**Assessing Prospective Elementary School Teachers'
Beliefs About Mathematics and Mathematics Learning: Rationale and Development of a
Constructed-Response-Format Beliefs Survey[1]**

Accepted for Publication in *School Science and Mathematics*

Rebecca Ambrose, School of Education, University of California-Davis
Lisa Clement, Randolph A. Philipp, Jennifer Chauvot,
San Diego State University

ABSTRACT

Because of the role beliefs play in the teaching and learning of mathematics, mathematics educators need to consider ways to assess beliefs and belief change. Beliefs, because they must be inferred, can be difficult to measure, particularly with a common metric that enables one to compare individuals. Because of the limitations of Likert scales, we developed a computer-based survey to assess beliefs; in this survey, prospective teachers interpret scenarios in a free-response format. The survey, used with more than 150 participants, captures qualitative data that are later quantified for purposes of comparison. In our work to quantify the qualitative data, we developed a systematic method for creating research rubrics. Results from a pre/post administration of the survey demonstrate that it is an effective tool for assessing belief change. We share the theory behind the development of the survey, some specific information about the survey and the way that responses are coded, and a description of our process for developing rubrics along with some recommendations for researchers interested in developing similar surveys.

Because of the important role beliefs play in the teaching and learning of mathematics (Leder, 2003; Pajares, 1992; Thompson, 1992), mathematics educators need to consider ways to assess beliefs and belief change. Beliefs, because they must be inferred, can be difficult to measure, particularly with a common metric that enables one to compare individuals. As part of our large-scale research project, Integrating Mathematics and Pedagogy, we needed a survey to assess the belief change for a large number of prospective teachers. In this article, we first explain why we considered Likert-scale surveys insufficient for assessing belief change. We then describe the web-based beliefs survey we developed, provide a rationale for its benefits, and describe the process we used to develop research rubrics to code the large set of survey data. Finally, we provide data that indicate that our survey is general enough to capture a range of positions on beliefs and sensitive enough to capture change.[2]

We identify two components of beliefs that account for the critical role beliefs play in teaching and learning and that are thus important for the way we attempt to measure beliefs. First, beliefs influence perception (Pajares, 1992). That is, beliefs serve to filter some complexity of a situation to make it comprehensible, shaping individuals' interpretations of events. Teachers and students are constantly faced with uncertain situations requiring interpretations. In our survey we provide respondents with complex situations that they are asked to interpret. Second, beliefs might be thought of as dispositions toward action, having a motivational force (Cooney, Shealy, & Arvold, 1998; Rokeach, 1968). When teachers face challenging decisions that often must be made spontaneously, their beliefs often compel them to act in particular ways. In measuring beliefs, we provide respondents with scenarios in which they are called on to make teaching decisions. Their dispositions to act in these situations provide us with evidence from which to infer their beliefs.

Beliefs are not all-or-nothing entities; they are, instead, held with different intensities (Pajares, citing Rokeach, 1968). When measuring beliefs, we provide tasks that offer multiple interpretation points; we ask a variety of questions, first general questions and then more specific questions about the complex situations. When we assign scores to responses, we differentiate among strong evidence, evidence, weak evidence, and no evidence of a belief to allow for the different intensities with which individuals hold beliefs. We are careful not to claim that an individual lacks a particular belief but instead state that we found no evidence for the belief in the responses the individual provided. Beliefs tend to be context specific, arising in situations with specific features (Cooney, et al., 1998), and, hence, we situate survey segments in contexts and infer a respondent's belief on the basis of his or her interpretation of the context.

The Problem of Assessing Beliefs of Large Numbers of Students

"For the purposes of investigation, beliefs must be inferred" because individuals can be unaware of some the beliefs that shape their actions (Pajares, 1992, p. 315). Mathematics education researchers have typically used case-study methodology to infer teachers' beliefs related to mathematics teaching and learning (e.g., Clarke, 1997; Cooney, et al.,1998; Raymond, 1997). Using this approach, researchers provide rich descriptions of the beliefs of a small number of prospective teachers by relying on thick data sets that include multiple observations, interviews, and surveys collected over a long period of time. The findings from this research provide details of the conceptions of small numbers of teachers, with conclusions that have multiple data points to support findings. These rich reports are important for theory building, but

---

[2] A browse version of the IMAP Web-Based Beliefs Survey and a manual explaining how to code the survey are available at http://www.sci.sdsu.edu/CRMSE/IMAP/pubs.html.

theory testing often requires tools for studying larger groups of individuals. Given the nature of our work, we faced two problems in assessing beliefs: We needed to assess the beliefs of prospective elementary school teachers years *before* they were in the classroom, and we needed an assessment that could be administered to more than 150 prospective teachers.

Typically, beliefs of large numbers of individuals have been measured using Likert scales (e.g., Bright & Vacc, 1994). However, we identified three problems with Likert scales and attempted to overcome them with the development of a survey that was faithful to the characteristics of beliefs mentioned above. First, with Likert scales, knowing how a respondent interprets the words used in items is difficult. For example, in the item "It is important for a child to be a good listener in order to learn how to do mathematics" (Fennema, Carpenter, Loef, 1990), determining how the respondent might be interpreting the idea of *good listener* is difficult. In addition the respondent must decide to what the child is listening: a teacher demonstrating a procedure, a teacher presenting a problem situation, or another child sharing a solution strategy. Because individual respondents are unable to explain their responses for Likert items, inferring what perceptions may have guided the responses is impossible. In our open-ended, computerized survey, respondents reacted in their own words to questions about learning situations, thus producing more information from which we could infer their interpretations of questions.

Second, Likert items provide no information for determining the importance of the issue to respondents; they may react to questions about beliefs that are unimportant to them. McGuire (1969) stated, "When asked, people are usually willing to give an opinion even on matters about which they have never previously thought" (p. 151). For example, a respondent may agree strongly that when learning mathematics, children need to be good listeners but may not believe that listening matters as much as speaking or other activities. We addressed this issue in our beliefs survey by drawing inferences from that to which respondents attended in learning episodes and when they attended to certain issues.

Third, we think that beliefs can be inferred by determining to what one attends in a complex situation, and Likert scales seldom provide contexts, as for example, in this item, "In mathematics, perhaps more than in other fields, one can find set routines and procedures" (Collier, 1972). Respondents may differentiate between eighth-grade algebra and second-grade arithmetic and believe that one has set routines and procedures and the other does not. This particular item does not supply a context, forcing respondents to consider mathematics in general when, in fact, they view different levels of mathematics in different ways. In our survey, each segment is embedded in a context, so we can better determine to what the respondent's attention was drawn.

Our Beliefs Survey

We set out to create a survey to assess beliefs that might affect prospective teachers' subsequent learning of mathematics: beliefs about mathematics and mathematics understanding and learning. We designed segments to measure only the seven beliefs we had identified (see Figure 1). We wanted a survey from which we could derive a common metric for measuring change in individuals and for comparing individuals to one another. We also wanted a survey that would provide qualitative data that could be used for more holistic analysis. To avoid the limitations of Likert scales, outlined above, and to better capture the characteristics of beliefs we deemed relevant (e.g., that they tend to be context specific and held with varying intensities), we developed a survey in which prospective teachers constructed responses instead of choosing from options provided. We later developed rubrics for quantifying these constructed responses. In addition to providing us with respondents' genuine responses, the constructed-response

feature of the survey engaged respondents in active thinking because they could not merely select from one of several options. In this way we think that our survey provides a more valid measure than one generated from a Likert-scale instrument. One advantage of using this kind of survey is that it can be used for dual purposes. Written responses of individuals can be used to provide insights into their beliefs and interpretations. The numerical scores can be used to statistically analyze differences among groups in different treatments.

Beliefs About Mathematics
Belief 1. Mathematics, including school mathematics, is a web of interrelated concepts and procedures.
Beliefs About Knowing/Learning Mathematics
Belief 2. One's knowledge of how to apply mathematical procedures does not necessarily go with understanding the underlying concepts. That is, students or adults may know a procedure they do not understand.
Belief 3. Understanding mathematical concepts is more powerful and more generative than remembering mathematical procedures.
Belief 4. If students learn mathematical concepts before they learn procedures, they are more likely to understand the procedures when they learn them. If they learn the procedures first, they are less likely ever to learn the concepts.
Beliefs About Children's [Students'] Doing and Learning Mathematics
Belief 5. Children can solve problems in novel ways before being taught how to solve such problems. Children in primary grades generally understand more mathematics and have more flexible solution strategies than their teachers, or even their parents, expect.
Belief 6. The ways children think about mathematics are generally different from the ways adults would expect them to think about mathematics. For example, real-world contexts support children's initial thinking whereas symbols do not.
Belief 7. During interactions related to the learning of mathematics, the teacher should allow the children to do as much of the thinking as possible.

Figure 1. Beliefs measured by survey.

Survey Development

This survey and the accompanying scoring rubrics were developed over a 2-year period by the authors with support from other staff members. We used what we call a recursive cycle of development that included piloting segments of the survey, analyzing prospective teachers' responses to the segments to develop quantification schemes for scoring responses, revising the segments, and piloting them again. We established the validity by having 15 mathematics educators (5 mathematics education researchers who were employed as university professors at various institutions and 10 mathematics education graduate students) complete and comment on the survey. They stated that the survey was reasonable and that it provided scenarios that elicited the beliefs we had identified. The university professors also reviewed student responses with us; they agreed that the responses provided information about the respondents' beliefs.

To further establish the validity of the instrument, we had participants in pilot studies complete the instrument. One group of 15 students was interviewed after completing the survey to determine whether their answers on the computer survey were similar to their spoken answers. Their spoken answers were probed to further establish their beliefs. The interview and computer responses did not differ significantly, and further probing did not reveal evidence that contradicted our conclusions from analyzing their survey responses. During the semester-long pilot study, the students' instructor and the IMAP research team (5 observers) got to know the students' beliefs through class discussions and written work. In addition, the students were

interviewed on a periodic basis. The instructor and researchers found that the students' beliefs-instrument scores were consistent with beliefs assessments based on these observational data.

The survey is comprised of seven segments, each of which includes several questions about a particular situation. Four segments are in the domain of whole number, two are in the domain of fractions, and one is a more general teaching segment. Whole numbers and fractions were the domains of focus for our experimental treatments and were important topics in the mathematics-for-teachers course in which the prospective teachers in our study were enrolled. Two segments include video clips of individual children solving mathematics problems with an interviewer. Each segment is associated with two or three beliefs, and each belief is assessed using a separate rubric for each of two or three segments. Overall, we developed 17 rubrics for the survey. Because of the length and the complexity of the survey, we are unable to provide the entire survey in this article. Four segments are included to provide readers with information on which to base their own conclusions about the face validity of the survey.

To illustrate how we assigned scores for each belief, we describe one segment, the rubric used to assign scores to prospective teachers' responses to that segment, and the scoring system used to combine scores on individual rubrics to determine an overall score for the belief. The scores on segments and on beliefs reflect the amount of evidence a respondent provided related to the belief. This scoring is in keeping with the idea that beliefs can be held with different intensities (Rokeach, 1968). In segment 7, prospective teachers watch a video clip of a teacher presenting a story problem to a 6-year-old child in a one-on-one setting: "There are 20 kids going on a field trip. Four children fit in each car. How many cars do we need to take all 20 kids on the field trip?" After a long pause, the child states 10 as his answer. He confirms that he had guessed when the teacher asks how he got his answer. The teacher then directs the child to show her the kids by counting out 20 cubes. She reminds him that 4 kids fit in one car and asked him to show her 4 kids in one car. She directs him to make another group of 4 for the next car, and he follows her directions. She continues in this fashion until he has made five groups of 4 cubes. She then reminds him that each group stands for a car and prompts him to count each of the cars. She counts along with him. The respondents were asked to react to the episode and to identify the strengths and weaknesses of the teaching in it (see Figure 2 for the complete text of the segment).

In this part of the survey, you will be watching an interview with a child.

The following problem is posed to the child:
There are 20 kids going on a field trip. Four children fit in each car. How many cars do we need to take all 20 kids on the field trip?

Click HERE to see the video

7.1 Please write your reaction to the videoclip. Did anything stand out for you?

(The next two questions are asked on a new screen after the respondents has submitted a response to 7.1)
7.2 Identify the strengths of the teaching in this episode.

7.3 Identify the weaknesses of the teaching in this episode.

Figure 2. Survey Item for Segment 7

Our interpretation of this clip was that the teacher, who conducted the session as she did at our request, was overly directive and focused the child's attention on counting cubes instead of on understanding the relationships among the quantities in the situation. She could have provided prompts that were less specific, to see whether the child could solve the problem with less help. For example, she might have invited him to try to use the cubes to represent the situation and then waited to see what he would do before providing him with additional help. The clip featured a familiar real-world context and manipulatives. The teacher was encouraging in her tone of voice and in providing the child with praise. These positive features of the clip were quite attractive to some respondents, leading them to focus on these aspects instead of on the excessive guidance offered by the teacher, as is evident in the following response:

> I thought it was good that she let him try and answer the problem first and then she showed him how to figure it out using the blocks. ... They need to test things out themselves and then see the different ways to approach a problem. … I think the strengths of this video were allowing the child to think on his own and solve the problem. … I didn't see any weaknesses in this video clip. I really liked it.

In addition to being impressed with the teacher's use of blocks, this respondent wrote about the importance of letting the child figure out the problem for himself. She used the rhetoric that we would like prospective teachers to employ, but in this case she applied the rhetoric in a context in which we believe it was inappropriate. Responses like this one reminded us of the critical role that context played in inferring beliefs from responses. Without knowing the context to which the comment "allowing the child to think on his own" was directed, one might interpret this response as providing strong evidence of Belief 7: *During interactions related to the learning of mathematics, the teacher should allow the children to do as much of the thinking as possible*.

No Evidence Overall satisfaction with guidance provided by teacher No teaching weaknesses identified	**No Evidence** Thought the teacher should explain more.
Weak Evidence In initial response, does not mention teacher's excessive guidance. In response to second prompt, expresses satisfaction with the guidance offered by the teacher. In response to third prompt, points out that child may not have needed so much help.	**Weak Evidence** In initial response, suggests that this problem was too hard and inappropriate for this child to solve.
Evidence In initial response, does not mention that the teacher did too much leading. In response to second prompt, identifies cubes or story problem or positive reinforcement as strengths but does not talk about teacher's guidance as strength. In third response, critiques teaching for being too leading.	
Strong Evidence In initial response, notes that the teacher was too leading. In response to third prompt, criticizes teaching for being too leading.	

Figure 3. Scoring rubric for Belief 7/Segment 7.

The complete rubric used to score this segment (for Belief 7) is provided in Figure 3. We were particularly concerned about whether the respondents noted that the teacher did too much leading and, if so, when they noted that fact. (See Figure 4 for examples of different kinds of responses to this segment.) Those who noted excessive guidance in their response to the first prompt ("Please write your reaction to the video clip. Did anything stand out for you?") provided strong evidence of this belief because the issue was of such importance to these respondents that it shaped their interpretations of the episode. For the subsequent prompt ("Identify the weaknesses of the teaching in the episode"), some respondents noted that the teacher might have provided too much guidance. In this case, we determined that the respondents provided some evidence of the belief. It was not strong evidence because the issue was not of sufficient significance to shape their initial interpretations of the clip. Other beliefs shaped their interpretations of the teaching situation. Because the survey was web-based, respondents could not change earlier responses. This feature of the survey allowed us to capture their first reactions before they had been affected by subsequent questions we might ask. For example, in this segment, we did not want the respondents to change their initial responses once they realized that we were interested in teaching weaknesses. Using open-ended questions allowed us to discern which issues were significant enough to respondents to affect their initial interpretations.

Each belief was measured by more than one segment to give a valid measure. In the case of Belief 7, the second segment used to measure the belief included the only general (not context-specific) segment: Respondents were asked, first, whether they would ever ask children to solve problems without first showing them how and, second, to explain their answer. Their explanations were assigned scores on the basis of the amount of autonomy they planned to provide for children and their rationale for doing so. Respondents who suggested that children understand more mathematics when they devise their own solution strategies scored highest for the segment.

To determine a final score for each belief, we combined individual scores from each rubric. Because we treated these scores as ordinal, we did not sum scores or calculate means. We developed a rubric-of-rubrics system that could be applied to each belief. In this system, we accounted for the differing strengths of beliefs by having a range of scores for the rubrics and the beliefs scores.

Rubric Development

We offer a description of rubric development to establish the validity of both the process and the rubrics that we developed. We also offer it so that other researchers who decide to develop their own rubrics can improve upon the process, having a clear idea of the intensity of this process before embarking on it. We found little in the research literature to guide us in the process of developing rubrics; therefore, describing this methodology is warranted. Sharing emergent methodologies is not a new idea. Glaser and Strauss (1967) called for others "to codify and publish their own methods for generating theory" (p. 8, cited in Cobb & Whitenack, 1996), and others have published their methods for analyzing data (e.g., Cobb & Whitenack, 1996).

	7.1 Please write your reaction to the videoclip. Did anything stand out for you?	7.2 Identify the strengths of the teaching in this episode.	7.3 Identify the weaknesses of the teaching in this episode.
Strong Evidence	Well first the child guessed when he couldn't think of it off the top of his head. Then he was asked to use blocks. It seemed like he was really being pushed in the direction the interviewer wanted him to go instead of going there on his own.	The teaching had a lot of positive reinforcement, I have to give it that. She was very positive when he did something and she seemed really positive throughout the entire video.	She pushed the child a little too hard instead of letting him go there on his own. Maybe it was because he wasn't going to find it on his own, but I really don't think he knew why he was counting all the groups of four, other than because he was told to.
Evidence	Before the child used manipulatives, he really couldn't even fathom what it was he needed to do. As soon as the manipulatives were introduced, he seemed to get more of a focus, and with the teacher's help, he got the problem right.	Manipulatives made the concept concrete. The child could actually see the 4 "kids" in each of the 5 "cars."	It seemed the teacher led the child to the answer, and I would want to see if the child, once introduced to the manipulatives, could have figured out what to do with a few questions from the teacher, rather than the teacher saying "now do this, and this etc"
Weak Evidence	Once the teacher scaffolded the student he was able to make some conjectures on his own.	The teacher guided the student well. She did not give too much information and solve the problem. She also did not give too little information so that the problem was difficult to do	I think the student could have completed more of the problem on his own.
No Evidence	Manipulatives are very powerful. They make the abstract concrete.	'The teacher used manipulatives and explained every step of the procedure as she thought out loud about the problem.	'No obvious weaknesses.

Figure 4. Sample responses for Segment 7 and their Scores for Belief 7.

Gronlund (1998) provided a basic definition of the term *scoring rubric*: "a set of scoring guidelines that describes the characteristics of the different levels of performance used in scoring or judging a performance." We developed 19 scoring rubrics for eight segments that assessed seven beliefs. Because of inadequate reliability measures on 2 of the rubrics, we eliminated them from our final data analysis, along with the segment with which they were associated. Additionally, we removed one belief from our original list of eight beliefs, because we could not develop segments or rubrics to validly and reliably assess responses about that belief.

The methods employed by our teams of researchers were different from those used by many who develop rubrics for use in classrooms. In particular, classroom rubrics are often created and then shared with students so that students have guidelines from which to construct responses. Also, classroom rubrics are often global in nature; for example, on a website dedicated to rubric development, a rubric is provided for assessing writing mechanics of students in Grades 1–6. The criteria for the highest score are "One or two minor errors. No major errors" (Herman, Gearhart, & Baker, 1993, cited on http://intranet.cps.k12.il.us/Assessments/Ideas_and_Rubrics/Rubric_Bank/rubric_bank.html). We needed to develop rubrics that captured detailed information about respondents' beliefs about mathematics and mathematics teaching and learning. Because we inferred their beliefs from responses to specific contexts, we needed to develop *task-specific* rubrics, specific to a particular segment about a particular belief (Moskal, 2000).

Two research teams of 2–4 members met two to four times per week in 3–4-hour sessions for 6 months. To get a wide range of responses for each segment, we initially gathered data from three groups: prospective-teacher participants in a pilot of an experimental treatment (pre and post data), expert mathematics educators, and mathematics education graduate students. We later gathered responses from prospective teachers in a second pilot of the treatment; thus, our rubric development was based on a set of about 80 responses. We used this set of responses to create rubrics for the purpose of assigning numerical codes.

To develop each rubric, we adopted a grounded-theory approach (Glaser & Strauss, 1967), using pilot data. Rubrics were designed to align with the data and the types of responses we received instead of being based on a set of predetermined criteria. To begin the process, each person on the team independently read the entire set of responses on a particular segment with a particular belief in mind and sorted the responses into categories. Those responses that provided the greatest evidence of the belief were placed into one category, whereas those that provided no evidence of the belief were placed into another category. The remaining responses were placed into one, two, or three groups, depending on how each team member categorized responses. To determine the appropriate category for each response, the team members looked for degrees of evidence for the belief in question.

After individual team members had sorted the responses, they met to compare their categories and to develop descriptions for the categories. During these first meetings, team members tended to agree on the responses that showed the greatest and least evidence for the belief but had greater difficulty coming to consensus on responses that provided only partial evidence for the belief or, as was sometimes the case, responses that provided evidence for the belief in response to one part of a segment only to provide disconfirming evidence for another part of the same segment. For example, in one segment designed to assess the belief that a person might be able to perform a procedure

without understanding, respondents were asked to state whether a student (Carlos) who could perform the standard algorithm for addition could understand and explain a student's (Sarah's) compensating strategy. (See Figure 5 for segment). One respondent wrote, "Yes because Sarah and Carlos show they understand although Carlos might not understand and might just know how to carry a 1." This response provides conflicting evidence about what Carlos understands; team members had to make decisions about how to categorize such responses.

Carlos	Elliott	Sarah
149 + 286	149 + 286	149 + 286
Written on paper	Written on paper	
1 1	149	Sarah says, "I know that 149 is only 1 away from 150, so 150 and 200 is 350, and 80 more is 430, and 6 more is 436. Then I have to subtract the 1, so it is 435."
149	+286	
+286	300	
435	120	
	15	
	435	

3.4. Do you think that Carlos could make sense of and explain Sarah's strategy? Why or why not?

3.5. Do you think that Carlos could make sense of and explain Elliott's strategy? Why or why not?

Figure 5. Segment 3.4.

During these discussions, descriptions of categories emerged. Quite often, the group agreed on which responses belonged in a particular category but had difficulty developing a written description for the category. The challenge became to make the implicit features of the category explicit. We needed descriptions that were robust, that described the nature of evidence for the belief that the responses provided, but that were also procedural and concrete, so that others using the rubrics could code the responses with a high degree of reliability. This rubric-negotiation process was lengthy, and the development process took approximately 72 person-hours per rubric (4 weeks x 6 hours per week x average of 3 people per team). Sometimes negotiations concerned the number of categories, whereas in other cases, negotiations concerned the descriptions of the categories. We often traversed the terrain from the theoretical to the practical. We described categories to one another, then re-analyzed the data to check that the descriptions aligned with all the responses assigned to the category.

Once we had reached consensus on a rubric's categories and descriptions, we re-analyzed the data to check for inter- and intra-rater reliabilities. We then shared the rubric with

the other team to test for coherence, reliability, and validity, a critical component of our work. The other team's members used the rubrics to code the data; we then compared the development group's codes with the testing group's codes to determine inter-rater reliability. We sought at least 80% agreement; if we did not achieve that level, we further clarified the rubric descriptions. We also discussed issues of validity to ensure that the scores were representative of the amount of evidence for the belief we were claiming to measure.

An Example of Rubric Design

Using two different rubrics, we measured one belief (Belief 6) about children's learning of mathematics:

> The ways children think about mathematics are generally different from the ways adults would expect them to think about mathematics. For example, real-world contexts support children's initial thinking whereas symbols do not.

We used responses to segment 8 about fractions to infer the respondents' support (or lack thereof) for Belief 6 (see Figure 6 for the segment, see Figure 7 for responses to items 8.1 and 8.2). Specifically we noted whether respondents recognized that children often misinterpret the sizes of fractions when they encounter symbolic representations and that contexts provide children with support in understanding the relative sizes of fractions (Mack, 1990). The greatest challenge in developing this rubric was to appropriately describe each of the three categories, particularly the middle category, that emerged from the data. We struggled to describe responses like that in Figure 7, which we had placed in the middle category

---

8.1) Place the following four problems in rank order of difficulty for children and explain your ordering (you may rank two or more items as being of equal difficulty). NOTE. Easiest = 1.

a) Understand $1/5 + 1/8$ Rank: ____   Please explain your rank.

b) Understand $1/5 \times 1/8$ Rank: ____   Please explain your rank.

c) Which fraction is larger, 1/5 or 1/8,

     or are they same size?   Rank: ____   Please explain your rank.

d) Your friend Jake attends a birthday party at which five guests equally share a very large chocolate bar for dessert. You attend a different birthday party at which eight guests equally share a chocolate bar exactly the same size as the chocolate bar shared at the party Jake attended. Did Jake get more candy bar, did you get more candy bar, or did you and Jake each get the same amount of candy bar?   Rank: _____   Please explain your rank.

Consider the last two choices:

8.2 Which of these two items did you rank as easier for children?

___ Item c is easier than Item d. ___ Item d is easier than Item c ___ The two items are equally difficult.

Please explain your answer.

Figure 6. Segments 8.1 and 8.2.

Explain Item c Comparing 1/5 and 1/8	Explain Item d Word problem	Choose *c* or *d*	Explain Choice of Which Item Is Easier
I think this problem is pretty simple once the child has it explained to him/her. They could use visual aids or any other method of viewing which fractions are larger and smaller.	This story problem paints the picture and is more understandable because you know why the answer is what it is.	d is easier	It illustrates the answer so that you can visualize the candy bar and the amount of children at the party which helps you visualize how much candy each child would receive.

Figure 7. Sample response to Segments 8.1 and 8.2 on Beliefs Survey.

In an early version of the rubric, we described the middle category in the following way: "Says Item d is easier than Item c but has a weak explanation." Group members realized that the term *weak* was insufficiently clear to describe this kind of explanation for future coders. In this instance, we assigned the response in Figure 7 the middle score because of the ease with which the respondent thought a child could compare the fractions 1/5 and 1/8 in symbolic form, showing that the respondent was insensitive to the challenge that symbolic representations can pose for children. In another draft we focused on the respondent's claims about Item c, stating, "Says that Item d is easier but tends to think that c is either relatively straightforward OR would be difficult for reasons that are NOT related to the ways children typically approach the problem." This description was later revised to also include the reasons that respondents provided for Item's d being easier than Item c. The final version (see Figure 8) was more specific than the earlier versions and more focused on the degree of evidence that the respondent provided for the belief. We devised at least six versions of this rubric, each more detailed and more focused than the previous one.

In rubric development, our first concern was validity; we asked ourselves the question "What does this kind of response tell us about this belief?" Given the open-ended nature of our questions, sometimes respondents provided insightful responses that did not provide us with evidence about the belief in question. We were tempted to assign high scores to such answers, but we learned to look not at the overall quality of the response but rather for evidence of a specific belief. Our second concern was reliability; thus, we sought to develop rubric descriptions that would be clear for others using them. For this rubric, our coders (10 graduate students external to the project) achieved 87.5% interrater reliability (the target for interrater reliability is typically set at 80%).

The coders worked in teams of two, focusing on one rubric at a time. The teams changed each day to code responses using a different rubric. The mean interrater reliability for all 17 rubrics was 84% for the responses of the 159 participants who completed this survey before and after taking part in one of five treatments. At least 20% of the responses were coded by two coders; the remaining responses were coded by one coder. We maintained high interrater reliability by employing an on-going training model in which each coder evaluated papers in groups of 40; 10 of the 40 papers were also being coded by another coder. Scores of the matching papers were evaluated for agreement. When agreement was below the 80% target, coders were retrained to ensure that their interpretations of the rubrics were consistent. In the case of disagreement, the coders discussed their scores and came to an agreement on the final score. The coders were also encouraged to identify responses that they wanted to discuss because they found them difficult to code. These discussions served not only to generate a final score for a response but also to deepen the coders' understanding of each rubric.

No Evidence	No Evidence
Says that Item c is easier (or that c and d are equally difficult), AND the explanation indicates no or little appreciation for the use of real-world context to support children's understanding.	Says that Item d is easier BUT gives either inconsistent explanations (that is, explanations that indicate that they think Item c might be easier) or a clear focus in 8.2 on the teacher's role in showing students how to solve the problem[a]
Evidence	**Evidence**
Says that Item c is easier BUT expresses great appreciation for real-world context	Says that Item d is easier (or that Items c and d are equally difficult) AND expresses some appreciation for real-world context or says that Item d is easier to visualize, BUT does not recognize the difficulty some students have comparing fractions in symbolic form.
Strong Evidence	
Says that Item d is easier (or that Items c and d are equally difficult) AND notes that Item d is easier because of the real-world context or says that d is easier to visualize AND notes that Item c is more difficult because of the abstractness or confusion of the numerals (or says that Item c is more difficult to visualize)	

[a] When the responses for Items 8.1c or 8.1d appear to conflict with the response for Item 8.2, give MORE weight to the response in Item 8.2.

Figure 8. Final rubric for Belief 6, Items 8.1 and 8.2.

How Effective Is the Survey?

We were interested in determining whether the beliefs survey was sensitive enough to yield a range of scores on each of the seven beliefs and whether the survey would measure belief change. For 159 prospective teachers enrolled in the first of four mathematics courses for prospective elementary school teachers, we administered the assessment as a pretest (at the beginning of the course) and as a posttest (at the end of the course). (In our Integrating Mathematics and Pedagogy study we assigned these 159 prospective teachers to treatments, but presentation of those data goes beyond the scope of this article.) Pretest results indicate that a large percentage of the prospective teachers initially showed *no evidence* of holding each belief (see Table 1), and nearly all the prospective teachers fell into one of two categories, showing either no evidence or weak evidence. We found variation in scores in the pretest, showing that the survey captured individual differences.

Posttest results indicate that many prospective teachers' beliefs changed over the semester, with far fewer *No Evidence* scores and a greater number of *Strong Evidence* scores on the posttest (see Table 1). Many prospective teachers' responses were still coded as showing no evidence of the beliefs. We interpret these results as indicating that our beliefs survey was not simply a measure of information that could be easily learned or parroted back to us by prospective teachers over the course of a semester in which they

took a mathematics course for prospective teachers. The prospective teachers could not clearly determine the preferred answers and they supplied a range of responses. The open-ended nature of the segments allowed prospective teachers' beliefs to emerge. Clearly, some had changed and some had not.

Table 1.*Pretest and Posttest Scores by Percentage (and Number [*n *= 159])*

Belief	No evidence		Weak evidence		Evidence		Strong evidence	
	%	*n*	%	*n*	%	*n*	%	*n*
B1 Pre	60	(95)	30	(47)	9	(15)	1	(2)
B1 Post	19	(30)	40	(64)	25	(39)	16	(26)
B2 Pre	77	(122)	14	(23)	7	(11)	2	(3)
B2 Post	40	(63)	20	(31)	21	(34)	19	(31)
B3 Pre	64	(102)	13	(20)	18	(29)	5	(8)
B3 Post	28	(44)	11	(18)	26	(41)	35	(56)
B4 Pre	72	(114)	23	(36)	5	(9)	0	(0)
B4 Post	36	(57)	35	(56)	14	(22)	15	(24)
B5 Pre	33	(53)	45	(72)	18	(28)	4	(6)
B5 Post	16	(25)	34	(54)	29	(46)	21	(34)
B6 Pre	41	(65)	36	(57)	18	(29)	5	(8)
B6 Post	18	(28)	35	(56)	25	(40)	22	(35)
B7 Pre	71	(112)	25	(39)	4	(7)	0	(0)
B7 Post	40	(64)	36	(58)	19	(30)	4	(7)

Conclusions

We were guided in the design of our beliefs survey and in rubric development by four features of beliefs. We situated segments within contexts because beliefs tend to relate to specific contexts. We assume that people hold beliefs with different intensities, with some beliefs being stronger than others, so we differentiated between strong evidence of beliefs and lesser evidence of beliefs. Beliefs shape interpretations, so we used respondents' interpretations of events as evidence of their beliefs. Beliefs dispose people toward particular actions, so we provided respondents with opportunities to make teaching decisions from which we could infer beliefs. We inferred beliefs from constructed rather than multiple-choice responses so that we could get respondents' genuine reactions to situations without their being influenced by

someone else's reaction. Through this practice, strong beliefs surfaced because those are the beliefs that drive interpretations and decisions.

We recognized our reliance on inference throughout this process and chose to label our categories according to the amount of evidence for a belief that a respondent provided rather that to make more definitive statements about whether the respondent held a particular belief. We acknowledged that given different scenarios, respondents' beliefs scores might look different. We measured what we considered core beliefs for the learning and teaching or elementary school mathematics and provided contexts in which these beliefs could emerge. From this sample of interpretations and decisions, we inferred the intensity with which respondents held those core beliefs. A major strength of our survey is that it uses video clips and learning episodes to create contexts to which users respond in their own words rather than choose from one of several options. This format provides qualitative data that can be used for a variety of purposes. It also provides detailed information about the respondents' interpretations of the questions they are asked. This strength comes with a cost in terms of time required for coders to learn to use the rubrics and translate the constructed responses into quantified responses. Whether this "price" of coding will be too high for those seeking a beliefs survey is an important question left to be answered. We might answer this question by creating multiple-choice answers for the segments on the existing survey and determining whether the results prove to be similar to those of the constructed-response survey.

The mathematics of the survey was whole number place value and rational number. We suspect that the survey would have been different were it intended for different content—say, geometry—or for a different population—say, preservice secondary school teachers. For example, we have given little thought to whether different approaches are needed for investigating the relationship between concepts and procedures in geometry or algebra. In developing segments for the survey, we were mindful that the immediate audience for the survey would be prospective elementary school teachers early in their course taking. We limited the survey to teaching scenarios with which they would be familiar and teaching decisions of the type they expect to make. Although the survey was designed for this specific population, we anticipated its use with a broader population, including students in mathematics methods courses and practicing teachers. No aspect of the survey precludes its use with these populations, and we have piloted it with practicing teachers as well as with community college instructors who teach mathematics for teachers courses. None of the teachers or instructors questioned the relevance of the segments or found them to be inappropriate. The nature of the responses from these groups might differ from those of the prospective-teacher population because of the teachers' more sophisticated knowledge of teaching mathematics. Rubrics might require adjustments to account for some of the responses that might emerge when used with other populations to make them more sensitive to the differences within each group.
Our pilot work using the survey with in-service teachers has been encouraging, because we have found that with that group we capture a range of scores. We believe that the survey has potential as a measure of in-service teachers' beliefs and that the use of the survey with in-service teachers warrants further research. We make no claim about the efficacy of this survey with secondary school teachers.

Although our survey measured change between the beginning and end of our treatments, it seemed neither "too easy" nor "too difficult"; that is, it measured neither a floor effect nor a ceiling effect. Although the high scores on the pretest were few, we did measure variation, and on the posttest we found low- and high-scoring prospective teachers.

Beliefs are inferred by someone who holds beliefs. The most those inferring the beliefs can do is to be clear about what those beliefs are and how those beliefs were operationalized so

that others considering using the survey can decide whether they value those beliefs and whether they agree with how those beliefs were measured. We would be presumptuous to claim that we have created a survey to measure all beliefs related to mathematics and mathematics learning, so we will state only that we have created a survey that measures seven specific beliefs about elementary school contexts. We think that we have developed a beliefs survey that can effectively measure quantitative differences while still capturing the individual voices of respondents that, in the past, have been captured only through more intense qualitative approaches.

We have found the use of rubrics for research purposes to be quite promising, because the responses offer a rich data set that is typically unheard of when studying conceptions of large numbers of participants. Yet, we provide a word of caution. The work of rubric development requires the resources of time, money, and large numbers of persons qualified to develop and code rubrics. The end result, however, is a survey that can be used for a variety of purposes, both quantitative and qualitative. We end with recommendations for those who may be considering developing their own rubrics.

Recommendations for Developing Rubrics

- Be clear on the definitions of the constructs being measured, because these definitions serve as the foundation for your work. We referred to our statements of beliefs over and over while we developed rubrics and found that the clearer these statements were, the easier it was to develop the rubrics.
- Use a team of two to four people when developing research rubrics. Rubric development cannot be done alone. Interpretations of responses can vary widely, but one person cannot know how others will interpret responses unless others are simultaneously examining them. We found three people per team to be ideal. Also, having two coding teams allowed us to have other experienced coders serve as reliability and validity checks while we finalized the rubrics.
- Decide on a particular number of categories beforehand, but do not feel constrained to use that number of categories; for example, begin with four categories: responses that provide no evidence, weak evidence, evidence, and strong evidence of the belief. Then decide on particulars for that belief. We did not use this strategy, but we think that the approach could have facilitated our work, because it might have guided our initial conversations about responses.
- Recognize the interpretative nature of the rubrics, and do not try to develop a rubric that captures every response. From any 100 responses, we typically found that 5–10 were unique and did not fall into any of our categories. In some cases a response was so brief that it was impossible to categorize. When responses did not fit any of the categories on the rubric, coders were advised to consider the belief and the amount of evidence for the belief supplied by the respondent.

References

Bright, G. W., & Vacc, N. N. (1994, April). *Changes in undergraduate preservice teachers' beliefs during an elementary teacher-certification program*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Clarke, D. M. (1997). The changing role of the mathematics teacher. *Journal for Research in Mathematics Education, 28,* 278–308.

Cobb, P., & Whitenack, J. (1996). A method for conducting longitudinal analyses of classroom videorecordings and transcripts. *Educational Studies in Mathematics 30*, 213–228.

Collier, C. P. (1972). Prospective elementary teachers' intensity and ambivalence of beliefs about mathematics and mathematics instruction. *Journal for Research in Mathematics Education, 3,* 155–163.

Cooney, T. J., Shealy, B. E., & Arvold, B. (1998). Conceptualizing belief structures of preservice secondary mathematics teachers. *Journal for Research in Mathematics Education, 29,* 306–333.

Fennema, E., Carpenter, T. P., & Loef, M. (1990). *Mathematics beliefs scales.* University of Wisconsin-Madison.

Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research*. New York: Aldine.

Gronlund, N. E. (2003). *Assessment of student achievement*. Boston, MA: Allyn & Bacon

Herman, J. L., Gearhart, M., & Baker, E. (1993). Assessing writing portfolios: Issues in the validity and meaning of Scores, *Educational Assessment*, *1* (3), (cited on http://intranet.cps.k12.il.us/Assessments/Ideas_and_Rubrics/Rubric_Bank/rubric_bank.html).

Mack, N. K. (1990). Learning fractions with understanding: Building on informal knowledge. *Journal for Research in Mathematics Education*, 21, 16–32.

McGuire, W. J. (1969). The nature of attitudes and attitude change. In G. Lindzey & E. Aronson (Eds.), *The handbook of social psychology* (pp. 136–314). Reading, MA: Addison-Wesley.

Moskal, B. (2000). Scoring rubrics: What, when and how? *Practical Assessment, Research, & Evaluation, 7* (3). Available online: http://ericae.net/pare/getvn.asp?v=7&n=3

Pajares, M. F. (1992). Teachers' beliefs and educational research: Cleaning up a messy construct. *Review of Educational Research, 62,*, 307–332.

Raymond, A. M. (1997). Inconsistency between a beginning elementary school teacher's mathematics beliefs and teaching practice. *Journal for Research in Mathematics Education, 28*, 550–576.

Rokeach, M. (1968). *Beliefs, attitudes, and values: A theory of organization and change.* San Francisco: Jossey-Bass.

Thompson, A. (1992). Teachers' beliefs and conceptions: A synthesis of the research. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 334–368). New York: Macmillan.