to form some of the predictions
2: The PST clearly used a model to form most or all of the predictions

References

Arbaugh, F., & Brown, C. A. (2005). Analyzing mathematical tasks: A catalyst for change? *Journal of Mathematics Teacher Education, 8*(6), 499–536.

Ball, D. L. (1994, November). *Developing mathematics reform: What don't we know about teacher learning—but would make good working hypotheses?* Paper presented at Conference on Teacher Enhancement in Mathematics K–6, Arlington, VA.

Carpenter, T. P., Fennema, E., Peterson, P. L., & Carey, D. A. (1988). Teachers' pedagogical content knowledge of students' problem solving in elementary arithmetic. *Journal for Research in Mathematics Education, 19*, 385–401.

Chamberlin, M. T. (2005). Teachers' discussions of students' thinking: Meeting the challenge of attending to students' thinking. *Journal of Mathematics Teacher Education, 8*, 141–170.

Cochran-Smith, M., & Zeichner, K. M. (Eds.). (2005). *Studying teacher education: The report of the AERA panel on research and teacher education*. Mahwah, NJ: Lawrence Erlbaum.

Confrey, J. (1988). *Multiplication and splitting: Their role in understanding exponential functions.* Paper presented at the tenth annual meeting of the North American chapter of the international group for the psychology of mathematics education, DeKalb, IL.

Confrey, J. (1993). Learning to see children's mathematics: Crucial challenges in constructivist reform. In K. Tobin (Ed.), *Constructivist perspectives in science and mathematics* (pp. 299–321). Washington, DC: American Association for the Advancement of Science.

Confrey, J. (1994). Splitting, similarity and rate of change: A new approach to multiplication and exponential functions. In G. Harel & J. Confrey (Eds.), *The development of multiplicative reasoning in the learning of mathematics* (pp. 291–330). Albany: State University of New York.

Confrey, J., & Lachance, A. (2000). Transformative teaching experiments through conjecture driven research design. In A. E. Kelly & R. A. Lesh (Eds.), *Handbook of research design in mathematics and science education* (pp. 231–265). Mahwah, NJ: Lawrence Erlbaum.

Crespo, S. (2000). Seeing more than right and wrong answers: Prospective teachers' interpretations of students' mathematical work. *Journal of Mathematics Teacher Education, 3*(2), 155–181.

Crespo, S., & Nicol, C. (2003). Learning to investigate students' mathematical thinking: The role of student interviews. In N. A. Pateman, B. J. Dougherty, & J. Zilliox (Eds.), *Proceedings of the 2003 joint meeting of PME and PMENA* (Vol. 2, pp. 261–267). Honolulu: College of Education, University of Hawaii.

Davis, B. (1997). Listening for differences: An evolving conception of mathematics teaching. *Journal for Research in Mathematics Education, 28*(3), 355–376.

diSessa, A. A. (2007). An interactional analysis of clinical interviewing. *Cognition and Instruction, 25*(4), 523–565.

Education Commission of the States. (2003a). Eight questions on teacher preparation: What does the research say? Retrieved July, 2006, from www.ecs.org/tpreport.

Education Commission of the States. (2003b). Eight questions on teacher preparation: What does the research say? A summary of the findings. Retrieved July, 2006, from http://www.ecs.org/ecsmain.asp?page=/html/educationIssues/teachingquality/tpreport/.

Educational Testing Service (2006). State requirements. Retrieved June 27, 2006, from http://www.ets.org/portal/site/ets/menuitem.22f30af61d34e9c39a77b13bc3921509/vgnextoid=d378197a484f4010VgnVCM10000022f95190RCRD.

Fennema, E., Carpenter, T. P., Franke, M. L., Levi, L., Jacobs, V. R., & Empson, S. B. (1996). A longitudinal study of learning to use children's thinking in mathematics instruction. *Journal for Research in Mathematics Education, 27*, 403–434.

Franke, M. L., Carpenter, T. P., Levi, L., & Fennema, E. (2001). Capturing teachers' generative change: A follow-up study of professional development in mathematics. *American Educational Research Journal, 38*(3), 653–689.

Goodman, L., & Kruskal, W. (1954). Measures of associations for cross-validations. *Journal of the American Statistical Association, 49*, 732–764.

Hill, H. C., Ball, D. L., & Schilling, S. G. (2008). Unpacking pedagogical content knowledge: Conceptualizing and measuring teachers' topic-specific knowledge of students. *Journal for Research in Mathematics Education, 39*(4), 372–400.

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

Kane, T. J., Rockoff, J. E., & Staiger, D. O. (2006). What does certification tell us about teacher effectiveness? Evidence from New York City. Retrieved July, 2006, from http://gseweb.harvard.edu/news/features/kane/nycfellowsmarch2006.pdf.

Kersting, N. (2008). Using video clips of mathematics classroom instruction as item prompts to measure teachers' knowledge of teaching mathematics. *Educational and Psychological Measurement, 68*(5), 845–861.

Koirala, H. P., Davis, M., & Johnson, P. (2008). Development of a performance assessment task and rubric to measure prospective secondary school mathematics teachers' pedagogical content knowledge and skills. *Journal of Mathematics Teacher Education, 11*, 127–138.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159–174.

Lewis, C. C. (2000, April). *Lesson study: The core of Japanese professional development.* Invited presentation to the Special Interest Group on Research in Mathematics Education at the annual meeting of the American Educational Research Association, New Orleans, LA.

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

Norton, A. H., & McCloskey, A. (2008). Teaching experiments and professional development. *Journal of Mathematics Teacher Education, 11*(4), 285–305.

Olive, J. (2000). Computer tools for interactive mathematical activity in the elementary school. *International Journal of Computers for Mathematics Learning, 5*, 241–262.

Philipp, R. A., Ambrose, R., Lamb, L. L. C., Sowder, J. T., Schappelle, B. P., Sowder, L., et al. (2007). Effects of early field experiences on the mathematical content knowledge and beliefs of prospective elementary school teachers: An experimental study. *Journal for Research in Mathematics Education, 38*(5), 438–476.

Piaget, J. (1970). *Science of education and the psychology of the child*. New York: Viking Press.

Schifter, D. (1998). Learning mathematics for teaching: From the teachers' seminar to the classroom. *Journal for Mathematics Teacher Education, 1*(1), 55–87.

Shulman, L. S. (1986). Those who understand: A conception of teacher knowledge. *American Educator, 10*(1), 9–15.

Star, J. R., & Strickland, S. K. (2008). Learning to observe: Using video to improve preservice mathematics teachers' ability to notice. *Journal of Mathematics Teacher Education, 11*, 107–125.

Steffe, L. P. (1991). Mathematics curriculum design: A constructivist's perspective. In L. P. Steffe & T. Wood (Eds.), *International perspectives on transforming early childhood mathematics education* (pp. 389–398). Hillsdale, NJ: Lawrence Erlbaum.

Steffe, L. P., & Thompson, P. W. (2000). Teaching experiment methodology: Underlying principles and essential elements. In A. E. Kelly & R. A. Lesh (Eds.), *Handbook of research design in mathematics and science education* (pp. 267–306). Mahwah, NJ: Lawrence Erlbaum.

Steinberg, R., Empson, S. B., & Carpenter, T. P. (2004). Inquiry into children's mathematical thinking as a means to teacher change. *Journal of Mathematics Teacher Education, 7*(3), 237–267.

Stockero, S. L. (2008). Using a video-based curriculum to develop a reflective stance in prospective mathematics teachers. *Journal of Mathematics Teacher Education, 11*, 373–394.

Thompson, P. W., Carlson, M. P., & Silverman, J. (2007). The design of tasks in support of teachers' development of coherent mathematical meanings. *Journal of Mathematics Teacher Education, 10*, 415–432.

Vacc, N. N., & Bright, G. W. (1999). Elementary preservice teachers' changing beliefs and instructional use of children's mathematical thinking. *Journal for Research in Mathematics Education, 30*, 89–110.

Van Es, E. A., & Sherin, M. G. (2002). Learning to notice: Scaffolding new teachers' interpretations of classroom interactions. *Journal of Technology and Teacher Education, 10*, 571–596.

Van Es, E. A., & Sherin, M. G. (2008). Mathematics teachers' "learning to notice" in the context of a video club. *Journal of Technology and Teacher Education*, *24*(2), 244–276.

von Glasersfeld, E., & Steffe, L. P. (1991). Conceptual models in educational research and practice. *Journal of Educational Thought, 25*(2), 91–103.

Wallach, T., & Even, R. (2005). Hearing students: The complexity of understanding what they are saying, showing, and doing. *Journal of Mathematics Teacher Education, 8*, 393–417.

Wright, R. J. (2000). Professional development in recovery education. In L. P. Steffe & P. W. Thompson (Eds.), *Radical constructivism in action: Building on the pioneering work of Ernst von Glasersfeld* (pp. 134–151). London: Falmer.