When Wrong Answers Receive Top Grades

By Arthur Eisenkraft and Noah Eisenkraft

To find out whether the education community shares a collective understanding about how students should be evaluated, we surveyed 202 educators (from all grade levels) and scientists attending assessment workshops (Pennsylvania, California, and Massachusetts) or judging a national student competition (Washington, DC). The educators and scientists graded hypothetical student responses to trivial math problems with definitive answers. Depending on the person grading the question, the same student answer received anywhere from no points to full credit. When the instructions preceding a question changed, the graders often changed how they evaluated the students, even though the evidence about what the student knew remained the same. After the instructions changed, some graders awarded more credit for three wrong answers than for three right answers. The graders shared no consensus about how student answers should be graded. If students are going to be evaluated using tests, the education community must create tighter rubrics that ensure a higher degree of inter- and intragrader reliability.

Imagine you are grading a test on fractions. How many of a possible 5 points would you give the student who produces the following answer?

$$\frac{16}{64} = \frac{1\cancel{6}}{\cancel{6}4} = \frac{1}{4}$$

How confident are you that you graded the answer correctly? Would you be less confident if you learned that 65% of your colleagues gave the student a different grade? The data we collected shows that no matter how many points you awarded the answer above, at least 65% of your colleagues believe the student deserved a different grade. At a time when students are increasingly forced to prepare for or take high-stakes tests because of No Child Left Behind (NCLB; 2002), it is imperative that the education community come to a consensus about what we are looking for when we evaluate assessments and attempt to assure consistency across different graders (American Educational Research Association [AERA], 1999).

To find out whether the education community shares a collective understanding about how students should be evaluated, we surveyed 202 educators (from all grade levels) and scientists attending assessment workshops (Pennsylvania, California, and Massachusetts) or judging a national student competition (Washington, DC). The educators and scientists graded hypothetical student responses to trivial math problems with definitive answers. The graders were first asked to grade one problem for which the instruction to the hypothetical student was "simplify the fraction" and then asked to grade the same problem when the new instruction was "simplify the fraction and show all work." The graders were then asked to grade three problems for which the instruction to the hypothetical student was "simplify the fraction" and then asked to grade the same three problems when the new instruction was "simplify the fraction and show all work." Depending on the person grading the question, the same student answer received anywhere from no points to full credit. When the instructions preceding a question changed, the graders often changed how they evaluated the students, even though the evidence about what the students knew remained the same. After the instructions changed, some graders awarded more credit for three wrong answers than for three right answers. The graders shared no consensus about how student answers should be graded. If students are going to be evaluated using tests, the education community must create tighter rubrics that ensure a higher degree of inter- and intragrader reliability.

Is there intergrader reliability in the education community?

If a test has intergrader reliability, the same student answer would receive the same number of points regardless of who grades the test (Heubert & Hauser, 1999). To test whether there is intergrader reliability in the education community, we first asked 202 scientists and educators to assign a grade of 0, 1, 2, 3, 4, or 5 to the following student answers.

ANSWER of Student A:

$$\frac{16}{64} = \frac{1}{4}$$

ANSWER of Student B:

$$\frac{16}{64} = \frac{1\cancel{6}}{\cancel{6}4} = \frac{1}{4}$$

Over 77% of the graders awarded Student A full credit for this answer. The answer is mathematically correct, but a significant minority of the graders decided that the answer did not deserve full credit. These graders all agreed that the student performed well—nobody disputed that 1/4 is the correct answer—but some wanted to see more proof that the student used a correct procedure before giving full credit.

When there was some evidence that a hypothetical student used an improper procedure to arrive at the answer, there was even less intergrader agreement. By crossing out the 6s, Student B's answer suggests that he improperly "cancelled" the 6s and only arrived at the correct solution by chance. Over 20% of the graders reacted so negatively to Student B's cancelling marks that they gave the student no credit for the correct answer. Another 32% of the graders were not at all troubled by the evidence and gave the student full credit. The remaining graders we surveyed assigned 1, 2, 3, and 4 points in almost equal numbers. Some graders gave Student B the equivalent of an F, whereas others gave him a perfect mark. There is no evidence that Student B should expect any level of intergrader reliability.

Although we have our own personal beliefs about how many points Student B deserves, we do not believe that our grading intuition is objectively better than any of the approaches the graders consciously or unconsciously used. The graders spent time thinking about how the student's answer should be graded and provided convincing rationales for their evaluations. The graders who awarded no points wrote that, by crossing out the 6s in the numerator and denominator, Student B showed that he did not know how to simplify fractions. Some of the other graders who awarded no points further wrote that, although the answer is correct, this is no more than a coincidence and should not be rewarded. In contrast, the graders who gave full credit often commented that Student B has the correct answer and this is enough. Some of the graders who gave full credit wrote that they gave the student "the benefit of the doubt": they gave Student B full credit because they believed that he solved the problem correctly but forgot to put slash marks through the entire numerator and denominator. Other graders based their assessment on the principle of equity and argued that Student B deserved full credit because, if there is a chance that Student A used the same method as Student B, both students should receive the same high grade.

When there is uncertainty about whether a student used the correct procedure, intergrader reliability dwindles. The graders did not share an understanding of how many (if any) of the question's 5 points should reward process.

One way to minimize the effect of process uncertainty is to ask many questions. If a student uses the wrong process to answer a question, he or she will probably use the incorrect process many times throughout the test. If there are multiple questions, the grader should be able to discern whether the student knows the correct procedure and provide a grade that accurately reflects the student's knowledge.

To test whether asking additional questions would increase intergrader reliability, we asked our graders to evaluate two three-question sets. As there are three questions within each set, the graders were allowed to award anywhere from 0 to 15 points to each.

Simplify the fractions:

$$\frac{16}{64} \qquad \frac{25}{75} \qquad \frac{13}{39}$$

ANSWER Student A:

$$\frac{16}{64} = \frac{1}{4}$$

$$\frac{25}{75} = \frac{1}{3}$$

$$\frac{13}{39} = \frac{1}{3}$$

ANSWER Student B:

$$\frac{16}{64} = \frac{1\cancel{6}}{\cancel{6}4} = \frac{1}{4}$$

$$\frac{25}{75} = \frac{2\cancel{5}}{7\cancel{5}} = \frac{2}{7}$$

$$\frac{13}{39} = \frac{1\cancel{3}}{\cancel{3}9} = \frac{1}{9}$$

With more information from each student, it is relatively clear that Student A knows how to simplify fractions, whereas Student B does not. There is now much more evidence that Student B's first answer is only correct by chance. As predicted, the graders reacted similarly to the decrease in uncertainty. Almost 86% of graders gave Student A the full 15 points. A similar proportion of the graders awarded 5 points or fewer to Student B. By reducing the uncertainty about whether the process was correct, we were able to increase intergrader reliability.

Although asking multiple questions is an acceptable solution for the assessment of a relatively simple skill, it may not be possible to ask multiple questions when (1) the questions are more complex or (2) the test covers a large amount of material. When only a few questions can be asked, we suggest that it is best to create and disseminate a well-defined rubric explaining the expectations associated with each question. We further recommend that graders be trained

on the use of the rubric. Rubrics have been shown to increase accuracy from grader to grader (Marzano, 2000).

Is there intragrader reliability in the education community?

Through the second set of questions, we also explored intragrader reliability. When there is intragrader reliability, the same grader will evaluate student answers consistently in different situations (AERA, 1999).

To learn more about intragrader reliability, we asked graders to evaluate the same sets of answers that were described in the previous section. For this grading exercise, however, the graders were told that the students were given a different set of instructions. Instead of being told to "simplify the fractions," the hypothetical students in this exercise were asked to "simplify the fractions and show all work." Graders were then asked to award between 0 and 15 points to each set of answers.

Before discussing how grades were affected by the change in instructions, we must reiterate that the students' answers have not changed at all. The graders should still be confident that Student A knows how to simplify fractions and that Student B does not. The student's understanding of fractions is the same as it was when the previous instructions were given. The graders internal rubrics have, however, changed.

Although they were no less confident that Student A knew how to simplify fractions, almost 80% of the graders lowered Student A's grade, sometimes by as much as the full 15 points. Student B, still unable to simplify fractions, benefits on average from the change in instructions. More than twice as many people increased Student B's grades than reduced them because of the new instructions.

Although the graders knew no more or less about the student's knowledge, a slight change in instructions was so powerful that almost 20% of the graders awarded Student B a higher grade than Student A. Although the old adage says that two wrongs don't make a right, a small change in instruction leads some graders to score three wrongs higher than three rights.

There are obviously merits associated with encouraging students to "show all work," merits we will not discuss here. However, the surprising finding in our data is that inserting these three words produced large changes in how graders evaluated identical answers. Graders who changed their evaluations decided to evaluate how well the student followed instructions in addition to rewarding the student's ability to simplify fractions. Student A's answers, previously shown, would receive dramatically different grades from different graders because of the varying interpretation and importance placed on the "show all work" instruction. This dispersion occurs even though no ambiguity existed as to Student A's ability to simplify fractions.

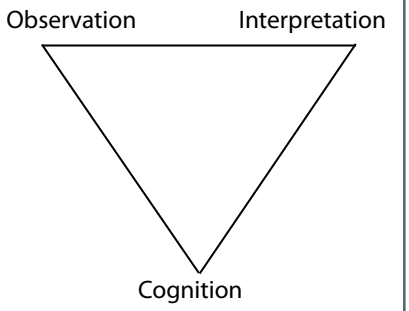
Using the assessment triangle to improve reliability

We can better understand the trends in our data by referring to the assessment triangle as a guide (Figure 1; Pellegrino, Chudowsky, & Glaser, 2001). The vertices of the assessment triangle are cognition, observation, and interpretation. These represent the critical factors underlying any assessment. Cognition represents the knowledge of the student. Observation represents the tasks chosen to explore what the student knows and is able to do. Interpretation represents how the assessor makes sense of the observation data to draw conclusions about the cognition or cognitive model of the student. The three vertices must work together for an effective assessment.

The need for coherence across the assessment triangle can be conveyed most clearly with an example. Imagine that a father takes his six-year-old daughter to the eye doctor. After sitting the girl down and putting drops in her eyes, the doctor asks the little girl to read the eye chart. The girl, forlorn,

FIGURE 1

Assessment triangle.



responds that she can't. Nodding his head in understanding, the doctor excuses himself from the room for a few minutes. As soon as he leaves, the little girl begins to cry. "Daddy," she says, "none of the words made sense."

In this example, the ophthalmologist, with limited observational data—"I can't read the chart"—incorrectly interpreted the data and concluded that the girl needs glasses. The father, with additional observational data—"None of the words made sense"—is better able to interpret the data available to him and correctly interprets the situation. The physiology of the girl's vision is what we want to know. The observation and interpretation can lead the doctor to an incorrect diagnosis or the parent to a correct diagnosis. When we are assessing the complex domain of student knowledge, understanding, and cognition, we are even more likely to make errors in interpretation.

To reduce the amount of error in our interpretation, the assessment triangle recommends clearly defining the cognition we want to ascertain. In the math problems involving fractions, we are trying to find out if the students are able to simplify fractions. When we observe three problems, we are quite confident in our interpretation that Student A is able to simplify fractions, whereas Student B is not able to simplify problems. When the instructions are changed to "simplify the fractions and show all work," the confidence we have in our interpretation of which student knows how to simplify fractions

has not changed. Yet, as we see from the data, our grades have changed.

The grades changed when the instructions changed because the graders did not a priori agree on the purpose of the assessment. With the first set of instructions to "simplify the fractions," it appears that the purpose of the assessment was to ascertain if the students can simplify fractions. With the second set of instructions to "simplify the fractions and show all work," some of the graders assumed and reported that the purpose of the assessment was to learn whether students could both simplify fractions and follow instructions. The reason that almost 20% of graders assigned more credit to 25/75 = 2/7 than 25/75 = 1/3 is that "showing all work" became more important in the assessment than finding the right answer for those graders and, by extension, the students they evaluate. If the graders had defined and communicated the cognition they hoped to ascertain (be it "simplify fractions" or "simply fractions and follow directions"), there would probably not have been such low intragrader reliability between two questions. A sample rubric for this trivial fraction problem is provided in Table 1. This sample rubric clearly shows the value of correct answers over incorrect answers and that the purpose of showing all work is to only assist in understanding the student work. This rubric ensures that precocious math students who can correctly simplify fractions do not receive lower grades for not having to "show all work." The rubric also ensures that incorrect answers do not receive higher grades than correct answers. Of course, an alternative rubric with a different set of values could similarly increase reliability.

If educators define what they are testing beforehand, it is possible to create tight rubrics that increase both inter- and intragrader reliability. If these rubrics exist, students may feel that the testing process is fairer and better able to assess their abilities. With tight rubrics and training on using those rubrics, we can increase grading reliability and be more confident of our assessments. This will help us all better prepare students for their high-stakes NCLB (2002) exams.

Even though college students are not given problems as simple as those in the study, the cautions about reliability measures are perhaps just as important at these institutions. In many instances, a pool of teaching assistants grades student exams. Only by having discussions about grading and creating rubrics can we ensure that student grades are consistent across teaching assistants. To help create consistency, some papers can be graded initially by multiple assistants and discussions can take place to ensure that all graders have the same interpretation for identical observations. This will not only ensure grading with better interrater and intrarater reliability, but will also be an important exercise and learning experience for the teaching assistants.

This "simplify the fraction" protocol is offered as a tool to help convince teachers, administrators, and policy makers of the need for rubric development, training of graders, and discussion of what we want to measure. It can also be used in methods courses as a memorable caution about the need for consistency in grading and clarity about grading and as a way to introduce the assessment triangle. In this way, our assessments will have a better chance of measuring what we want and better informing us of what our students know. ■

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: Author.

Heubert, J. P., & Hauser, R. M. (Eds.). (1999). *High stakes: Testing for tracking, promotion, and graduation.* Washington, DC: National Academies Press.

Marzano R. J. (2000). *Transforming classroom grading*. Alexandria, VA: Association for Supervision and Curriculum Development.

No Child Left Behind (NCLB) Act of 2001, Pub. L. No. 107-110, § 115, Stat. 1425 (2002).

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know.* Washington, DC: National Academies Press.

Arthur Eisenkraft (arthur.eisenkraft@umb.edu) *is the Distinguished Professor of Science Education and Director of the Center of Science and Mathematics in Context at the University of Massachusetts Boston.* ***Noah Eisenkraft*** *is an assistant professor of Organizational Behavior at the Kenan-Flagler Business School at the University of North Carolina, Chapel Hill.*

TABLE 1

Sample rubric for "Simplify the fractions and show all work."

	With no work shown	With work shown that demonstrates understanding	With work shown that demonstrates an identical confusion across all problems	With work shown that demonstrates multiple confusions
3 problems correct	15 points	15 points	9 points	0 points
2 problems correct	10 points	10 points	6 points	0 points
1 problem correct	5 points	5 points	4 points	0 points
0 problems correct	0 points	0 points	0 points	0 points