

GRADING STUDENT PROJECTS AND FREE-RESPONSE QUESTIONS CONSISTENTLY, THROUGH SCORING GUIDES

The Need

Valid assessment of students' mathematical achievement requires using a variety of measures. In addition to mastering content, students should learn to think critically and to synthesize information. Evaluating their development in these areas is accomplished most validly through projects and extended response items.

It is not easy even for an individual teacher to maintain consistency when grading a set of examinations or homework; teachers often hear comments like, "You gave him three points and I only got two points for the same answer." Further, some high schools have adopted the practice of giving uniform departmental tests in specific courses. A fair evaluation on such examinations requires that all teachers in the department grade consistently with one another. Among common grading errors studied by Frank E. Saal, Ronald G. Downey, and Mary Anne Lahey (1980) are severity and leniency. Some teachers take pride in being "easy" graders, believing that students will be motivated by this "kindness." Others boast of being "tough" graders, claiming that this motivates students, and that students who succeed in their classes really "know their stuff." It is not good to be at either of these extremes -- teachers should set reasonable expectations, and assess student achievement based upon those expectations. Nevertheless, a broad spectrum of grading practices can exist in a single department, and thus a special effort may be required to achieve consistent grading on a common examination.

The College Board's Advanced Placement (AP) Program has been grading free-response questions for many years, and can grade over 600,000 free response questions in only a week, using approximately 500 readers. How do AP graders do it? They use a formal scoring guideline they develop for a problem, which every grader then uses for grading the problem. This uniform guide encourages consistency among the graders. Many teachers regularly use scoring guides, but may not realize they are doing so. For

example, they may break the solution of a problem into steps and award a fixed number of points for successfully completing each of the steps. This is a form of a scoring guide. Using a scoring guide can make grading easier, faster and more consistent. Furthermore, according to the National Council of Teachers of Mathematics, "Scoring guides, or rubrics, can help teachers analyze and describe students' responses to complex tasks and determine students' levels of proficiency. They can also help students understand the characteristics of a complete and correct response" (NCTM 2000, 22). Teachers should create and use these types of scoring guidelines and apply them to free-response items, short answer questions, and course projects.

Developing a Scoring Guide

Scoring guides are as applicable in classroom assessment as in large-scale assessment (Webb 1998). *Analytic* and *holistic* schemes are two common forms of scoring guides. Analytic scoring guides, ones that provide rules for assigning point values to specific features of a student's work, are the focus of this paper.

The Standards for Educational and Psychological Testing (AERA, APA, and NCME 1999) assert that the "criteria used for scoring test takers' performance on extended-response items should be documented." In addition, these standards require that a well-documented process for training test scorers be followed. For example, if a group of scorers are to grade an extended-response item, the procedures for training those scorers should: 1) include reviewing examples of test-taker responses that illustrate the various levels on the scale; and, 2) result in a degree of agreement among scorers that will allow for the scores to be interpreted as intended by the test writers. When scoring is done locally and requires scorer judgment, the test user is responsible for providing adequate training and instruction to the scorers and for examining scorer agreement and accuracy. The expected level of scorer agreement and accuracy should be documented (ibid).

In an effort to help teachers 1) *understand the mathematical content being addressed by a specific open-ended question*; 2) *assess what students understand about this specific*

mathematics content; and, 3) *create and use scoring guides to score responses to this specific item*, a scoring guide development and usage workshop was held in summer, 2001, for twelve Miami, Florida, high school teachers involved in a statewide collaborative mathematics project. The collaborative involved several units of the State University System of Florida and nine selected Florida counties. It was designed to improve the mathematics opportunities of Florida high school students by preparing their teachers, through needs-specific professional development, to offer more and stronger advanced mathematics courses. The scoring guide workshop was a component of this professional development program. This paper describes the structure and outcomes of that workshop.

The first step in designing the workshop was to create a suitable free-response item to comprise the assessment. The authors have used other free-response items for similar workshops. Other questions for which teachers need to design scoring guides can be found in the Appendix, but for this workshop the item in Figure 1 was presented. It is a modified item selected from the Florida Comprehensive Achievement Test (FCAT). The problem involved ratios and proportional reasoning -- a standard included in the NCTM *Principles and Standards* (2000).

A simulated “examination” consisting only of this problem was administered to a group of students from several of the high schools in the collaborative. From these papers, samples of student work were collected and workshop leaders created three training packets, as follows:

- A three-sample packet illustrating excellent, medium, and poor mathematical achievement on the problem.
- A ten-sample packet for training the participants in using the scoring guide the workshop participants would develop.
- A final thirty-sample packet for participants to grade once training was completed, to determine how consistently participants followed the scoring guide.

The workshop began with an introductory session on basic principles of assessment to set the stage for how a scoring guide can foster high-quality, consistent assessment. Next, the participants discussed and solved the assessment problem. Then, using a ten-point total, the group developed an analytic scoring guide for evaluating student performance on the problem.

The workshop presenters had been readers for the advanced placement calculus examination, and explained the AP reading procedure. At the AP reading, a scoring guide is designed for each free response problem, and every reader must follow it whether or not it represents the way he or she would prefer to evaluate the problem. In the workshop, participants designed their own scoring guide, and had to agree among themselves that they all would follow it. The scoring guide they designed assigned points to problem components as follows:

- 1) How far did Teresa row?
 - 1 point for use of similar triangle proportions;
 - 1 point for using the correct proportion;
 - 1 point for solving the proportion;
 - 1 point for obtaining the correct answer.
- 2) If the average speed for both rowers was 30 yards per minute, given that there was no break in the rowing, how long did it take to row the boat from Point A to Point B?
 - 1 point for correct use of the formula relating distance, time and speed;
 - 1 point for obtaining the correct answer.
- 3) Write a paragraph explaining how you decided how far Teresa rowed.
 - 1 point for an attempt to give an explanation;
 - 1 point for explaining the answer to the first question;
 - 1 point for correctness of the answer.
- 4) 1 point for use of correct units of measurement throughout the problem.

Following the AP training model, the participants first saw how the scoring guide was applied to the packet of sample questions, under the direction of the workshop coordinators. After this discussion, the teachers graded the papers in the second packet one by one, and discussed as a group the points they had assigned to each part of the problem, to be sure they were in agreement on how to apply the scoring guide. Once it was clear that all participants understood the scoring guide, identical packets of 30 papers were distributed for the teachers to grade independently. These scores then were entered into a spreadsheet, and the mean score awarded to each student was calculated. From this data, statistical indices measuring consistency of scoring among the graders were determined.

Validity and Reliability Considerations

Validity, the *degree* to which a test measures what it is designed to measure, and **reliability**, the *consistency* with which the test results measure whatever is being measured, are critical features of any assessment. A well-designed and uniformly applied scoring guide can help to optimize these aspects of a particular assessment.

Validity is the most important consideration in test evaluation. “The gathering of evidence may involve not only the examination of the present instrument in the present situation but also the available evidence on the use of the same or similar instruments in similar situations” (AERA, APA, and NCME 1999, CPA 1996). The validity of an achievement test is usually assessed by either: 1) the examination of test content by teachers or other educational professionals; or, 2) a statistical analysis which compares the results of an administration of the test with the results from related measures. Mathematics achievement includes knowing algorithms, definitions, and a collection of techniques associated with specific mathematics content. It also includes demonstrating the processes of problem solving, reasoning and proof; making connections; and, communicating ideas. Therefore, in addition to tests that employ multiple-choice, true/false, short answer, or matching formats, valid evaluation of mathematics achievement must include free-response questions and projects that require the demonstration of the above processes.

For a test to be valid, it must be reliable -- it is difficult to assess the validity of an instrument that has low reliability. For achievement tests that consist of multiple-choice questions, reliability is often assessed using statistical indices like the KR-20 (Lord and Novick 1968). If the test consists of open-ended questions, however, reliability is often assessed by comparing the scores assigned to student responses by several independent graders. "When test scoring involves a high level of judgment, indices of scorer consistency are commonly obtained" (AERA, APA, and NCME 1999). The choice of techniques for accomplishing this and deciding upon the minimum acceptable level for any index remain a matter of professional judgment. However, reliability data ultimately bear on the consistency of scores, so the importance of grader consistency cannot be overstated. Whereas a high degree of reliability is easily accomplished when multiple-choice, true/false, short answer, or matching formats are used, it is not so easily accomplished with free-response items or complicated projects. Scoring guides, when properly applied, offer a means of achieving this desired reliability.

Statistical Indices

Among common rater errors are severity or leniency, and lack of inter-rater reliability or agreement (Saal, Downey, and Lahey 1980). Leniency refers to a tendency to rate considerably higher than is warranted by the test taker, whereas severity refers to a tendency to assign scores lower than warranted. Three statistical indices were used in the workshop analysis to assess severity, leniency, and inter-rater consistency. They were:

- 1) The correlation between the set of rater's scores for the thirty students and the corresponding average scores for the students across all twelve raters;
- 2) The sum of the differences over all thirty students between a rater's score for a student and the across-raters average score for that student; and,
- 3) The sum of the absolute values of the differences between a rater's score and the across-raters average score.

Correlation across raters

One approach to assessing inter-rater reliability uses correlations between pairs of raters who evaluate the same examinee on an identical dimension (Saal, Downey, and Lahey 1980, Popham 1990). For the workshop, the correlation between the scores assigned by each particular rater and the average across-raters score for each student response was calculated to measure the consistency between each rater and the group average. A value of the correlation coefficient near 1.0 shows that the rater agrees with the other raters in terms of assigning relatively high scores to high-achieving students and relatively low scores to low-achieving students. However, this correlation coefficient for this workshop has the drawback that the group average score includes the score for the rater; the coefficient tends to be somewhat inflated by this contamination.

Sum of the Differences

One way to assess severity vs. leniency of a grader is through use of the sum of differences between the grader's scores for each student and the group average scores for that student. A large positive value indicates that the rater tends to be more lenient than the group as a whole, whereas an extreme negative value indicates that the rater tends to be more severe than the group as a whole. A value of the sum near 0 indicates that the rater is, on average, neither more severe nor more lenient than the group as a whole. However, this index being close to zero does not necessarily indicate that the rater is consistent with the group. For example, if he or she awarded 4 points higher than the mean on paper 1, 4 points lower than the mean on paper 2, etc., the sum of the differences would be zero, but the rater clearly would not be awarding scores that were consistent with those of the rest of the group.

Sum of the Absolute Values of the Differences

The sum of absolute differences is another measure of consistency between the rater and the group as a whole. For this index, a high value indicates inconsistency, but not necessarily leniency or severity, while a value close to zero indicates consistency.

Data and Analysis

The values of the above three indices for each grader in the workshop are given in the table 1. They are labeled as follows: r = correlation coefficient, SD = sum of differences, and SA = sum of absolute differences.

The data reveal some information about the individual raters. For example, grader tt was very inconsistent with the group average, as shown by the low correlation ($r = 0.60$), and by the high value of the sum of absolute differences ($SA = 32.4$). This grader was the most severe, with an SD value much less than that of any of the other graders. Grader rr had the second lowest correlation coefficient (0.82), and the value of SA was greater than that of any other grader. Also, this grader was the most lenient, as evidenced by the value of SD (27.9). It was noted that this grader did not agree with the scoring guide, but agreed to use it. Grader vv was among the most consistent ($r = 0.95$, $SA = 19.1$), but was somewhat more lenient ($SD = 13.9$) than the group average.

Another approach to assessing inter-rater reliability is to examine the standard deviations of the set of ratings assigned to a particular examinee by the raters (Saal, Downey, and Lahey 1980). The standard deviations of the scores given the students are given in table 2. The data reveal an inconsistency problem with the scores for student 18, whose ratings were: 6, 7, 8, 6, 9, 6, 6, 10, 10, 2, 10, and 5. A review of that student's paper indicated a correct response, although the student's paragraph explaining this student's solution was somewhat minimal. Some of the raters gave the student full credit for the solution, while other graders deleted points -- perhaps because of the brevity of the explanation.

Conclusions

The correlation coefficients showed a high degree of consistency among the graders, although the few graders who were not consistent lowered the values of r somewhat for all graders (due to self-contamination). The sum of absolute differences (SA) also shows that there was consistency across graders -- only two graders had SA values of more than one point per student paper. More training and experience in scoring-guide use may have reduced this moderate inter-rater inconsistency.

Based on the values of sum of differences (SD) between the individual grader's scores for each student and the average across-grader score for that student, only a few graders were either much too lenient or much too strict in assigning points. This problem also may be corrected through training and practice with using scoring guides.

Examination of the standard deviation of the scores for each student response across graders helped to pinpoint some of the inconsistencies, in particular with respect to the scores assigned to one student. This index was useful in helping graders to understand their own cognitive processes in assigning or subtracting points.

This workshop left the teachers confident that they would be able to use scoring guides as an assessment strategy in their classrooms. Further, the application of statistical indices as indicators of good grading practices was a worthwhile activity. The use and proper interpretation of these indices will provide teachers with feedback on their own consistency and leniency/severity when applying departmental scoring guides. This should enhance the reliability of those assessments.

At the end of the workshop, the participants were asked: "On a scale of 1 to 10 (10 being totally valid), do you believe the scoring guide you created and used measured validly the students' mathematical understanding of this problem in the context of similar triangles/proportional reasoning/units/equations/ratios/communications?" Most of the participants gave the exercise high marks -- nine of the responses were "8" or above. Some participants expressed concern about the way points were assigned when developing the scoring guide. One participant said, "The only drawback is the equal parts assigned (it seems not appropriate)." Another said, "I gave credit for answers that I would normally mark wrong. If given more time to make a scoring guide, as an assessment tool it is outstanding." A third participant was more positive about the exercise: "When I finished grading the set of papers I felt comfortable with the judgments I had made about the students' work."

The statistical analysis of the results of this training workshop demonstrate that the use of scoring guide in a departmental testing situation may prove valuable in enhancing

grading consistency and in helping teachers to understand how they, both individually and collectively, grade student papers.

REFERENCES

"The authors wish to express their appreciation to Dr. William Caldwell, Senior Fellow, Florida Institute of Education at the University of North Florida, for his encouragement and assistance."

American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME). *The Standards for Educational and Psychological Testing*. Washington, D. C.: American Educational Research Association. 1999.

Canadian Psychological Association. *Guidelines for Educational and Psychological Testing*. First edition 1987 © CPA 1996. Ottawa, Ontario. Retrieved on the WWW on September 11, 2001 at <http://www.cpa.ca/guide9.html>.

Lord, Frederick M. and Melvin R. Novick. *Statistical Theories of Mental Test Scores*. Reading, Mass: Addison-Wesley. 1968.

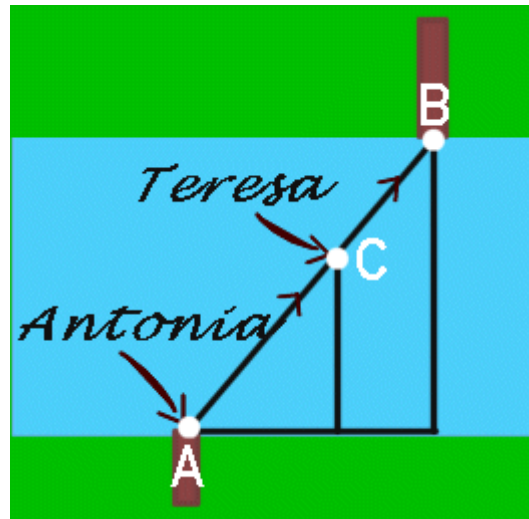
National Council of Teachers of Mathematics (NCTM). *Principles and Standards for School Mathematics*. Reston, Va.: NCTM. 2000.

Popham, James W. *Modern Educational Measurement: A Practitioner's Perspective*. Englewood Cliffs, N. J.: Prentice-Hall. 1990.

Saal, Frank E., Ronald G. Downey, and Mary Anne Lahey. "Rating the Ratings: Assessing the Psychometric Quality of Rating Data." *Psychological Bulletin*, 88(2), 413-428. 1980.

Webb, Norman L. *Thoughts on Assessment in the Mathematics Classroom*. In George W. Bright & Jeane M. Joyner (eds.) *Classroom Assessment in Mathematics: View from a National Science Foundation Working Conference*. 101-114. Lanham, MD: University Press of America, Inc. 1998.

Antonia and Teresa shared rowing "duty." They traveled in a straight line from one dock to another dock on the other side of the Seine River. For this section of the river its two shorelines are parallel. The width of this section of the river is 240 yards. Antonia began rowing at Point A. When Antonia rowed 200 yards (to Point C); Teresa immediately began to row. The shortest distance from Point C to the shore from which they started was 150 yards. Teresa rowed the rest of the way in a straight line to the dock (Point B). Use this information to respond to the three items below.



1. How far did Teresa row?
2. If the average speed for both rowers was 30 yards per minute, given that there was no break in the rowing, how long did it take to row the boat from Point A to Point B?
3. Write a paragraph explaining how you decided how far Teresa rowed.

Fig. 1 The Free-Response Item



Fig. 2. Workshop leader discusses scoring guide design in the workshop

	Grader											
Index	rr	sd	df	de	aa	ss	ff	ee	dd	tt	vv	ww
r	0.82	0.88	0.90	0.86	0.93	0.89	0.86	0.91	0.86	0.60	0.95	0.87
SD	27.9	1.9	-4.1	-7.1	3.9	-0.1	1.9	-2.1	-0.1	-28.1	13.9	-8.1
SA	33.6	15.8	20.8	20.8	16.4	19.9	19.8	23.3	24.8	32.4	19.1	18.6

Table 1. Grader statistics

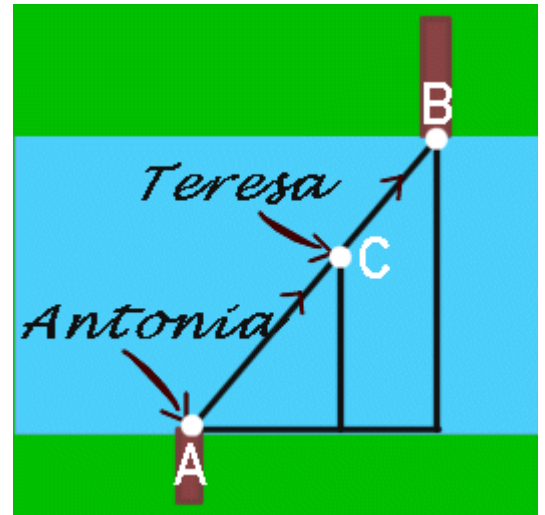
Student	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>	<i>6</i>	<i>7</i>	<i>8</i>	<i>9</i>	<i>10</i>
Std. Dev.	1.73	0.58	0.90	0.83	0.90	0.89	0.94	1.22	0.98	0.87
Student	<i>11</i>	<i>12</i>	<i>13</i>	<i>14</i>	<i>15</i>	<i>16</i>	<i>17</i>	<i>18</i>	<i>19</i>	<i>20</i>
Std. Dev.	0.79	0.90	1.03	1.06	0.90	1.08	0.79	2.43	0.90	1.24
Student	<i>21</i>	<i>22</i>	<i>23</i>	<i>24</i>	<i>25</i>	<i>26</i>	<i>27</i>	<i>28</i>	<i>29</i>	<i>30</i>
Std. Dev.	0.67	0.97	0.51	0.52	0.87	0.90	0.97	0.90	0.78	0.51

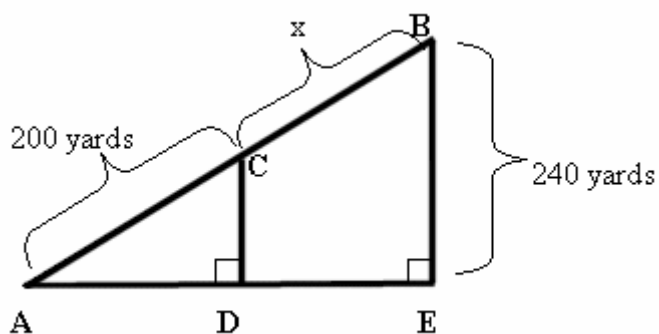
Table 2. Standard deviations of raters' scores per item

Appendix: Model Solution

Antonia and Teresa shared rowing "duty." They traveled in a straight line from one dock to another dock on the other side of the Seine River. For this section of the river its two shorelines are parallel. The width of this section of the river is 240 yards. Antonia began rowing at Point A. When Antonia rowed 200 yards (to Point C), Teresa immediately began rowing. The shortest distance from Point C to the shore from which they started was 150 yards. Teresa rowed the rest of the way in a straight line to the dock (Point B).

Use this information to respond to the three items below.





1. How far did Teresa row?

240 yards

Let x = distance Teresa rowed

$$\frac{200 + x}{x} = \frac{200}{150}$$

$$150(200 + x) = 200(240)$$

$$30000 + 150x = 48000$$

$$150x = 18000$$

$$x = 120 \text{ yards}$$

2. If the average speed for both rowers was 30 yards per minute, given that there was no break in the rowing, how long did it take to row the boat from Point A to Point B?

distance = (rate) (time)

T_T = time Teresa rowed

$$T_{\text{Total}} = \frac{120}{30} + \frac{200}{30}$$

$$\frac{\text{distance}}{\text{rate}} = \text{time}$$

T_A = time Antonia rowed

$$T_{\text{Total}} = \frac{320}{30}$$

T_{Total} = total time rowed

T_{Total} = total time rowed

$$T_{\text{Total}} = 10\frac{2}{3} \text{ minutes}$$

3. *Write a paragraph explaining how you decided how far Teresa rowed.*

To determine how far Teresa rowed, I labeled the point D on the shore that was 150 yards from Point C and Point E that was 240 yards from Point B. Similar Triangles have corresponding parts that are proportional. Right triangles, ACD and ABE are similar triangles (all angles are equal). Setting up a proportion using the corresponding parts and solving gave me my answer.

Let x = the distance Teresa rowed

$$\frac{200 + x}{240} = \frac{200}{150} \quad \frac{(\text{hypotenuse of Triangle ABE})}{(\text{leg of Triangle ABE})} = \frac{(\text{hypotenuse of Triangle ACD})}{(\text{leg of Triangle ACD})}$$

Appendix

Questions For Which Teachers Need To Design Scoring Guides

Questions For Which Teachers Need To Design Scoring Guides

Below are sample questions for which teachers should collaboratively design scoring guides.

1. Use the definition of the derivative of a function to show that $f'(x) = 6x$ for the function, $f(x) = 3x^2$.
2. Water boils at 212 degrees Fahrenheit, which is 100 degrees Celsius. Water freezes at 32 degrees Fahrenheit, which is 0 degrees Celsius. There is a linear relationship between measurements on the Fahrenheit scale and measurements on the Celsius scale. Use this information to **develop** the equation that expresses that relationship. Show your work.
3. A quality-control executive at a soft-drink bottling plant regularly gathers data on the amount of cola in 12-oz. cans produced at the plant. For a sample of 36 cans, he finds that the sample mean is 11.97 ounces and that the sample standard deviation is 0.08 ounces. Test the hypothesis that the population mean differs from 12 ounces. Use a 0.05 significance level.
4. The room temperature of the Coffee House is 70 degrees Fahrenheit. The temperature of the coffee in your cup is 200 degrees Fahrenheit. Since you don't want to scorch your tongue, you want to wait until the temperature of your coffee is 150 degrees Fahrenheit. You checked the temperature after 2 minutes and found that the temperature is 180 degrees. How much longer must you wait for your coffee to cool down to 150 degrees Fahrenheit?
5. Explain the method to find the sum of a and b for all integers a and b ?

6. You are given a rectangular sheet of plastic. The plastic sheet has dimensions 12 centimeters by 9 centimeters. Equal-sized squares are to be cut out of each corner of the sheet of plastic so that the sheet can be made into a container the shape of a rectangular prism. There will be no top on it. You want to make this container so that it holds the most amount of sugar so a waiter does not have to refill it as often as if the container holds a smaller amount of sugar. You are to tell the person making the plastic sugar containers exactly how long to make the side of one of the squares that will be cut out. Explain the process that you use to determine the length of the side of the square you cut out of each corner. Include in your explanation, why your process is helpful in determining the answer to this question.