A METHOD FOR DEVELOPING RUBRICS FOR RESEARCH PURPOSES

Lisa Clement, Jennifer Chauvot, and Randolph Philipp

San Diego State University

Rebecca Ambrose

University of California at Davis

Preparation of this paper was supported by a grant (REC-9979902) from the National Science Foundation (NSF). The views expressed are those of the author and do not necessarily reflect the views of NSF.

Citation:
Clement, L., Chauvot, J., Philipp, R., & Ambrose, R. (2003). A method for developing rubrics for research purposes. In N. A. Pateman, B. J. Dougherty, & J. T. Zilliox (Eds.), *Proceedings of the 2003 joint meeting of PME and PMENA* (Vol. 2, pp. 221–227). Honolulu: CRDG, College of Education, University of Hawaii.

A METHOD FOR DEVELOPING RUBRICS FOR RESEARCH PURPOSES[1]

Lisa Clement, Jennifer Chauvot, Randolph Philipp, San Diego State University

Rebecca Ambrose, University of California at Davis

A methodological approach that emerged during the design of task-specific research rubrics to code large sets of open-ended survey data fills the void in scholarship about developing rubrics for research purposes. A brief rationale for using this method rather than other, often-used, data analysis methods is provided, with a description of the methodology, using an example to support the description. Finally, recommendations are included for those who plan to undertake the task of rubric development for research purposes.

RATIONALE AND CONTEXT

Beliefs are a central construct to those interested in research on teaching and learning mathematics. At least as difficult as defining *belief*, is devising instruments that operationalize a definition. We describe the process we used to develop research rubrics to code large sets of belief-survey data, but we first state the four components of beliefs that guided us in the development of our instrument: Beliefs influence perception; they are not all-or-nothing entities; they are context specific; and they are dispositions toward action (for more information about the beliefs we assessed, see Ambrose, Philipp, Clement, and Chauvot, 2003)

Mathematics education researchers have typically used case-study methodology to analyze teachers' beliefs related to mathematics teaching and learning (e.g., Clarke, 1997; Cooney, Shealy, & Arvold, 1998; Raymond, 1997). This approach provides rich descriptions of the beliefs of a small number of preservice elementary school teachers (PSTs), typically no more than four in one study. Its strength is that it relies on thick data sets that include multiple observations, interviews, and surveys that are collected over a long period of time. The findings from this research inform the community about details of the conceptions of small numbers of teachers, with conclusions that have multiple data points to support findings. These rich reports of a small number of cases are important for theory building, but theory testing often requires tools for studying larger groups of individuals.

One means for studying the beliefs of large groups of individuals is through Likert scales, which are often used in a pre- and post-testing to measure change before and after some treatment (e.g., Bright & Vacc, 1994). Likert scales are used in many fields, and are widely accepted research instruments. However, the Likert scales typically used in mathematics education use statements that are decontextualized, so that results are difficult to interpret, while the voices of the individuals go unheard,

[1] Preparation of this paper was supported by a grant from the National Science Foundation (NSF) (REC-9979902). The views expressed are those of the authors and do not necessarily reflect the views of NSF.

thereby making inferring what perceptions may have guided the responses difficult (for more information about our rationale for not using Likert scales, see Ambrose, Philipp, Clement, and Chauvot, 2003).

As part of our large-scale research and design project, Integrating Mathematics and Pedagogy (IMAP), we needed to measure beliefs, about mathematics and children's learning of mathematics, held by large groups of PSTs. Because we were interested in studying changes in beliefs of large numbers of PSTs, case-study methods were neither manageable nor appropriate. Our concern about limitations of Likert-scale surveys caused us to seek a different approach. We subsequently designed an open-ended, computerized survey that provided various contexts to assess PSTs' beliefs. Using piloted data, we then developed rubrics to assign numerical codes to the responses. One advantage of using this kind of instrument 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. We offer this 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 perfect the process, but have a clear idea of the intensity of this process before embarking on it. A paucity of scholarship about developing rubrics to be used for research purposes is available; 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).

RUBRIC DEVELOPMENT

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 items that assessed seven beliefs. Because of inadequate reliability measures on 2 of the rubrics, we eliminated them, along with the item with which they were associated, from our final data analysis.

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 8th and 10th graders' writing mechanics. The criteria for the highest score are "There are few or no minor errors. There are no major errors." 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 scenarios to which respondents reacted, either in the role of

the teacher or in commenting on teaching scenarios, we needed to develop *task-specific* rubrics, specific to a particular segment about a particular belief (Moskal, 2000).

Two research teams met two to four times per week in 3–4-hour sessions for 6 months. During this rubric-development phase, the 2–4-member teams each had two members who formed a consistent core, to provide grounding to the team and expertise with development. To get a wide range of responses for each item, we began with data gathered 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 adopted a grounded-theory approach (Glaser & Strauss, 1967), using pilot data, to develop each rubric. To begin the process, each person on the team independently analyzed the entire set of responses on a particular item with a particular belief in mind and sorted them 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 of 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 of the belief but had greater difficulty coming to consensus on responses that provided only partial evidence of the belief or, as was sometimes the case, responses that provided evidence of the belief in response to one part of an item only to provide disconfirming evidence of the belief for another part of the same item. 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 another student's (Sarah's) compensating strategy. 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.

During these discussions, descriptions of grouped responses emerged. Quite often, the group agreed on which responses belonged together 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 both robust, describing aspects of the belief that the responses provided,

and 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 three 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 provided the glue that held the category together with regard to the belief in question.

Once we had reached consensus on a rubric's categories and descriptions, we re-analyzed the data to check for interrater and intrarater reliability. 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 interrater 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 claimed 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, manipulatives, and drawings support children's initial thinking whereas symbols often do not.

We used responses to a survey segment about fractions to infer the respondents' support (or lack thereof) for Belief 6 (see Figure 1 for Segments 8.1 and 8.2). The greatest challenge in developing this rubric was to appropriately describe each of the three categories, particularly the middle category. We struggled to describe responses like the following, which we had placed in the middle category:

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.

In an early version of the rubric, we described it 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 some explanations for future coders; in cases similar to the example above, the term *weak* did not capture the reasons that we determined that the response was a middle score. In another draft focused on the respondent's claims about Item c, we stated, "Says Item d is easier but tends to think that Item 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 because the focus had shifted from the aspects that the respondent provided about the belief to aspects that the respondent did not provide. The final version (see Table 1) focused on specific ways that the respondent provided some evidence of the belief. We felt that the final version was more concrete than the earlier version and was more focused on the evidence that the respondent provided with respect to the belief.

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 X 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 there are five guests who equally share a very large chocolate bar for dessert. You attend a different birthday party at which there are eight guests who 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 did you rank as easier for children?

___c is easier than d __d is easier than c ___both items are equally difficult

Please explain your answer.

Figure 1. Segment 8.1 and 8.2.

We successively 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 the extent to which this respondent holds this 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 (external to the project) achieved 87.5% interrater reliability (the target for interrater reliability is typically set at 80%). The mean interrater reliability for all 17 rubrics was 84% for the responses of our 159 participants who completed this survey before and after taking part in one of five treatments; 20% of the responses were coded by two coders, and all responses were blinded.

Table 1. *Final Rubric for Belief 6, Segments 8.1 and 8.2*

Score	Description	
0	• 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 OR	• Says that Item d is easier BUT gives either inconsistent explanations (that is, explanations that indicate that they think c might be easier) or a clear focus in 8.2 on the teacher's role in showing students how to do the problem[a]
1	• Says that Item c is easier BUT expresses great appreciation for real-world context OR	• Says that Item d is easier (or that c and d are equally difficult) AND expresses some appreciation for real-world context or says that Item d is easier to visualize
2	• Says that Item d is easier (or that 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)	

Note. When the responses in 8.1c or 8.1d appear to conflict with the response in 8.2, give MORE weight to the response in 8.2.

[a]Our perspective is that a respondent who writes, "I would teach the children how to think about this idea," does not hold the belief that children think about mathematics differently from the way adults might expect, because, from this respondent's perspective, children need to be shown how to think about the mathematics.

CONCLUSIONS

Our definition of *belief* guided the development of our instrument by causing us to operationalize four components of beliefs in our rubrics. Our items are situated within contexts. We assume that people hold these beliefs at varying levels, and we

infer these levels either by observing to what respondents attend in contexts or by placing respondents in positions to act and inferring the extent to which they hold a belief by their purported actions within these contexts. We have provided the reader a specific example of this process.

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 your constructs, because these constructs serve as the foundation to which you must return when you attempt to operationalize these constructs.
- 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—at least three per team is highly recommended. Also, having two coding teams allowed us to have others who were experienced coders but not familiar with the particulars of a rubric code the data.
- 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.
- Accept that when dealing with the written word, some responses will be challenging to code. From any 100 responses, we typically found that 5–10 could fall into one of two categories, either because of differences in interpretation of the response or because an insufficient amount of information was provided by the respondent.

References

Ambrose, R. C., Philipp, R. A, Clement, L. L., & Chauvot, J. (2003). *A web-based survey to assess prospective elementary school teachers' beliefs about mathematics and mathematics learning: An alternative to Likert scales*. Manuscript submitted for publication.

Bright, G. W., & Vacc, N. N. (April, 1994). *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.

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.

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

Gronlund, N. E. (2003). *Assessment of Student Achievement*, Boston, MA.

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.

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.