

An Investigation of How Teachers Score Constructed-Response Mathematics Assessment Tasks

Ning Wang, nwang@mail.widener.edu, Widener University
Jinfa Cai, University of Delaware

This study identified some factors associated with teachers' knowledge and beliefs that are related to scoring mathematics constructed-response (CR) assessment tasks. Five groups of teachers (n = 274) who either had different teaching experiences or had different cultural beliefs about teaching and learning mathematics were selected to score 28 students' responses to seven CR math tasks. Among the 274 teachers, the first four groups (n=222) were selected from China. Group 1 was composed of pre-service elementary school teachers; group 2, pre-service secondary teachers; group 3, elementary in-service teachers; and group 4, secondary in-service teachers. The fifth group (n=52) was composed of in-service middle school teachers from the United States. A number of analyses of variance (ANOVA) on teachers' scores of the 28 responses and subsets of the responses were conducted to examine the mean differences across different groups of teachers. Four factors were found to have significant relationships with the rating differences. They were teaching experience, experience with students at particular grade levels, the nature of students' responses, and beliefs in teaching and learning mathematics. The identification of the factors has implications both for promoting validity of test scores and for examining teachers' understanding of student learning targets.

Keywords: cross-culture study, mathematics assessment, performance tasks, scoring constructed-response questions

Purpose of the Study

To ensure the reliability and validity for scores from constructed-response (CR) assessment tasks, great efforts have been made in minimizing raters' effects on scoring those assessment tasks. So far, research has mainly focused on training raters to address the concerns of rating consistency and objectivity (Fitzpatrick, Ercikan, Yen, & Ferrara, 1998; Mashburn & Henry, 2004; Moon & Hughes, 2002; Schafer, Swanson, Bene, & Newberry, 2001). However, little attention has been given to investigating what factors may have influenced raters' scoring of students' responses to CR assessment tasks. The purpose of this study is to identify some factors associated with teachers' beliefs about teaching and learning mathematics, as well as their knowledge and experience with students, that may influence their scoring of student responses to CR assessment tasks in mathematics. As defined by the National Council of Teachers of Mathematics (NCTM) and other researchers, teachers' beliefs in teaching and learning mathematics are the philosophy and views that teachers hold about the nature of mathematics and the nature of doing mathematics (Maaß & Schlöglmann, 2009; NCTM, 2017; Smith, 2014).

Literature Review and Research Questions

Classroom assessments often serve as diagnostic tools to inform teachers about what their students know, understand, and think (Balck, 2015; Brookhart & Nitko, 2015; [DeLuca](#), [Valiquette](#), [Coombs](#), [LaPointe-McEwan](#), & [Luhanga](#), 2016). Because CR tasks allow students to produce their own answers, to display the processes used to obtain their answers, to explain the thinking or reasoning associated with their answers, and to exhibit alternative approaches to problem-solving, students' thinking can become more visible and easier for teachers to grasp (Moyer, Cai, Nie, & Wang, 2011; Ni, Cai, & Zhou, 2014). As a result, using these assessment tasks in classroom assessments and scoring student written responses to such tasks can provide

teachers with great opportunities in gaining insightful information about students' thinking, understanding, and developmental status (Bennett, 2011; Kane, Crooks, & Cohen, 1999).

Standards developed by many states and professional organizations also direct teachers to use CR assessment tasks on a daily basis to assist in classroom instruction and student learning. For example, *Curriculum and Evaluation Standards for School Mathematics* developed by the National Council of Teachers of Mathematics (NCTM) (1989) heavily focuses on the development of students' mathematical proficiencies in problem-solving, communication of mathematical ideas, critical reasoning, and connections to ideas and procedures, both within mathematics and from other content areas. Such complex learning targets require the use of complex assessments, including CR assessment tasks, to actively engage students in the mastery of complex skills (Arter, 1999).

Effective teaching and learning are directly affected by teachers' formative and summative assessment decisions in the classroom (Bennett, 2011; Looney, Cumming, [van Der Kleij](#), & Harris, 2017; McMillian, 2003). Therefore, it is important to know whether those decisions validly and reliably reflect the teaching standards and learning targets, and whether teachers recognize the valid assessment standards by reflection on their values and by engagement in a shared development of assessments. However, researchers claim that in scoring student responses to CR assessment tasks, teachers may differ in their judgment of the quality of a student's response (Bennett, 2011; Klein, Stecher, Shavelson, McCaffrey, Ormseth, Bell, Comfort, & Othman, 1998). For example, it has been constantly reported that teachers' attention to score validity could be undermined by the external test regimes, or even individual assessment policies that reflect their own individualized values and beliefs about teaching (Black, 2015; Black, Harrison, Hodgen, Marshall, & Serret, 2010; Chen & Bonner, 2016).

Research has been conducted concerning teachers' interpretations and uses of students' classroom assessments results. It has been found that teachers do not always follow recommended grading practices; instead, the constructs teachers use in grading are highly influenced by their beliefs and value judgments (Brookhart, 1993, 2004, 2009; McMillian, 2003). A "hodgepodge" of subjective and objective factors could be used when teachers assessed and graded students (Chen & Bonner, 2016; McMillian, Myran, & Workman, 2002; Sun & Cheng, 2014). Research also indicated that there were grade-level differences in teachers' grading practices (e.g., elementary school teachers assigned higher grades than did their middle school counterparts) (Randall & Engelhard, 2009). Even when teachers have extensive training in analyzing student responses using the same scoring criteria, the inter-rater agreement between two teachers is not as high as would be preferable (Klein et al., 1998; Lane, Stone, Ankenmann, & Liu, 1994). The situation could be worse in classroom assessments since teachers do not always have written scoring criteria available to guide their analyses of student responses to CR assessment tasks. Instead, what they rely on are invisible criteria written in their mind, and these invisible criteria might be influenced by a number of factors.

Given the fact that teachers' scoring of student responses have a substantial impact on the enhancement of student learning process and on the improvement of student achievement, it is essential to know in what ways teachers think of student performance on the assessment tasks, how teachers score students' responses to the tasks, how teachers interpret the assessment results, and whether teachers' scoring aligns with the learning targets. Answers to these questions not only provide valuable information about effective instruction, but also provide evidence for validating test score interpretations. In particular, this study attempts to answer the following research questions:

- 1) Do teachers' beliefs in teaching and learning mathematics affect their scoring?
- 2) Do teachers' familiarity and experience with students and student learning of mathematics affect their scoring?
- 3) Does the nature of students' responses affect teachers' scoring?

Methods

Participants

Cross-cultural studies have revealed differences between U.S. and Chinese students' mathematical thinking and reasoning in their problem-solving process (Cai, 1995, 2000; Cai & Hwang, 2002; Chen & Bonner, 2016; Singer, Ellerton, & Cai, 2015). They have also found that the differences between the students in the two nations were likely due to teachers' differential beliefs in teaching and learning mathematics (Cai, 2014). As research indicates, teachers draw upon their cultural beliefs as a normative framework of values to guide their teaching (Bruner, 1996; Chen & Bonner, 2016). To explore whether teachers' beliefs in teaching and learning mathematics would affect their scoring (i.e., to answer the first research question of this study), teachers from both the U.S. and China were selected to participate in this study.

To explore whether teachers' familiarity and experience with students and student learning of mathematics would affect their scoring (i.e., to answer the second research question of this study), teachers with different teaching experience (pre-service versus in-service teachers) as well as teaching at different grade levels (elementary versus secondary) were selected for this study. Therefore, as shown in Table 1, five groups of teachers were selected to participate in the study. The first four groups of teachers were from a city in southwest China, including 53 pre-service elementary mathematics teachers (group 1), 60 pre-service secondary mathematics teachers (group 2), 59 in-service elementary school mathematics teachers of fourth, fifth, and

sixth grade mathematics (group 3), and 50 in-service secondary school mathematics teachers (group 4). The fifth group of teachers included 52 in-service middle school mathematics teachers of sixth, seventh, and eighth grade mathematics from Delaware, North Carolina, Pennsylvania, and Wisconsin in the United States (group 5). The members of group 5 were chosen because they teach math contents similar to those of the Chinese in-service elementary school teachers (group 3). In total, 274 teachers participated in this study.

The average number of years of teaching experience was 21 years for the teachers in group 3, 17 years for the teachers in group 4, and 19 years for the teachers in group 5. The two groups of pre-service teachers (groups 1 and 2) were in their senior year of college when the data were collected.

Table 1.

Study Participants

	<u>Group 1</u>	<u>Group 2</u>	<u>Group 3</u>	<u>Group 4</u>	<u>Group 5</u>
Sample size ($N = 274$)	53	60	59	50	52
Teaching status	Pre-service math teachers	Pre-service math teachers	In-service math teachers	In-service math teachers	In-service math teachers
Teaching experience	Senior year of college	Senior year of college	21 years teaching (in average)	17 years teaching (in average)	19 years teaching (in average)
School level	Elementary school	Secondary school	Elementary school	Secondary school	Middle school
Country	China	China	China	China	U.S.A.

Instrumentation

To examine whether the nature of students' responses would affect teachers' scoring (to answer the third research question of the study), a set of 28 student responses to seven CR mathematics assessment tasks were selected. The CR assessment tasks used in this study were

developed by various research projects (Cai, 2000; Lane, 1993). The student responses were selected from Cai (2000). Each of the seven CR tasks and 28 student responses are described in the Appendix of this study. These tasks were embedded in various content areas, covered in the Chinese fourth, fifth, and sixth grade math curriculums and in the U.S. sixth, seventh, and eighth grade math curriculums. As a result, Chinese elementary math teachers (teaching fourth, fifth, and sixth grade math) and U.S. middle school math teachers (teaching sixth, seventh, and eighth grade math) were selected for the study. Each of the student responses had a correct answer (or a reasonable estimate for the answer) and an appropriate strategy that yielded the correct answer (or estimate), but representations and solution strategies in the responses were different.

Although the 28 student responses were selected from actual sixth grade students' work, both the Chinese and English versions of these responses were re-written clearly by a math educator to avoid possible biases and misinterpretations. The presentations of students' work and explanations were identical to all groups of teachers except that there were in Chinese or English as appropriate.

Data Collection

Data collection consisted of two phases: in the first phase, each of the 274 teachers was asked to score the 28 student responses using a general 5-point scoring rubric (0-4):

4 points - correct and complete understanding

3 points - correct and complete, except for a minor error, omission, or ambiguity

2 points - partial understanding of the problem or related concept

1 point - a limited understanding of the problem or related concept

0 point - no understanding of the problem or related concept

The purpose of using the general and brief scoring rubric, rather than a specific and analytic scoring rubric for each task, was based on the consideration that this would allow for a better examination about whether teachers' familiarity, experience, and beliefs are related to their scoring. After they completed their initial scoring, nine Chinese in-service elementary teachers from group 3 and eleven U.S. in-service middle school teachers from group 5 were selected for the second phase. Each of these teachers was interviewed and asked to explain the reasons for his or her scoring of each of the 28 responses. All interviews were videotaped and transcribed. In both data collection phases, teachers were informed that the responses were from sixth grade students.

Results

Results from Two Groups of In-Service Teachers (Groups 3 and 5)

Table 2 provides mean ratings for each of the 28 responses and overall mean ratings across the 28 responses for teacher group 3 (in-service Chinese elementary teachers) and group 5 (in-service U.S. middle school teachers). The analysis of variance (ANOVA) shows that the overall mean ratings across the 28 responses for the U.S. teachers ($mean = 3.358, SD = .293$) is significantly higher than that for the Chinese teachers ($mean = 3.052, SD = .519$), $F(1, 109) = 15.063, p < .001$. An additional twenty-eight ANOVAs were conducted to compare mean rating differences between the two groups of teachers for each of the 28 responses. Significant differences are shown between the two groups of ratings for 8 out of the 28 responses at the significant level equal to or lower than 0.001. Also, significant difference is shown for one response at the level of .005, and for another response at the level of 0.01.

Table 2.

Mean Ratings and Associated Standard Deviations (In Parenthesis) from Teacher Groups 3 and 5 for Each of the 28 Students' Responses and Across the 28 Responses

<u>Response</u>	Chinese In-Service Elementary Teachers (n=59)*	U.S. In-Service Middle School Teachers (n=52)*	<u>p-value</u>
1	3.53 (0.728)	3.44 (0.826)	0.574
2**	3.02 (1.075)	3.69 (0.544)	< 0.001
3	3.25 (0.883)	3.4 (0.774)	0.347
4**	3.36 (1.047)	3.69 (0.506)	0.037
5**	2.15 (1.412)	3.38 (0.932)	< 0.001
6	3.63 (0.763)	3.81 (0.487)	0.146
7	3.64 (0.663)	3.46 (0.779)	0.185
8	1.71 (1.115)	2.16 (0.886)	0.031
9	2.93 (0.868)	3.23 (0.757)	0.058
10	3.80 (0.406)	3.79 (0.457)	0.921
11	1.88 (1.100)	2.34 (0.978)	0.034
12	3.73 (0.639)	3.62 (0.718)	0.38
13	2.54 (1.023)	3.12 (0.784)	0.001
14	2.97 (1.017)	3.62 (0.631)	< 0.001
15**	2.39 (1.260)	3.62 (0.565)	< 0.001
16	3.95 (0.222)	3.92 (0.269)	0.577
17	2.78 (1.100)	3.48 (0.700)	< 0.001
18**	3.10 (1.109)	3.69 (0.579)	0.001
19	2.58 (0.875)	2.92 (0.882)	0.04
20	3.73 (0.715)	3.77 (0.469)	0.729
21**	3.17 (1.020)	3.85 (0.364)	< 0.001
22	3.73 (0.739)	3.63 (0.595)	0.465
23	3.88 (0.326)	3.77 (0.469)	0.143
24	2.78 (1.001)	2.96 (1.028)	0.348
25	1.92 (1.368)	2.43 (0.962)	0.039
26**	3.32 (1.025)	3.79 (0.498)	0.004
27	2.97 (1.414)	2.23 (1.604)	0.012
28	3.02 (1.196)	3.21 (0.848)	0.331
Overall**	3.052 (0.519)	3.358 (0.293)	<.001

*A number in parentheses indicates the standard deviation associated with each mean.

**The difference of the standard deviations between the two group means for the task is significant at the level equal to or less than .001.

An examination of the differences between each of the 28 individual responses indicates that the nature of the student responses influenced the teachers' scoring. Table 3 summarizes the average mean ratings for four types of students' responses that show significant differences between the two groups of teachers. These types are the eight responses that involve visual drawings or concrete solution strategies, $F(1, 109) = 34.608, p < .001$, Response 15 that consists of mathematics errors, $F(1, 109) = 41.767, p < .001$, Response 18 that uses a guess-and-check solution strategy, $F(1, 109) = 11.879, p = .001$, and Response 27 that allows for multiple correct answers, $F(1, 109) = 6.591, p = .012$. Among all of the 28 responses, Response 15 accounts for the largest difference in the ratings between the two groups of teachers. For the first three types of responses (i.e., except Response 27), the U.S. teachers provide significantly higher ratings than Chinese teachers. Response 27 is the response for which the Chinese teachers score higher than the U.S. teachers. The mean difference of 0.74 for Response 27 is on the borderline of being statistically significant ($p = .012$) and is the largest one as compared to other responses for which the Chinese teachers score higher. Also, the U.S. teachers have the highest variation ($SD = 1.604$) in scoring this response as compared to the Chinese teachers ($SD = 1.414$ which is relatively high too) and their ratings on the other responses. Further analysis of teachers' scoring reveals that Response 27 involves a task that allows for multiple correct answers. The approach used in Response 27 is not as common as the other two solutions as used in Responses 26 and 28. As a result, low scores (either 0 or 1 point) are given to the response by a number of teachers, particularly by nearly one third of the U.S. teachers, which results in a large score variation as well.

Table 3.

Mean Ratings and Associated Standard Deviations for Four Types of Students' Responses from Teacher Groups 3 and 5

<u>Response Type</u>	<u>Responses</u>	<u>Mean Ratings for Chinese In-Service Elementary Teachers (n=59)*</u>	<u>Mean Ratings for U.S. In-Service Middle School Teachers (n=52)*</u>
Responses Involving Visual			
Drawings or Concrete Solution Strategies	2, 4, 5, 13, 14, 17, 21, & 26	2.913 (0.763)	3.577 (0.301)
Responses Involving Mathematics Errors	15	2.39 (1.260)	3.62 (0.565)
Responses Using Guess-and-Check Solution Strategy	18	3.10 (1.109)	3.69 (0.579)
Responses Involving Multiple Correct Answers	27	2.97 (1.414)	2.23 (1.604)

*A number in parentheses indicates the standard deviation associated with each mean.

The analysis from the interview data indicates that not only are the nature of students' solution strategies related to teachers' ratings, but also that the differences in teachers' beliefs in teaching and learning mathematics affect teachers' scoring of students' responses.

It is found that Chinese teachers consistently take the nature of the solution strategies into account in their scoring. If a response involves a drawing or making a list, the Chinese teachers usually give a relatively lower score even though the strategy is appropriate with the correct answer. The general reason Chinese teachers give for their lower scores to the responses with visual or concrete approaches is that the strategy does not find regularities. Although most U.S. teachers realize that drawing is not a sophisticated—yet very time-consuming—strategy, they

also comment that the drawing in some cases is a viable approach producing a correct answer. Moreover, almost all U.S. teachers state that these visual strategies clearly show how students think and solve the problems, while Chinese teachers seem to have a clear goal: students should learn more generalized strategies.

The errors in Response 15 are related to the written communication of the student's thinking. The process reflects the chain of thoughts that a student used to solve the problem. Although the result is correct, the mathematical expression consists of errors. Interviews with the teachers reveal that Chinese teachers tend to be more stringent in scoring students' responses like this, and that they expect students' written expression to be mathematically appropriate. They believe that the use of appropriate expressions in mathematics can help students develop their mathematical proficiencies and logical thinking. Meanwhile, this type of error does not affect U.S. teachers' ratings as much. They tend to be more tolerant and allow students to write what they think without paying much attention to students' written expression.

The other noticeable difference between Chinese and U.S. teachers is that the scoring variation among the Chinese teachers is considerable larger than the U.S. teachers, in addition to the mean score differences. As indicated in Table 3, for the first three types of responses, the standard deviations of the ratings from the Chinese teachers are all larger than that from the U.S. teachers. This is also seen in Table 2 for the individual responses; there, Levene's test of homogeneity of variance indicates that the standard deviations for Chinese teachers' ratings are significantly larger than that for the U.S. teachers for 7 responses and for the overall means at the level equal to or less than .001. This indicates a higher scoring consistency among U.S. teachers than Chinese teachers.

By examining the other types of responses, the results show that, for responses using conventional solution strategies, both U.S. and Chinese teachers score them similarly. For example, for the 10 responses involving algebraic or arithmetic approaches (i.e., Responses 6, 7, 9, 10, 12, 16, 20, 22, 23, and 28), the averaged mean scores across the 10 responses are 3.60 ($SD = .356$) for the Chinese teachers and 3.62 ($SD = .383$) for the U.S. teachers. In particular, both U. S. and Chinese teachers highly value the responses using algebraic approaches (Responses 10, 16, and 20) and consistently provide high scores on the responses.

Results from Four Groups of Chinese Teachers (Groups 1, 2, 3, and 4)

Table 4 provides mean ratings for each of the 28 student responses, as well as across the 28 responses for the four groups of Chinese teachers. The analysis of variance (ANOVA) across the four groups of teachers shows that there is a significant difference in their overall mean ratings across the 28 responses, $F(3, 218) = 7.091, p < .001$. A post hoc comparison indicates that significant differences exist between the following pairs of the teacher groups:

In-service elementary	vs.	Pre-service elementary ($p = .007$)
In-service elementary	vs.	Pre-service secondary ($p < .001$)
In-service elementary	vs.	In-service secondary ($p = .008$)

This result reveals a contrast between the teachers who have the same cultural beliefs in teaching and learning mathematics but with different levels of familiarity and knowledge about the students being assessed and the mathematics content being assessed. For the in-service teachers, one group teaches students at the level being assessed in this study (group 3) and the other group teaches at a different level (group 4). The other two groups are pre-service teachers and do not have as rich teaching experience as the other two groups.

Table 4.

Mean Ratings and Associated Standard Deviations (In Parenthesis) from the Four Groups of Chinese Teachers for Each of the 28 Students' Responses and Across the 28 Responses

<u>Response</u>	Pre-Elementary Teachers (<i>n</i> =53)*	Pre-Secondary Teachers (<i>n</i> =60)**	Elementary Teachers (<i>n</i> =59)**	Secondary Teachers (<i>n</i> =50)**	<u>Significant Difference*</u>
1	2.96 (0.759)	3.30 (0.743)	3.53 (0.728)	3.22 (0.840)	ET > PE (<i>p</i> =.002)
2	3.00 (0.832)	2.68 (0.948)	3.02 (1.075)	2.90 (0.863)	
3	2.91(0.883)	2.58 (1.293)	3.25 (0.883)	2.92 (0.752)	ET > PS (<i>p</i> =.002)
4	2.68 (1.015)	3.37 (0.843)	3.36 (1.047)	3.00 (1.050)	ET > PE (<i>p</i> =.002) PS > PE (<i>p</i> =.002)
5	2.26 (1.195)	2.07 (1.326)	2.15 (1.412)	2.12 (1.189)	
6	3.30 (0.890)	3.73 (0.607)	3.63 (0.763)	3.60 (0.700)	
7	3.08 (0.997)	2.92 (0.787)	3.64 (0.663)	3.16 (0.934)	ET > PE (<i>p</i> =.003) ET > PS (<i>p</i> <.001)
8	1.62 (1.042)	1.17 (0.905)	1.71 (1.115)	1.32 (1.168)	
9	2.64 (1.194)	2.30 (1.306)	2.93 (0.868)	2.38 (1.141)	
10	3.68 (0.613)	3.58 (0.944)	3.80 (0.406)	3.78 (0.418)	
11	1.43 (1.047)	1.17 (1.196)	1.88 (1.100)	1.18 (1.044)	ET > PS (<i>p</i> =.003) ET > ST (<i>p</i> =.007)
12	3.30 (0.822)	3.47 (0.676)	3.73 (0.639)	3.76 (0.476)	ET > PE (<i>p</i> =.005) ST > PE (<i>p</i> =.004)
13	2.79 (0.817)	2.58 (0.962)	2.54 (1.023)	2.56 (1.033)	
14	2.72 (1.063)	2.28 (1.059)	2.97 (1.017)	2.68 (0.957)	ET > PS (<i>p</i> =.002)
15	2.38 (1.213)	1.88 (1.303)	2.39 (1.260)	1.70 (1.313)	
16	3.68 (0.872)	3.65 (0.606)	3.95 (0.222)	3.92 (0.274)	
17	2.47 (0.973)	2.60 (1.061)	2.78 (1.100)	3.26 (0.664)	ST > PE (<i>p</i> <.001) ST > PS (<i>p</i> =.003)
18	2.77 (1.368)	2.99 (0.846)	3.10 (1.109)	3.00 (0.881)	
19	1.87 (1.093)	1.90 (0.986)	2.58 (0.875)	2.32 (0.935)	ET > PE (<i>p</i> =.001) ET > PS (<i>p</i> <.001)
20	3.66 (0.618)	3.70 (0.743)	3.73 (0.715)	3.48 (0.863)	
21	3.42 (0.663)	3.33 (0.681)	3.17 (1.020)	2.90 (1.015)	
22	3.09 (1.061)	3.47 (0.982)	3.73 (0.739)	3.62 (0.530)	ET > PE (<i>p</i> =.001)
23	3.60 (0.689)	3.62 (0.715)	3.88 (0.326)	3.78 (0.545)	
24	2.79 (1.199)	2.80 (0.988)	2.78 (1.001)	2.30 (1.093)	
25	2.11 (1.050)	1.88 (1.180)	1.92 (1.368)	1.50 (0.974)	
26	3.21 (0.793)	2.53 (0.911)	3.32 (1.025)	2.64 (0.921)	ET > PS (<i>p</i> <.001) ET > ST (<i>p</i> =.001) PE > PS (<i>p</i> =.001)
27	1.81 (1.429)	2.25 (1.068)	2.97 (1.414)	2.78 (1.282)	ET > PE (<i>p</i> <.001) ST > PE (<i>p</i> =.001)
28	3.04 (0.784)	2.65 (1.022)	3.02 (1.196)	2.52 (1.035)	
Overall	2.80 (0.350)	2.73 (0.325)	3.05 (0.519)	2.80 (0.407)	ET > PE (<i>p</i> =.007) ET > PS (<i>p</i> <.001) ET > ST (<i>p</i> =.008)

*PE—Pre-elementary teachers; PS—Pre-secondary teachers; ET—Elementary teachers; & ST—Secondary teachers.

**A number in parentheses indicates the standard deviation associated with each mean.

First, it is interesting to find that all of the significant differences exist between the in-service elementary teachers (i.e., the targeted assessment level in this study) and each of the other three groups. There is no significant difference in the mean ratings between any pairs of the other three groups of teachers. Also, in-service elementary teachers rate students' responses more leniently and provide higher scores than the remaining three groups of teachers do. Because the CR assessment tasks used in this study and the students' responses selected for scoring are within the in-service elementary teachers' teaching level, this result clearly indicates that teachers' familiarity with students, their expectations of student performance, and their knowledge about students' understanding and ability have influenced their scoring of students' responses. This result indicates that raters' experience with students and student learning of mathematics is one factor that influences the scoring of the CR assessment tasks.

The influence of this factor can be found on the individual responses as well. Not only do the significant differences exist on the overall mean ratings of the 28 responses between the groups of teachers, but also similar differences are found on many individual responses. Of the 28 responses, 12 responses show significant differences in the mean ratings among the different groups of teachers, while the in-service elementary teachers have significantly higher ratings than do at least one of the other three groups on 11 of the responses.

Second, it is even more interesting to find that there is a significant difference between in-service elementary and in-service secondary teachers on the overall mean ratings, but no significant difference between in-service secondary teachers and any other groups of pre-service teachers. The in-service secondary teachers provide ratings similar to both pre-service

elementary and secondary teachers. In fact, at the individual response level, the in-service secondary teachers provide mean ratings similar to both pre-service elementary and secondary teachers on 22 of the 28 responses. This result suggests that teaching experience alone does not necessarily distinguish raters from others in scoring students' responses to CR assessment tasks. Although in-service secondary math teachers have extensive teaching experience, such experience does not guarantee their familiarity with and knowledge about teaching and learning mathematics for the grade level of the students being assessed. Also, their expectations of student learning of mathematics may differ. As a result, this group of in-service teachers scored student responses similarly as both pre-service elementary and secondary teachers. This result indicates that not only raters' experience with students, but also raters' experience at different grade levels and their expectations of student mathematics learning can be factors that influence the scoring of the CR assessment tasks.

As for the nature of the students' responses, unlike the findings from the Chinese and U.S. teachers, no systematic difference of teachers' ratings is found among the four groups of Chinese teachers. For example, for the 8 responses that involve visual drawings or concrete solution strategies, the mean scores for groups 1, 2, 3, and 4 are 2.818 ($SD = .483$), 2.681 ($SD = .454$), 2.913 ($SD = .763$), and 2.758 ($SD = .571$), respectively, which does not indicate a statistically significant difference, $F(3, 218) = 1.668, p = .175$. Similarly, Table 4 indicates that the four groups of teachers provide consistently high scores for the responses using algebraic approaches (Response 10, 16, and 20). By examining the 11 responses that show significant differences between the in-service elementary teachers and at least one of the other three groups, no apparent pattern is found with respect to types of student responses. However, it is interesting that, on average, the scoring variation among the in-service elementary teachers ($SD = .519$) is

larger than the other three groups ($SD = .350, .325, \text{ and } .407$ for groups 1, 2, and 4, respectively). Levene's test of homogeneity of variances shows a significant difference at the level less than .001.

Discussions

In order to enhance the validity and objectivity of test scores obtained from assessments that consist of CR tasks, minimizing rater effects has been a particular concern. This study identifies four factors associated with teachers' knowledge and beliefs that have affected scoring math CR assessment tasks. These factors are beliefs in teaching and learning mathematics, the nature of students' responses, teaching experience, and experience with students at particular grade levels.

Teachers' beliefs influence their scoring of various responses. In fact, the results of this study clearly show that overall, U.S. teachers are much more lenient than Chinese teachers in scoring student responses. However, the leniency is not reflected in their evaluation of students' responses involving conventional approaches, such as using algebraic equations and other mathematical expressions. Both U.S. and Chinese teachers consistently give higher scores for responses using algebraic approaches. In particular, almost all of the U.S. and Chinese teachers value the responses with algebraic approaches the highest when compared to other responses to the same math tasks.

U.S. teachers' leniency is reflected in their ratings of responses involving visual strategies. Chinese teachers consistently take the nature of the solution strategies into account in their scoring. If a response involves a visual or concrete strategy, Chinese teachers usually give a relatively lower score even though the strategy is appropriate for a correct answer. While U.S. teachers realize that the drawing strategy is not a sophisticated strategy and is very time

consuming, they recognize that drawing is a viable approach that produces correct answers. Therefore, a response with concrete drawing strategy should not be penalized. Chinese teachers seem to have a clear goal that students should learn more generalized strategies. Interviews of the teachers do not provide evidence that U.S. teachers have such a clear goal. Instead, the goal of U.S. teachers is to have students solve a problem no matter what strategies they use.

In addition, Chinese teachers seem to be much more concerned about details of the writing format and inclusion of units for answers than U.S. teachers. Chinese teachers believe that the use of an appropriate writing format and units in problem solving can help students develop their abilities to think logically. Such a requirement is also demanded in examinations.

Teachers' familiarity and experience with students and assessment content clearly influence their scoring of student responses. The mathematical content being assessed and the student responses selected in the study are within the grade level taught by the in-service elementary school teachers. This group of teachers provides significantly different scores from any of the other three groups of teachers while there are no significant overall differences among the other three groups of teachers. Because the in-service elementary school teachers are familiar with the grade level of students and the assessment content, one would naturally expect that this group of teachers should score much more consistently than the other three groups of teachers. However, the results show that on average, the in-service elementary school teachers have the highest scoring variation as compared to any of the other three groups. One plausible explanation is that this group of teachers' experience and familiarity with the students and the assessment content might prevent them from following the scoring guide strictly. Instead, they may use their own interpretations of the student's responses more than the teachers in any of the other three groups. As a result, their scoring may be highly influenced by their experience and

value judgment, as concluded by Brookhart (1993). In contrast, the other three groups of teachers may follow the scoring guide more closely in their scoring, resulting in the higher scoring consistency than the in-service elementary teachers.

This finding suggests that while it is important to choose teachers who are familiar with the content and the students being assessed for scoring student responses to CR tasks, rater training is also critical to ensure scoring consistency. It is also noted that in this study, we have only given teachers a general and brief scoring guide. The purpose of using the general and brief scoring guide is based on the consideration that this would allow for a better examination about whether teachers' familiarity, experience, and beliefs are related to their scoring. While using the general and brief scoring guide serves the purpose of this study well, we would suggest future research to explore whether the use of detailed and task-specific scoring guides would increase scoring consistency for raters who are less familiar and less experienced with the students and the mathematical contents being assessed. In addition, since teachers are more likely to use their beliefs and own interpretations in scoring student responses, it would be interesting for future research to explore effective training strategies on how to advise raters to avoid individualized interpretations of student responses and to follow scoring guides more closely.

It is the authors' hope that the identification of these factors in this study not only provides suggestions to the field of performance assessment in minimizing rater effects, assisting in rater training, and promoting scoring objectivity as well as validity of test scores, but also helps to examine teachers' understanding of student learning targets in mathematics and further promotes effective instruction.

References

- Arter, J. (1999). Teaching measures of performance. *Educational Measurement: Issues and Practice, 18*(2), 30-44.
- Bennett, R. E. (2011). Formative assessment: a critical review. *Assessment in Education: Principles, Policy & Practice, 18*(1), 5-25.
- Black, P. (2015). Formative assessment – an optimistic but incomplete vision. *Assessment in Education: Principles, Policy & Practice, 22*(1), 161-177.
- Black, P., Harrison, C., Hodgen, J., Marshall, B., & Serret, N. (2010). Validity in teachers' summative assessments. *Assessment in Education: Principles, Policy & Practice, 17*(2), 215-232.
- Brookhart, S. (1993). Teachers' grading practices: meaning and values. *Journal of Educational Measurement, 30*(2), 123-142.
- Brookhart, S. (2004). *Grading*. Upper Saddle River, NJ: Pearson.
- Brookhart, S. (2009). Teachers' grading: practice and theory. *Applied Measurement in Education, 7*(4), 279-301.
- Brookhart, S., & Nitko, A. J. (2015). *Educational assessment of students* (7th ed.). Upper Saddle River, NJ: Pearson Education.
- Bruner, J. (1996). *The Culture of Education*. Cambridge, MA: Harvard University Press.
- Cai, J. (1995). A cognitive analysis of U.S. and Chinese students' mathematical performance on tasks involving computation, simple problem solving, and complex problem solving. *Journal for Research in Mathematics Education Monograph*, i-151.
- Cai, J. (2000). Mathematical thinking involved in U.S. and Chinese students' solving process-constrained and process-open problems. *Mathematical Thinking and Learning, 2*, 309-340.

- Cai, J. (April, 2014). *Content representations and pedagogical moves in Chinese and U.S. mathematics classrooms: Lesson video analyses at elementary, middle, and high school levels*. Discussant at the annual meeting of American Educational Research Association, Philadelphia, PA.
- Cai, J., & Hwang, S. (2002). U.S. and Chinese students' generalized and generative thinking in mathematical problem solving and problem posing. *Journal of Mathematical Behavior*, 21(4), 401-421.
- Chen, P. P., & Bonner, S. M. (in press). Teachers' beliefs about grading practices and a constructivist approach to teaching. *Educational Assessment*. doi: 10.1080/10627197.2016.1271703.
- [DeLuca](#), C., [Valiquette](#), A., [Coombs](#), A., [LaPointe-McEwan](#), D., & [Luhanga](#), U. (in press). Teachers' approaches to classroom assessment: a large-scale survey. *Assessment in Education: Principles, Policy & Practice*. doi: [10.1080/0969594X.2016.1244514](#).
- Fitzpatrick, A., Ercikan, K., Yen, W., & Ferrara, S. (1998). The consistency between raters scoring in different test years. *Applied Measurement in Education*, 11(2), 195-208.
- Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice*, 18(2), 5-17.
- Klein, S. P., Stecher, B. M., Shavelson, R. J., McCaffrey, D., Ormseth, T., Bell, R. M., ... Othman, A. R. (1998). Analytic versus holistic scoring of science performance tasks. *Applied Measurement in Education*, 11(2), 121-137.
- Lane, S. (1993). The conceptual framework for the development of a mathematics assessment instrument for QUASAR. *Educational Measurement: Issues and Practice*, 12(2), 16-23.
- Lane, S., Stone, C. A., Ankenmann, R. D., & Liu, M. (1994). Reliability and validity of a

mathematics performance assessment. *International Journal of Educational Research*, 21(3), 247-266.

Looney, A., Cumming, J., [van Der Kleij](#), F., & Harris, K. (in press). Reconceptualising the role of teachers as assessors: teacher assessment identity. *Assessment in Education: Principles, Policy & Practice*. doi: [10.1080/0969594X.2016.1268090](https://doi.org/10.1080/0969594X.2016.1268090).

Maaß, J., & Schlöglmann, W. (2009). *Beliefs and Attitudes in Mathematics Education: New Research Results*. Rotterdam, Netherlands: Sense Publishers.

Mashburn, A. J., & Henry, G. (2004). Assessing school readiness: validity and bias in preschool and kindergarten. *Educational Measurement: Issues and Practice*, 23(4), 16-30.

McMillian, J. H. (2003). Understanding and improving teachers' classroom assessment decision making: Implications for theory and practice. *Educational Measurement: Issues and Practice*, 22(4), 34-43.

McMillian, J. H., Myran, S., & Workman, D. (2002). Elementary teachers' classroom assessment and grading practices. *The Journal of Educational Research*, 95(4), 203-213.

Moon, T., & Hughes, K. (2002). Training and scoring issues involved in large-scale writing assessments. *Educational Measurement: Issues and Practice*, 21(2), 15-19.

Moyer, J. C., Cai, J., Nie, B., & Wang, N. (2011). Impact of curriculum reform: Evidence of change in classroom instruction in the United States. *International Journal of Educational Research*, 50(2), 87-99.

National Council of Teachers of Mathematics (1989). *Curriculum and Evaluation Standards for School Mathematics*. Reston, VA: National Council of Teachers of Mathematics.

National Council of Teachers of Mathematics (2017). *Statement of Beliefs*. Retrieved January 30, 2017, from <https://www.nctm.org/About/At-a-Glance/Statement-of-Beliefs/> Reston,

VA: Author.

Ni, Y., Cai, J., & Zhou, D. (April, 2014). *Instructional tasks of high-cognitive demands improve affective attitudes toward mathematics in Chinese fifth grade classroom*. Paper presented at the annual meeting of American Educational Research Association, Philadelphia, PA.

Randall, J., & Engelhard, G. (2009). [Differences between teachers' grading practices in elementary and middle schools](#). *The Journal of Educational Research*, 102(3), 175-186.

Schafer, W. D., Swanson, G., Bene, N., & Newberry, G. (2001). Effects of teacher knowledge of rubrics on student achievement in four content areas. *Applied Measurement in Education*, 14(2) 151-170.

Singer, F. M., Ellerton, N., & Cai, J. (2015). Mathematical problem posing today: A cross-cultural view. In F.M. Singer, N. Ellerton, & J. Cai (Eds.), *Mathematical problem posing: From research to effective practice*. New York, NY: Springer.

Smith, K. (2014). *How teacher beliefs about mathematics affect student beliefs about mathematics* (Senior honors thesis). Retrieved from <https://scholars.unh.edu/honors/193/>

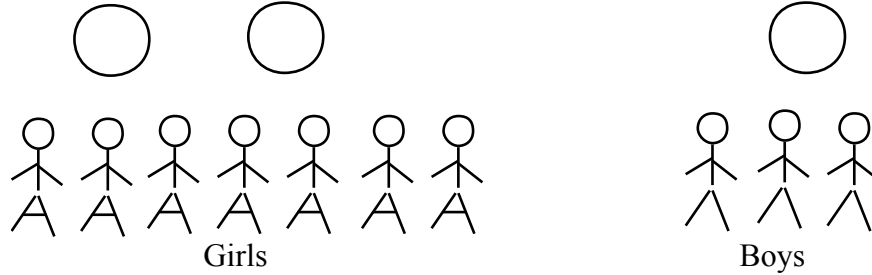
Sun, Y., & Cheng, L. (2014). Teachers' grading practices: meaning and values assigned. [Assessment in Education: Principles, Policy & Practice](#), 21(3), 326-343.

Appendix

Seven CR Assessment Tasks and Descriptions of Student Responses to the Assessment Tasks

Task 1: Pizza Ratio Problem

Here are some children and pizzas. 7 girls share 2 pizzas equally and 3 boys share 1 pizza equally.



- A. Does each girl get the same amount as each boy?
Explain or show how you found your answer.
- B. If each girl does not get the same amount as each boy, who gets more?
Explain or show how you found your answer.

Response 1: 7 Girls get two pizzas, and 3 boys get one pizza. The girls have twice as many pizza as boys. But the number of girls is more than twice as many than boys. So the boys get more.

Response 2: Three girls share one pizza and the remaining four share one pizza. Each piece that each of the remaining four girls get is smaller than the boys get. So the boys get more.



Response 3: $7/2 = 3.5$ and $3/1 = 3$. Therefore 3.5 girls will share one pizza and 3 boys will share one pizza. Thus, each boy gets more.

Response 4: Each pizza was cut into 4 pieces. Each girl gets 1 piece with 1 piece left over. Each boy gets 1 piece with 1 piece left over. The one piece left over must be shared by the 7 girls, but the 1 piece left over will be shared by three boys. So the boys get more.



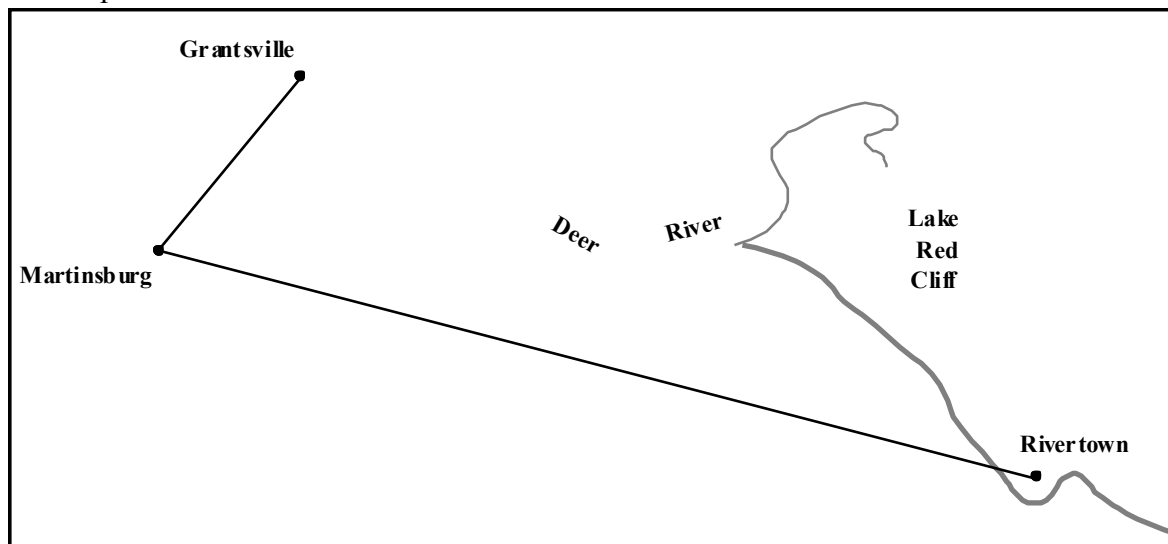
Response 5: Three girls share one pizza, and another three girls share another pizza. Each of these six girls will get the same amount of the pizza as each of the three boys. But one of the girls has no pizza. So, each boy will get more.



Response 6: Each boy will get $\frac{1}{3}$ of a pizza and each girl will get $\frac{2}{7}$ of a pizza. If you compared $\frac{1}{3}$ with $\frac{2}{7}$, you would know that $\frac{1}{3}$ is bigger than $\frac{2}{7}$ since $\frac{1}{3} = .33$ and $\frac{2}{7} = .29$. $.33 - .29 = .04$. Therefore, each boy gets more than each girl.

Task 2: Map Ratio Problem

The map below shows the locations of three cities.



The actual distance between Grantsville and Martinsburg is 54 miles. On the map, Grantsville and Martinsburg are 3 centimeters apart. On the map, Martinsburg and Rivertown are 12 centimeters apart.

What is the actual distance between Martinsburg and Rivertown?
Show how you found your answer.

Response 7: A student first found how many miles a centimeter on the map represents ($54/3=18$), then multiply by 12 to get the actual distance that the 12 centimeters on the map represents ($18 \times 12=216$ miles).

Response 8: A student used a finger to measure the distance between Martinsburg and Grantsville on the map, then used the measurement unit to measure the length between Martinsburg and Rivertown on the map and to find the number of the unit of the length. By multiplying the number of unit by 54, the student found the distance between Martinsburg and Rivertown.

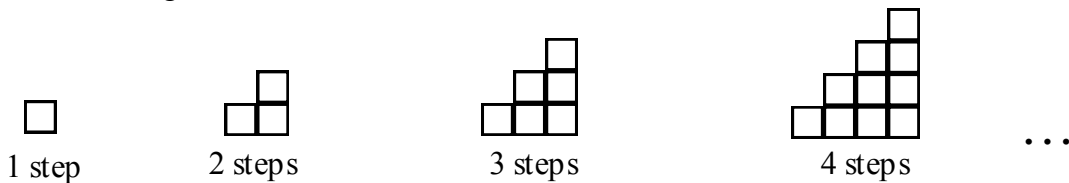
Response 9: A student first divided 12 by 3 and got 4. Since 3 centimeters represent 54 miles, the actual distance the 12 centimeters represents was 216 ($4 \times 54 = 216$ miles).

Response 10: A student set up a formal proportional relationship to find the actual distance (i.e., $3/12 = 54/x$, $x = 216$ miles).

Response 11: A student used a paper clip to mark or measure the distance between Martinsburg and Grantsville on the map, then used the measurement unit to measure the length between Martinsburg and Rivertown on the map and to find the number of the unit of the length. By dividing 54 by the number of the measurement unit between Martinsburg and Grantsville, the student found the number of actual miles per measurement unit. By multiplying the number of the measurement unit between Martinsburg and Rivertown by the number of actual miles per measurement unit, the student found the number of actual miles between Martinsburg and Rivertown.

Task 3: Block Pattern Problem

Look at the figures below.



- A. How many blocks are needed to build a staircase of 5 steps?
Explain how you found your answer.
- B. How many blocks are needed to build a staircase of 20 steps?
Explain how you found your answer.

Response 12: The student realized that the staircase of 5 steps is made of 1 block, 2 blocks, 3 blocks, 4 blocks, and 5 blocks. Therefore, the number of blocks for 5-step staircase is 15. The staircase of 20 steps is made of 1 block, 2 blocks, 3 blocks, ..., and 20 blocks. Therefore the number of blocks needed to build a staircase of 20 steps can be found by adding 1, 2, 3, 4, ..., and 20 (i.e., $1 + 2 + 3 + 4 + \dots + 20 = 210$).

Response 13: The student realized that the number of blocks in a 5-step staircase has five more blocks than the number of blocks in a 4-step staircase. Through a recursive process, the student found the number of blocks in a 19-step staircase, and by adding 20 to it finally found the number of blocks needed to build a staircase of 20 steps.

Response 14: The student correctly drew staircases of 5 and 20 steps on the paper and counted the number of blocks in them, which are 15 and 210, respectively (also see attached).

Response 15: The thinking process in Response O is similar to that in Response L. However, in this response, the student wrote $1 + 2 = 3 + 3 = 6 + 4 = 10 + 5 = 15$ to find the number of blocks needed for 5-step staircase. The student also wrote $1 + 2 = 3 + 3 = 6 + 4 = 10 + 5 = 15 + 6 = 21 + 7 = 28 + 8 = 36 + 9 = 45 + 10 = 55 + 11 = 66 + 12 = 78 + 13 = 91 + 14 = 105 + 15 = 120 + 16$

$= 136 + 17 = 153 + 18 = 171 + 19 = 190 + 20 = 210$ to find the number of blocks needed for 20-step staircase.

Task 4: Odd Number Pattern Problem

Sally is having a party.

The first time the doorbell rings, 1 guest enters.

The second time the doorbell rings, 3 guests enter.

The third time the doorbell rings, 5 guests enter.

The fourth time the doorbell rings, 7 guests enter.

Keep going in the same way. On the next ring a group enters that has 2 more persons than the group that entered on the previous ring.

- A. How many guests will enter on the 10th ring?
Explain or show how you found your answer.
- B. In the space below, write a rule or describe in words how to find the number of guests that entered on each ring.
- C. 99 guests entered on one of the rings. What ring was it?
Explain or show how you found your answer.

Responses 16: The student found the general rule $(2n - 1)$ to find the number of guests that entered on the n th ring. For answering question C of the problem, the student set $2n - 1 = 99$. Solve for n which is 50.




Response 17: The student said that the pattern goes by 2's. And then the student listed detailed tables to answer questions in the problem. In particular, to find the ring number when 99 guests entered, the student listed a table from the first ring to the ring number when 99 guests entered.

Response 18: The student found the general rule: number of guests = ring number + (ring number - 1). And then the student used guess-and-check strategy to correctly found the ring number when 99 guests entered.

Response 19: The student got all correct answers, but the descriptions are general. For example, "keep adding 2's until getting 99" is the description provided when the student correctly found the ring number when 99 guests entered.

Task 5: Hats Averaging Problem

Angela is selling hats for the Mathematics Club. This picture shows the number of hats Angela sold during the first three weeks.

Week 1	
Week 2	
Week 3	
Week 4	?

How many hats must Angela sell in Week 4 so that the average number of hats sold is 7? Show how you found your answer.

Responses 20: The student correctly used the average formula to solve the problem algebraically (e.g., $(9 + 3 + 6 + x) = 7 \times 4$, then solve for x).

Response 21: The student used "leveling-off processes" to solve the problem. The student viewed the average (7) as a leveling basis to "line up" the numbers of hats sold in the week 1, 2, and 3. Since 9 hats were sold in week 1, it has two extra hats. Since 3 hats were sold in week 2, 4 additional hats are needed in order to line up the average. Since 6 hats were sold in week 3, it needs 1 additional hat to line up the average. In order to line up the average number of hats sold over four weeks, 10 hats should be sold in week 4 (See attached).

Response 22: The student correctly used the average formula to solve the problem arithmetically (e.g., $7 \times 4 - (9 + 3 + 6) = 10$).

Task 6: Score Average Problem

The average of Ed's ten test scores is 87. The teacher throws out the top and bottom scores, which are 55 and 95.

What is the average of the remaining set of scores?
Show how you found your answer.

Response 23: $10 \times 87 = 870$. $870 - 55 - 95 = 720$. $720/8 = 90$. The average of the remaining 8 scores is 90.

Response 24: The student first used one of the properties of average and determined that the average for the remaining eight scores must be between 55 and 95. Then the student drew ten circles and put 95 in the first and 55 in the last, leaving eight empty circles. Using a modified sharing approach, the student realized that 55 and 95 contributed 15 to the average $[(95 + 55) \div 10 = 15]$. So the student said that each of the eight blank spaces should get 15. But 15 is 72 less than 87 (the average for the ten scores), the student then multiplied 10 by 72 and got 720. $720 \div 8 = 90$. Thus, 90 became the average of the remaining eight scores after the top and bottom scores were thrown away.

Response 25: I think that the average for the remaining set of scores is between 55 and 95. But 87 is closer to 95 than 55. So the average for the remaining must be about 90.

Task 7: Number Theory Problem

Yolanda was telling her brother Damian about what she did in math class.

Yolanda said, "Damian, I used blocks in my math class today. When I grouped the blocks in groups of 2, I had 1 block left over. When I grouped the blocks in groups of 3, I had 1 block left over. And when I grouped the blocks in groups of 4, I had 1 block left over."

Damian asked, "How many blocks did you have?"

What was Yolanda's answer to her brother's question?

Show how you found your answer.

Response 26: The student constructed three separate diagrams showing sets of blocks and attempted to make all the sets the same size. For example, the first set of blocks had 12 blocks, which were grouped in groups of 2 with 1 block left over; the second set had 12 blocks, which were grouped in groups of 3 with 1 block left over; and the third set had 12 blocks, which were grouped in groups of 4 with one left over.

Response 27: The student found 24 as a common multiple of 2, 3, and 4 by direct computation ($2 \times 3 \times 4 = 24$), and then added one to the common multiple.

Response 28: The student found 12 as a common multiple of 2, 3, and 4 by direct computation ($2 \times 6 = 12$, $3 \times 4 = 12$, $4 \times 3 = 12$), and then added one to the common multiple.