Lew Romagnano

# The Myth of Objectivity in Mathematics Assessment

**THE ASSESSMENT PRINCIPLE:** Assessment should support the learning of important mathematics and furnish useful information to both teachers and students.

—*Principles and Standards for School Mathematics* (NCTM 2000, p. 22)

A wide array of alternatives to traditional quiz-and-test assessment of students' mathematical understanding has been proposed in the last decade (e.g., Stenmark [1991]; NCTM [1995]; Greer et al. [1999]). Adding open-ended problems, performance tasks, writing assignments, and portfolios to teachers' assessment repertoires is important, these documents argue, because "assembling evidence from a variety of sources is more likely to yield an accurate picture of what each student knows and is able to do" (NCTM 2000, p. 24).

The decision by teachers to incorporate some of these less familiar assessment techniques is often framed as a trade-off between objectivity and subjectivity. Traditional assessment methods, which are sometimes narrowly focused on skills and procedures, are at least objective measures of those skills and procedures. By contrast, alternative approaches—which have the potential to assess students' conceptual understanding and their problem-solving and reasoning ability—are unfortunately subjective.

What does it mean for an assessment technique to be objective? The *American Heritage Dictionary of the English Language* defines the word as follows:

**Ob·jec·tive** *adj.* 1. Of or having to do with a material object as distinguished from a mental concept, idea, or belief. Compare subjective. 2. Having actual existence or reality. 3. a. Uninfluenced by emotion, surmise, or personal prejudice. b. Based on observable phenomena; presented factually: *an objective appraisal*.

A student's mathematical understanding—for example, knowledge of linear functions or the capacity to solve nonroutine problems—is a "mental concept" and as such can be observed only indirectly. Further, a teacher's appraisal of this knowledge cannot help but be influenced by emotion or surmise. Objectivity, like the mythical pot of gold at the end of the rainbow, would be wonderful if we could have it, but it does not exist. All assessments of students' mathematical understanding are subjective.

A more useful way to characterize methods of assessment would be with respect to their *consistency,* or reliability, and the *meaning,* or validity, of the information that they provide. When different teachers use a consistent method to assess the knowledge of a given student, the teachers' assessments agree. When two students have roughly the same level of understanding of a set of mathematical ideas, consistent assessment of these students' understandings is roughly equal, as well.

Meaningful methods give teachers information about students' understanding of specific mathematical ideas and how this understanding changes over time. This information can be used to make appropriate instructional decisions.

The following examples of information collected by using three familiar methods—a teacher-made quiz, the Advanced Placement calculus test, and

***Objectivity would be wonderful if we could have it, but it does not exist***

*Edited by* ***Vena Long***
*vlong@utk.edu*
*University of Tennessee*
*Knowville, TN 37996-3400*

*Lew Romagnano, romagnal@mscd.edu, teaches at Metropolitan State College of Denver, Denver, CO 80217-3362. His current interests include teaching mathematics, supporting teacher professional development in an era of reform, and research on teaching and learning to teach.*

*The Editorial Panel welcomes readers' responses to this article or to any aspect of the* Assessment Standards for School Mathematics *for consideration for publication as an article or as a letter in "Reader Reflections."*

the SAT-I Mathematics test—illustrate both the inherent subjectivity of these methods and the value of considering, instead, the consistency and meaning of the methods.

## A TEACHER-MADE QUIZ

An algebra teacher who is hoping to assess students' ability to solve quadratic equations might include the following task on a quiz:

Solve: $x^2 + x - 6 = 0$.

***A conclusion about a student's knowledge would require the teacher's judgment***

**Figure 1** shows one student's response to this task. Before reading any further, assess this second-year algebra student's work and assign a point value, assuming that "full credit" is five points and that partial credit is allowed.

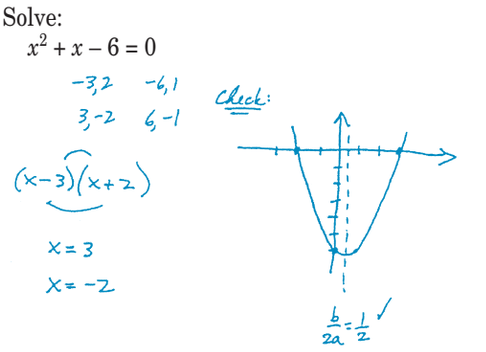


Fig. 1
A typical task and one student's solution

This student has correctly listed all factors of the constant coefficient of the expression on the left side of the equation. She used the first of these factor pairs to construct two potential binomial factors of the quadratic expression. She seems to have checked the "outside" and "inside" products to determine whether multiplying these binomials produces a quadratic expression with the proper middle term. Her first misstep is that the product of these binomials does not produce the correct middle term. She seems satisfied, though, and she proceeds to write the solutions of the equation. Then, in her "check," she shows that a graph of the quadratic function $y = x^2 + x - 6$ has $x$-intercepts at $-2$ and 3. Her graph, with its incorrect axis of symmetry, confirms her answers.

What does this student know about solving quadratic equations? She seems to know that one way to solve them involves factoring the quadratic expression. She also seems to know a method. She knows the factors of $-6$. She might know that if a product of two terms is zero, then at least one of the terms is zero. She does know that the solutions of this quadratic equation are specific points on the graph of a quadratic function.

The teacher could conclude that this student knows a great deal about solving quadratic equations but has some trouble keeping signs straight, since both mistakes are sign errors. Or the teacher could conclude that this student has tried to memorize a procedure for solving quadratic equations and has—perhaps without any understanding—reproduced most, but not all, the steps correctly. A conclusion about this student's knowledge of quadratic equations and how to solve them would require the teacher's judgment. This judgment would have to be exercised in the face of incomplete and ambiguous evidence furnished by the student and without any explicit guidance.

What score did you assign to this paper? Why did you assign that score? These questions have been asked of practicing teachers in many classes, workshops, and conference sessions in the last few years. The responses have been distributed more or less evenly among the scores 2, 3, and 4. This 40 percent variation is attributable to judgments that individual teachers made about the relative importance of each aspect of this student's work described previously. In other words, these scores are subjective.

Thus, an apparently straightforward question of the most common and traditional type produced assessment information that says as much about the scorer as it does about the student. The scores on quizzes and tests that consist of such items as this example are inconsistent and may not offer much information about the mathematical knowledge of the student.

## THE ADVANCED PLACEMENT CALCULUS TEST

Advanced Placement calculus tests have been taken by high school students for four decades. These tests include multiple-choice items, the staple of standardized tests, and a set of free-response questions for which students must supply answers, show their work, and explain their reasoning. This respected measure of students' knowledge of elementary calculus is thus, in part, an alternative assessment.

The 1998 Advanced Placement Calculus AB test contained the free-response question shown in **figure 2.** Students' solutions to free-response questions such as this one are scored by at least two readers, who follow an explicit set of guidelines for assigning points and must agree on the score assigned to each paper. The rubric used to score this problem is shown in **figure 3.**

In this scoring rubric, the nine points allocated for this problem are assigned as follows: two points for finding the derivative implicitly and verifying it,

Consider the curve defined by $2y^3 + 6x^2y - 12x^2 + 6y = 1$.

*a*) Show that

$$\frac{dy}{dx} = \frac{4x - 2xy}{x^2 + y^2 + 1}.$$

*b*) Write an equation of each horizontal tangent line to the curve.

*c*) The line through the origin with slope –1 is tangent to the curve at point *P*. Find the *x*- and *y*-coordinates of point *P*.

Fig. 2
1998 Advanced Placement Calculus AB free-response question 6
(Source: College Board)

| | |
|---|---|
| *a*) Show that $\frac{dy}{dx} = \frac{4x - 2xy}{x^2 + y^2 + 1}$. | 1: implicit differentiation<br>1: verifies expression for $dy/dx$ |
| *b*) Write an equation of each horizontal tangent line to the curve. | 1: sets $dy/dx = 0$<br>1: solves $dy/dx = 0$<br>1: uses solutions for $x$ to find equations of horizontal tangent lines<br>1: verifies which solutions for $y$ yield equations of horizontal tangent lines |
| *c*) The line through the origin with slope –1 is tangent to the curve at point *P*. Find the *x*- and *y*-coordinates of point *P*. | 1: $y = -x$<br>1: substitutes $y = -x$ into equation of curve<br>1: solves for $x$ and $y$<br>or<br>1: sets $dy/dx = -1$<br>1: substitutes $y = -x$ into $dy/dx$<br>1: solves for $x$ and $y$ |

Fig. 3
Scoring rubric for Advanced Placement calculus free-response question
(Source: College Board)

four points for finding where the derivative has the value of zero and verifying that the tangent lines are horizontal there, and three points for using one of two different specified approaches to find the point of tangency of the line $y = -x$. Use this rubric to score the work shown in **figure 4.**

On part (*a*), the student's correct implicit differentiation would earn two points. Setting the derivative equal to 0 and, after a false start, solving for $x$ and $y$ would earn two more points for part (*b*). Finally, in part (*c*), setting the derivative equal to –1 is worth an additional point. The score for this student would be five points out of a possible nine points.

This example shows a consistent assessment method. Unlike the previously discussed quadratic-equation task, for which arguments could be made for a wide range of scores, the Advanced Placement calculus task itself, for which predictable routes to the solution exist, combines with the rubric that specifies the routes and assigns points, thereby facilitating agreement on a single score.

How useful is this score? What does five points out of nine mean on this task? How much of the calculus that this task is meant to assess does this student know? Will everyone who obtains a five-point score on this problem know the same amount? This student is clearly able to differentiate implicitly. The student also seems to know that the derivative is related to the slope of the tangent to the curve at a point. Given the difficulty that this student had in completing parts (*b*) and (*c*), any other inferences about the student's mathematical knowledge would be difficult.

Another student who earned the same score for parts (*a*) and (*b*) could have earned three points for part (*c*) by successfully completing the first of the two solution strategies outlined in the rubric. However, that strategy makes no use of calculus. Therefore, a score of seven points out of nine could be earned without furnishing any additional evidence of understanding of calculus. To put these scores in context, the average score of all 1998 Advanced Placement Calculus AB test-takers on this item was 2.86, and 80 percent of those test-takers scored 4 or lower (College Board).

As this example illustrates, the specificity required for consistent scoring can reduce the usefulness of the scores themselves. Taken together, these two assessment examples show that, although consistency is necessary, it is not sufficient to ensure that assessment information is useful.

## THE SAT-I MATHEMATICS TEST

The Scholastic Assessment Test (SAT) is a widely used example of a standardized, norm-referenced test. The test is administered under *standardized* conditions, including the amount of time allotted and the directions and resources provided for the test-takers. The scores are *norm-referenced*: the student is told how his or her performance compared with that of a comparison group of students who already took the test instead of being told how many questions he or she answered correctly and incorrectly.

The mean score on the SAT-I Mathematics test is 500, the standard deviation of scores is 100, and the test items are chosen so that the scores of the comparison group are approximately normally distributed, as shown in **figure 5** (Crocker and Algina

*a*) Show that

$$\frac{dy}{dx} = \frac{4x - 2xy}{x^2 + y^2 + 1}.$$

$6y^2\frac{dy}{dx} + 6x^2\frac{dy}{dx} + 12xy - 24x + 6\frac{dy}{dx} = 0$

$(6y^2 + 6x^2 + 6)\frac{dy}{dx} = -12xy + 24x$

$\frac{dy}{dx} = \frac{-12xy + 24x}{6y^2 + 6x^2 + 6} = \frac{12(-xy + 2x)}{6(y^2 + x^2 + 1)} = \frac{2(-xy + 2x)}{y^2 + x^2 + 1}$

$\frac{dy}{dx} = \frac{-2xy + 4x}{y^2 + x^2 + 1} = \frac{4x - 2xy}{y^2 + x^2 + 1}$ ✓

*b*) Write an equation of each horizontal tangent line to the curve.

$\frac{-2xy + 4x}{y^2 + x^2 + 1} = 0$

$-2xy + 4x = 0$

$4x = 2xy$

$2 = y$

$2(2)^3 + 6x^2(2) - 12x^2 + 6(2) = 1$

$16 + 12x^2 - 12x^2 + 12 = 1$

$28 = 1$

$-2xy + 4x = 0$

$-2x(y - 2) = 0$

$x = 0, \quad y = 2$

$2y^3 + 6(0)^2y - 12(0)^2 + 6y = 1$

$2y^3 + 6y = 1$

$2y^3 + 6y - 1 = 0$

*c*) The line through the origin with slope –1 is tangent to the curve at point *P*. Find the *x*- and *y*-coordinates of point *P*.

$\frac{-2xy + 4x}{y^2 + x^2 + 1} = -1$

$-2xy + 4x = -y^2 - x^2 - 1$

$-2xy + y^2 = -4x - x^2 - 1$

Fig. 4
Sample student work

***The specificity required for consistent scoring can reduce the usefulness of the scores***

1986). A student who receives a score of 600 on this test actually earned a raw score that placed him or her one standard deviation above the mean raw score of the comparison group. That student scored higher than about 84 percent of the students with whom he or she is being compared.

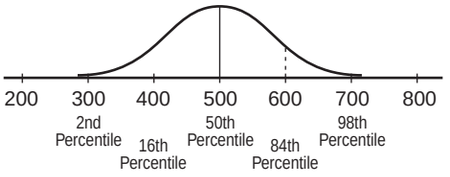


Fig. 5
SAT-I Mathematics test-score distribution
(Source: Crocker and Algina 1986)

Suppose that student *x* scores 470 on the SAT-I Mathematics test, whereas student *y* scores 530 on the same test. What can you conclude about the mathematical knowledge of these two students? Most consumers of these scores—the students themselves, their parents or guardians, teachers and administrators, college admissions officers, and newspaper reporters—would be confident that student *y* knows more. What is the meaning of these two scores? To answer this question, understanding how these tests are designed is important.

The creators of such tests as the SAT base their work on the assumption that students *x* and *y* each possess a certain amount of knowledge, ability, or (for the SAT) potential to succeed in the first year of college. If they could ask students all possible questions, the resulting "true scores" on this complete test would accurately measure their knowledge, ability, or potential. However, constructing and administering such a test is impossible. Instead, the designers create a test that consists of questions that are, in effect, a random sample drawn from the universe of all possible questions.

Like the results of any survey that is based on a sample drawn from some population, the actual scores that students earn on this test are only approximations of their true scores. Each actual score has some measurement error associated with it. A full report of a student's performance on this test would use the actual score and the measurement error to build an interval estimate.

For the SAT-I Mathematics test, the standard error of measurement is about thirty points. Student *x*'s actual score of 470, combined with this measurement error, tells us that we can be 95 percent sure that her or his true score is somewhere between 410 and 530, an interval that extends sixty points, that is, two "standard errors," on either side of the actual score. Similarly, student *y*'s true score is, with 95 percent certainty, between 470 and 590. See **figure 6.**

These confidence intervals overlap; students *x* and *y* would need actual scores that differed by at least eighty-four points for us to be 95 percent sure that their true scores were different. Because their

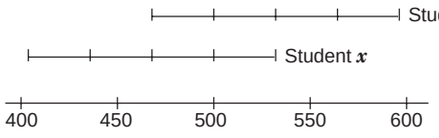


Fig. 6
Interval estimates of two SAT scores

actual scores differ by only sixty points, we do not have enough evidence to conclude that their knowledge differs at all. See the **appendix** for a derivation of these statistics.

The consistency of the assessment information furnished by the SAT is reduced by the seldom-reported variability introduced by measurement error. What do the scores mean? What mathematical ideas are being assessed by this test? How much mathematics is known by students *x* and *y,* whose scores are statistically the same? The norm-referenced score reported for each student—a score that simply describes how that student did relative to students in the comparison group—carries little information about how much that student understands of the arithmetic, elementary algebra, and geometry content of the test.

In the eyes of parents, administrators, and other consumers of assessment information, standardized, norm-referenced tests are the "gold standard" of objective assessment. However, objectivity—even in these tests—does not exist. Human judgment about mental constructs is introduced when test designers and consumers decide "what items to include on the test, the wording and content of the items, the determination of the 'correct' answer, . . . how the test is administered, and the uses of the results" (FairTest: The National Center for Fair and Open Testing), as well as when designers assume that at any given time, each student possesses a certain amount of knowledge, ability, or potential that can be measured, with some measurement error, by a single instrument. Such a test is only one way to conceptualize knowledge, ability, or potential. If knowledge is multifaceted, complex, individually constructed, and inextricably tied to the context in which the learning occurs—as more than two decades of research on learning indicate (Davis, Maher, and Noddings 1990; Battista 1999)—then no single instrument is likely to "measure" that knowledge in any consistent and meaningful way.

## DISCUSSION

In educational assessment—the myriad processes by which humans try to determine what other humans "know"—objectivity is a term that simply does not apply. Alternatively, we can strive for "agreed-on subjectivity." The following two specific suggestions can help improve the consistency and usefulness of assessment information gathered by teachers.

First, design classroom assessment tasks that are likely to elicit the information that you seek. Ask yourself the following questions: What is the mathematics that I am trying to assess here? What tasks will tap this mathematics most directly? A teacher of second-year algebra might want to know what students understand about quadratic equations and the techniques for solving them. The question in the previously discussed example—consisting of the one-word imperative "solve"—does not directly ask students to supply much information about their understanding. The set of tasks in **figure 7,** for example, does so more specifically. Greer et al. (1999) offer guidelines for creating and adapting tasks for classroom assessment.

*a*) Use one of the symbolic methods that were developed in class to find solutions to the equation

$$x^2 + x - 6 = 0.$$

*b*) Explain the method that you used in part (*a*).

*c*) Use the graph of a function to illustrate the solutions that you found in part (*a*).

*d*) Finding no real solutions to a quadratic equation is possible. Explain how this result could happen. Give an example that illustrates your explanation.

Fig. 7
A revised quadratic-equation task

Second, before assigning any task to students, devise—and share with the students—guidelines for scoring their work. See Thompson and Senk (1998) and Greer et al. (1999). Ask yourself what types of responses you are likely to get from students to these tasks and what you will accept as evidence of adequate understanding. Thinking these questions through before giving the tasks to students helps clarify the tasks themselves. It also helps align the tasks with the in-class instruction. Sharing these guidelines with students communicates expectations and makes meeting them more likely. One set of guidelines for scoring student work on the tasks in **figure 7** is proposed in **figure 8.**

## CONCLUSION

Such false dichotomies as "objective versus subjective" and "traditional versus alternative" derail

***No single instrument is likely to measure knowledge in any consistent and meaningful way***

5 – All the characteristics of 4, plus either a valid example with a clear explanation for part (*d*) or exceptional responses to parts (*a*) through (*c*) along with a response to part (*d*) that might have some minor flaws.

4 – Correct responses to parts (*a*) through (*c*): correct equation solutions, along with a valid explanation of the method; sketch of graph with all important features correct and labeled.

3 – Substantial evidence of understanding of quadratic equations: some minor errors (not central to understanding quadratic equations) are the only information that is missing from the characteristics of a 4.

- - - - - - - - - - - - - - - - - - - - - - - - -

2 – Some evidence of understanding of quadratic equations is present: either a symbolic solution or a graphical illustration, perhaps with some minor errors.

1 – Little understanding of quadratic equations is shown: major errors in all parts of the problem.

0 – No attempt made.

Fig. 8
A scoring rubric for the revised task

***Teachers should consider ways to make assessment more consistent and useful***

meaningful discussions of the important issues in mathematics assessment. The labels "traditional" and "alternative" are meaningless; a five-question classroom quiz can give detailed information about what students know, or it can furnish very little information, depending on how it is designed, scored, and used. No "objective" assessment occurs; subjective—that is, human—knowledge, beliefs, judgments, and decisions are unavoidable parts of any assessment scheme. Teachers should consider ways to make assessment of students' mathematical understanding, as well as the information gathered through that assessment, more consistent and useful.

## APPENDIX

Computing the measurement error and confidence interval around test scores depends on the concept of *reliability*. The reliability of a test is an answer to the question, How accurately does this test measure what it intends to measure? In other words, if the test is administered many times to the same student, how close will the results be? If we gave the test to two students who possess the same amount of the knowledge or ability being measured, how close would the scores be? As another example, if we were discussing the reliability of a thermometer, we would ask how close thermometer readings are to the actual temperature and how consistently the thermometer produces these readings.

One way to determine the reliability of a test is to correlate students' scores on repeated administrations of that test. A perfectly reliable test—one that reports students' true scores with no error—would have a "test-retest" reliability $\rho_{XX'} = 1$. However, no test is perfectly reliable. If you could repeatedly administer a test, the set of scores for a particular student would be distributed around the student's true score. See **figure A-1.** The more reliable the test, the higher the test-retest correlation and the tighter the distribution of scores. For the very reliable SAT-I Mathematics test, $\rho_{XX'} = 0.91$.

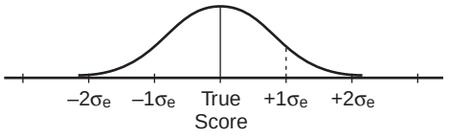


Fig. A-1
The distribution of actual scores around a student's true score
(Source: Crocker and Algina 1986)

The standard deviation of this distribution of scores is the *standard error of measurement,* $\sigma_e$. It can be calculated using the standard deviation of the test scores, $\sigma_X$, and the test's reliability, $\rho_{XX'}$, using

$$\sigma_e = \sigma_X\sqrt{1-\rho_{XX'}}.$$

For the SAT-I Mathematics test, $\sigma_X = 100$ and $\rho_{XX'} = 0.91$, so

$$\begin{aligned}\sigma_e &= 100\sqrt{1-0.91}\\ &= 100\sqrt{0.09}\\ &= 100(0.3)\\ &= 30 \text{ points.}\end{aligned}$$

Therefore, for this student, 68 percent of the scores that she would earn if she were to take the test repeatedly would be within thirty points on either side of her true score. Similarly, 95 percent of her scores would be within sixty points on either side of her true score.

Imagine that the test could be administered repeatedly to two different students. If the difference between these two students' scores is computed every time that the test is administered, these difference values would also lie on a distribution, this time around the true difference score for these students. Because the distributions of the two scores are independent, the variance of this difference distribution is equal to the sum of the two individual variances:

$$\sigma^2_{X-Y} = \sigma^2_X + \sigma^2_Y$$

For the SAT-I Mathematics test,

$$\begin{aligned}\sigma^2_X &= \sigma^2_Y\\ &= \sigma^2_e,\end{aligned}$$

so

$$\sigma^2_{X-Y} = \sigma^2_e + \sigma^2_e$$
$$= 2\,\sigma^2_e.$$

Therefore, the "standard error of the difference" between two SAT-I Mathematics test scores is

$$\sigma_{X-Y} = \sqrt{2}\sigma_e$$
$$\approx 1.4(30)$$
$$= 42 \text{ points.}$$

To be 95 percent sure that two actual scores represent different true scores, the actual scores would have to differ by at least eighty-four points.

## REFERENCES


*The American Heritage Dictionary of the English Language,* s.v. "objective."

Battista, Michael T. "The Mathematical Miseducation of America's Youth: Ignoring Research and Scientific Study in Education." *Phi Delta Kappan* 80 (February 1999): 424–33.

College Board. "1998 Free-Response Questions." www.collegeboard.org/ap/calculus/frq98/index.html. World Wide Web.

Crocker, Linda, and James Algina. *Introduction to Classical and Modern Test Theory.* New York: Holt, Rinehart, & Winston, 1986.

Davis, Robert B., Carolyn A. Maher, and Nel Noddings, eds. *Constructivist Views on the Teaching and Learning of Mathematics.* JRME Monograph no. 4. Reston, Va.: National Council of Teachers of Mathematics, 1990.

FairTest: The National Center for Fair and Open Testing. "What's Wrong with Standardized Tests." www.fairtest.org/facts/whatwron.htm. World Wide Web.

Greer, Anja S., Helen L. Compton, Alice B. Foster, Jo Ann Mosier, Lew Romagnano, and Carmen Rubino. *Mathematics Assessment: A Practical Handbook for Grades 9–12.* Assessment Standards for School Mathematics Addenda Series, edited by William S. Bush and Jean Kerr Stenmark. Reston, Va.: National Council of Teachers of Mathematics, 1999.

National Council of Teachers of Mathematics (NCTM). *Assessment Standards for School Mathematics.* Reston, Va.: NCTM, 1995.

———. *Principles and Standards for School Mathematics.* Reston, Va.: NCTM, 2000.

Stenmark, Jean Kerr, ed. *Mathematics Assessment: Myths, Models, Good Questions, and Practical Suggestions.* Reston, Va.: National Council of Teachers of Mathematics, 1991.

Thompson, Denisse R., and Sharon L. Senk. "Implementing the *Assessment Standards for School Mathematics:* Using Rubrics in High School Mathematics Courses." *Mathematics Teacher* 91 (December 1998): 786–93.