

Posted with permission from the *Mathematics Teacher*, © 2001 by the
National Council of Teachers of Mathematics. All rights reserved.

The Myth of Objectivity in Mathematics Assessment

Lew Romagnano
The Metropolitan State College of Denver
Department of Mathematical and Computer Sciences
Campus Box 38, P.O. Box 173362
Denver, CO 80217-3362
romagnal@mscd.edu

To appear in the *Mathematics Teacher*, January 2001

The Myth of Objectivity

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)

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 repertoire 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).

The decision by teachers to incorporate some of these less familiar assessment techniques is often framed as a tradeoff 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. On the other hand, alternative approaches—which have the potential to assess students' conceptual understanding, problem-solving and reasoning ability—are unfortunately subjective.

What does it mean for an assessment technique to be objective? The American Heritage dictionary defines the word this way:

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—knowledge of linear functions, or the capacity to solve non-routine problems—is a "mental concept," and as such can only be observed indirectly. Further, a human 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 great 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 they provide. When a consistent method is used by different teachers to assess the knowledge of a given student, the teachers' assessments will agree. When two students have roughly the same level of understanding of a set of mathematical ideas, consistent assessment of these students' understandings will be roughly equal as well.

Meaningful methods provide teachers with information about student understanding of specific mathematical ideas and how this understanding changes over time, information that can be used to make appropriate instructional decisions.

The following examples of student evidence collected using three familiar methods—a teacher-made quiz, the AP Calculus test, and the SAT-I Mathematics

The Myth of Objectivity

test—illustrate both the inherent subjectivity of these methods and the value of thinking instead of their consistency and meaning.

A Teacher-Made Quiz

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

$$\text{Solve: } x^2 + x - 6 = 0.$$

Figure 1 shows one student's response to this task. Before reading any further, assess this Algebra II student's work (and assign a point value), assuming that "full credit" is five points, and that partial credit is allowed.

Solve:

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

$-3, 2$ $-6, 1$
 $3, -2$ $6, -1$

$(x-3)(x+2)$

$x = 3$
 $x = -2$

Check:

$\frac{b}{2a} = \frac{1}{2}$ ✓

Figure 1. A typical task, and one student's solution.

This student has listed, correctly, 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 see if multiplying these binomials produces a quadratic expression with the proper middle term. Here is her first misstep; the product of these binomials does not produce the correct middle term. She seems satisfied, though, and she proceeds to write the solutions to 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 way to do this. She knows the factors of -6 . She might know that if a

The Myth of Objectivity

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.

We could conclude that this student knows a great deal about solving quadratic equations, but has some trouble keeping signs straight. (Note that both mistakes are sign errors.) Or, we 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) of the steps correctly. In any case, a conclusion about this student’s knowledge of quadratic equations and how to solve them would require the judgment of the teacher. This judgment would have to be exercised in the face of incomplete and ambiguous evidence provided 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 put to 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 is 40 percent variation, due to judgments made by individual teachers about the relative importance of each of the aspects of this student’s work described above. 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 made up of items such as this example are inconsistent and may not carry much information about the mathematical knowledge of the student.

The AP 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 have to supply answers, show their work, and explain their reasoning. (Thus, this respected measure of students’ knowledge of elementary calculus is, in part, an alternative assessment.)

The 1998 AP Calculus AB test contained the free-response question shown in figure 2. Student solutions to free-response questions like this are scored by at least two readers who follow an explicit set of guidelines for assigning points and have to agree on the score assigned to each paper. The “rubric” used to score this problem is shown in figure 3.

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 .

Figure 2. 1998 AP Calculus AB free-response question 6.

The Myth of Objectivity

<p>(a) Show that $\frac{dy}{dx} = \frac{4x - 2xy}{x^2 + y^2 + 1}$.</p> <p>(b) Write an equation of each horizontal tangent line to the curve.</p> <p>(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.</p>	<p>1: implicit differentiation</p> <p>1: verifies expression for $\frac{dy}{dx}$</p> <p>1: sets $\frac{dy}{dx} = 0$</p> <p>1: solves $\frac{dy}{dx} = 0$</p> <p>1: uses solutions for x to find equations of horizontal tangent lines.</p> <p>1: verifies which solutions for y yield equations of horizontal tangent lines.</p> <p>1: $y = -x$</p> <p>1: substitutes $y = -x$ into equation of curve.</p> <p>1: solves for x and y</p> <p>– or –</p> <p>1: sets $\frac{dy}{dx} = -1$</p> <p>1: substitutes $y = -x$ into $\frac{dy}{dx}$</p> <p>1: solves for x and y</p>
--	--

Figure 3. AP Calculus free-response question scoring rubric.

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; 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 student work shown in Figures 4a - c.

On part (a), the student's correct implicit differentiation would garner 2 points. Setting the derivative equal to 0 and (after a false start) solving for x and y would earn 2 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 5 out of a possible 9 points.

This is an example of a consistent assessment method. Unlike the quadratic equation task discussed above, for which arguments could be made for a wide range of scores, the AP Calculus task itself, for which there are predictable routes to the solution, and the rubric that specifies the routes and assigns points, combine to make it easy to agree on a single score.

The Myth of Objectivity

(a) Show that $\frac{dy}{dx} = \frac{4x - 2xy}{x^2 + y^2 + 1}$.

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

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

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

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

Figure 4a.

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

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

$$-2xy + 4x = 0$$

$$4x = 2xy$$

$$x = y$$

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

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

$$28 = 1$$

$$-2xy + 4x = 0$$

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

$$x = 0, y = 2$$

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

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

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

Figure 4b.

(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 .

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

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

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

Figure 4c.

How useful is this score? What does 5 out of 9 mean on this task? How much of the calculus this task is meant to assess does this student know? Will everyone who scores 5 on this problem know the same amount? It is clear that this student is 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 this student had completing parts (b) and (c), any other inferences about mathematical knowledge would be difficult.

Note that another student who earned the same score for parts (a) and (b) could have earned 3 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 7 out of 9 could be earned without providing any additional evidence of understanding of calculus. (To put these scores in context, the average score of all 1998 AP Calculus AB test-takers on this item was 2.86, and 80 percent of those test-takers scored 4 or lower¹.)

As this example illustrates, the specificity required for consistent scoring can have the effect of reducing the usefulness of the scores themselves. Taken together, these two assessment examples show that, while consistency is necessary, it is not sufficient to ensure that assessment information will be 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*: rather than being told how many questions she got right and wrong, a student is told how her performance compared to those of a comparison group of students who already took the test.

¹ This information is available from the College Board at: www.collegeboard.org/ap/calculus/frg98/index.html.

The Myth of Objectivity

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. (See fig. 5.) A student who receives a score of 600 on this test actually earned a raw score that placed her one standard deviation above the mean raw score of the comparison group. This student scored higher than about 84 percent of the students she is being compared to.

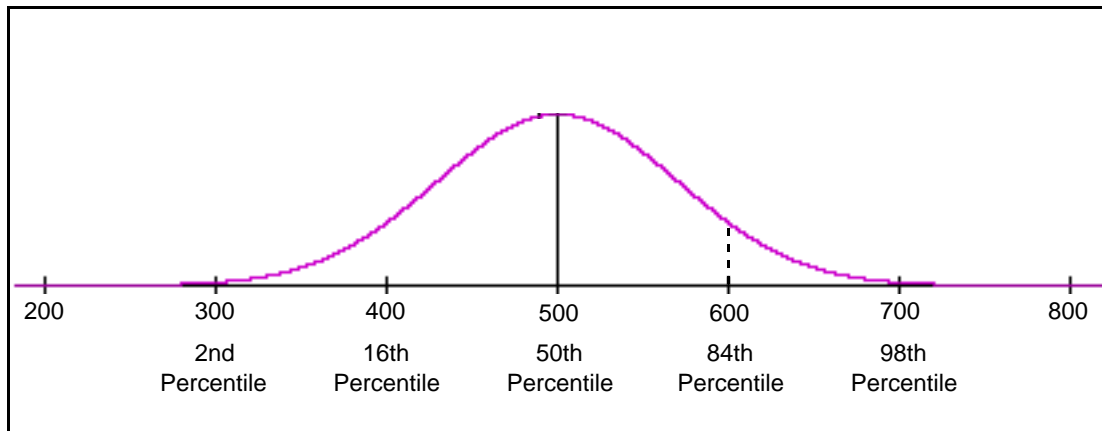


Figure 5. SAT-I Mathematics test score distribution

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

The creators of tests such as the SAT base their work on the assumption that students x and y each possess a certain amount of knowledge, ability, or (in the case of 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 be accurate measures of their knowledge, ability or potential. However, constructing and administering such a test is impossible. Instead, the designers create a test made up of questions that are, in effect, a random sample drawn from the universe of all possible questions.

Like the results of any survey based on a sample drawn from some population, the actual scores 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 30 points. Student x ’s actual score of 470, combined with this measurement error, tells us we can be 95 percent sure that her true score is somewhere between 410 and 530, an interval that extends 60 points (i.e. two “standard errors”) on either side of the actual score.

The Myth of Objectivity

Similarly, student y 's true score is, with 95 percent certainty, between 470 and 590. (See fig. 6.)

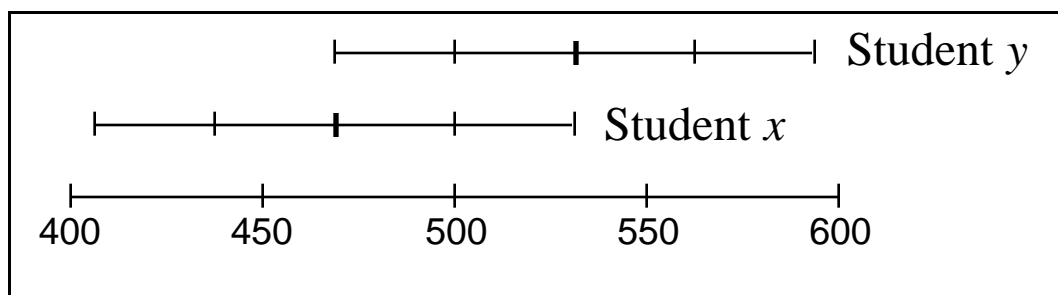


Figure 6. Interval estimates of two SAT scores.

These confidence intervals overlap; students x and y would need actual scores that differed by at least 84 points for us to be 95 percent sure that their true scores were different. Because their actual scores differ by only (only!) 60 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 provided 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 and 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. In this case, 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” (“What’s Wrong with Standardized Tests?” 1996), and when designers assume that at any point in time each student possesses a certain amount of knowledge, ability, or potential that can be measured (with some measurement error) by a single instrument. This is only one way to conceptualize knowledge, ability or potential. If knowledge is multifaceted and complex, individually constructed and inextricably tied to the context in which the learning occurs—as over 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-upon subjectivity.” Here are two specific

The Myth of Objectivity

suggestions for improving the consistency and usefulness of assessment information gathered by teachers.

1. Design classroom assessment tasks that are likely to elicit from students the information you seek. Ask yourself the questions, what is the mathematics I am trying to assess here? and what task(s) will tap this mathematics most directly? An Algebra II teacher might want to know what students understand about quadratic equations and the techniques for solving them. The question in the example discussed above—consisting of the one-word imperative “solve”—does not directly ask students to provide much information about their understanding. The set of tasks in figure 7, for example, do so more specifically. Greer et al. (1999) offer guidelines for creating and adapting tasks for classroom assessment.

(a) Use one of the symbolic methods developed in class to find solutions to the equation

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

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

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

(d) It is possible to find no real solutions to a quadratic equation. Explain how this could happen. Provide an example that illustrates your explanation.

Figure 7. A revised quadratic equation task.

2. Before any task is given to students, devise—and share with the students—guidelines for scoring their work. (See Thompson and Senk 1998; Greer et al. 1999.) Ask yourself, what types of responses am I likely to get from students to these tasks? and what will I accept as evidence of adequate understanding? Thinking through these questions before the tasks are given to students helps clarify the tasks themselves. It also helps align the tasks with the instruction that went on in class. Sharing these guidelines with students communicates expectations and makes it more likely that they will be met. One set of guidelines for scoring student work on the tasks in figure 7 is proposed in figure 8.

Conclusion

False dichotomies like “objective vs. subjective” and “traditional vs. alternative” derail meaningful discussions of the important issues in mathematics assessment. The labels “traditional” and “alternative” are meaningless; the five-question classroom quiz can provide detailed information about what students know, or provide very little information, depending on how it is designed, scored and used. There is no “objective” assessment; subjective—that is, human—knowledge, beliefs, judgments and decisions are an unavoidable part of any assessment scheme. It is more productive to consider ways to make assessment of students’ mathematical understanding, and the information gathered through that assessment, more consistent and useful.

<p>5 – All of 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.</p> <p>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.</p> <p>3 – Substantial evidence of understanding of quadratic equations: some minor errors (not central to understanding quadratic equations) are all that is missing from the characteristics of a 4.</p> <p>-----</p> <p>2 – Some evidence of understanding of quadratic equations is present: either a symbolic solution, or a graphical illustration, with perhaps some minor errors.</p> <p>1 – Little understanding of quadratic equations is shown: major errors in all parts of the problem.</p> <p>0 – No attempt made.</p>
--

Figure 8. A scoring rubric for the revised task.

References

- Battista, Michael T. "The Mathematical Miseducation of America's Youth: Ignoring Research and Scientific Study in Education." *Phi Delta Kappan* 80 (February 1999): 424 – 433.
- Davis, Robert B., Carolyn A. Maher and Nel Noddings, eds. *Constructivist Views on the Teaching and Learning of Mathematics: JRME Monograph Number 4*. Reston, Va.: National Council of Teachers of Mathematics, 1990.
- 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, ed. William S. Bush and Jean Kerr Stenmark. Reston, Va.: National Council of Teachers of Mathematics, 1999.
- Hoover, H. D., A. N. Hieronymus, D. A. Frisbie and S. B. Dunbar. *Iowa Test of Basic Skills: Norms and Score Conversions, with Technical Information*. Itasca, IL: Riverside Publishing, 1996.
- National Council of Teachers of Mathematics. *Principles and Standards for School Mathematics*. Reston, Va.: National Council of Teachers of Mathematics, 2000.

The Myth of Objectivity

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

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 Senk. "Using Rubrics in High School Mathematics Courses." *Mathematics Teacher* 91 (December 1998): 786 – 793.

"What's Wrong with Standardized Tests?" Hp. 16 June 1999 [Last Update]. FairTest: The National Center for Fair and Open Testing. Available: www.fairtest.org/facts/whatwron.htm. 18 June 1999.

Appendix

Computation of 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 we could administer the test 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? (If we were talking about the reliability of a thermometer, we would ask, How close will thermometer readings be to the actual temperature, and how consistently will the thermometer produce 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 $r_{XX} = 1$. However, no test is perfectly reliable. Repeated administrations of a test (if you could do this) would produce a set of scores for a particular student that would be distributed around the student's true score. (See fig. A-1.) The more highly reliable the test, the higher the test-retest correlation, and the tighter the distribution of scores. For the very reliable SAT-I mathematics test, $r_{XX} = 0.91$.

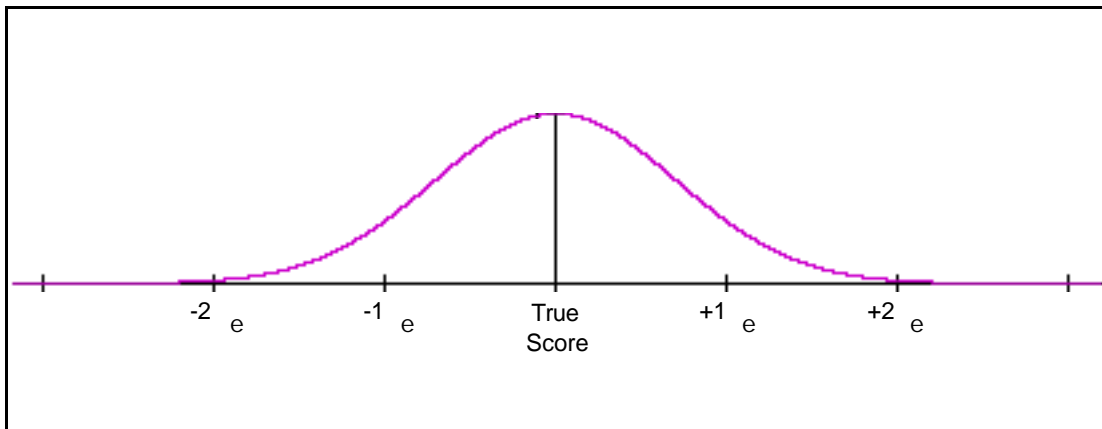


Figure A-1. The distribution of actual scores around a student's true score.

The standard deviation of this distribution of scores is the “Standard Error of Measurement,” e . It can be calculated, using the standard deviation of the test scores, s_x , and the test's reliability, r_{XX} , using:

$$e = s_x \sqrt{1 - r_{XX}}$$

In the case of the SAT-I Mathematics test, $s_x = 100$ and $r_{XX} = 0.91$, so

$$e = 100\sqrt{1 - 0.91} = 100\sqrt{0.09} = 100(0.3) = 30 \text{ points.}$$

Therefore, for this student 68 percent of the scores she would earn if she were to take the test repeatedly would be within 30 points on either side of her true score. Similarly, 95 percent of her scores would be within 60 points on either side of her true score.

Now imagine that we could administer this test repeatedly to two different students. If we were to compute the difference between these two students' scores every time the test

The Myth of Objectivity

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:

$$\sigma_{X-Y}^2 = \sigma_X^2 + \sigma_Y^2$$

In the case of the SAT-I Mathematics test, $\sigma_X^2 = \sigma_Y^2 = \sigma_e^2$, so $\sigma_{X-Y}^2 = \sigma_e^2 + \sigma_e^2 = 2\sigma_e^2$. Therefore, the “standard error of the difference” between two SAT-I Mathematics test scores is

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

To be 95 percent sure that two actual scores represent different true scores, the actual scores would have to differ by at least 84 points.