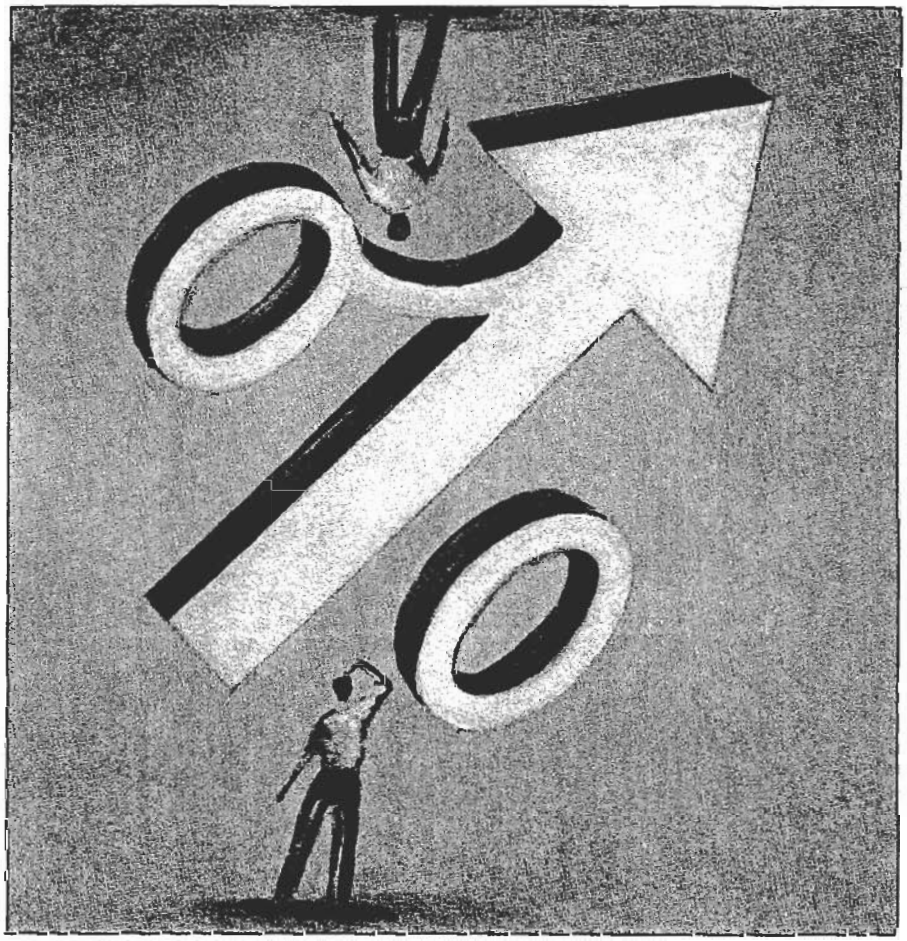
High Percentages Are *Not* The Same as High Standards

Mr. Guskey reminds us that, even when complex statistical formulas are used in setting cutoff scores, their mathematical precision is not a substitute for sound professional judgment.

BY THOMAS R. GUSKEY



HOW TO set appropriate cutoff scores for student performance on state assessments and other high-stakes examinations is a widely debated issue in education today. Typically these debates focus on what percentage of items students should be expected to answer correctly in order to have their performance judged "proficient" or "competent." On the Texas Assessment of Academic Skills (TAAS), for example, students must answer 70% correct in order to attain a passing score. This debate often extends to the classroom level, where teachers set cutoff scores for different grades. What percentage correct should students be expected to attain, for instance, to receive a grade of A or a grade of B, and so on?

Both policy makers and teachers generally assume that higher cutoff percentages mean more rigorous standards and higher expectations for student performance. A cutoff of 80% correct for proficiency in mathematics, for instance, is considered more rigorous than a 70% correct cutoff for proficiency in language arts. Similarly, the teacher who sets 95% correct as the cutoff for a grade of A is considered to be more demanding and to have higher standards than the teacher who uses a cutoff of only 90% or 92% correct for an A. This reasoning leads to the belief that raising the percentage for a cutoff is one way to raise both the standards and the expectations we set for student performance.

Unfortunately, it isn't quite that simple. Setting cutoff percentages for assessments and for grades is an arbitrary decision that says little about the standards or the expectations set for students' learning. A much more important consideration is the difficulty of the tasks students are asked

THOMAS R. GUSKEY is a professor of Educational Policy Studies and Evaluation, College of Education, University of Kentucky, Lexington.

Illustration by Karen Stolper

to perform or the cognitive complexity of the questions they are required to answer.

The cutoff percentage representing an excellent level of performance on an extremely challenging task or a very difficult set of questions might be quite different from the cutoff percentage considered excellent on a relatively simple task. This does not imply that the challenge is determined strictly by how well other students perform (i.e., norm-referenced). Rather, it means that tasks or items designed to assess a given learning goal (i.e., criterion-referenced) can vary widely in their intricacy and cognitive complexity.

Suppose, for example, we were interested in assessing students' basic knowledge about the Presidents of the United States. We could ask an open-ended, constructed-response question (also known as a "short-answer" or "completion" item):

> Who was the 17th President of the United States?[1]

Fewer than 10% of students are able to answer this question correctly. Its high level of difficulty is actually rather odd because most people know that Abraham Lincoln was the 16th President, and they know that the name of the President who succeeded him was Johnson. Putting these two pieces of information together, however, proves quite difficult for the vast majority.

We might then consider framing the same question as a multiple-choice, selected-response item. For example:

> Who was the 17th President of the United States?
> A. Abraham Lincoln
> B. Andrew Johnson
> C. Ulysses S. Grant
> D. Millard Fillmore

This remains a fairly difficult item for most students. Because of the multiple-choice format, however, about 30% are now able to answer correctly. Of course, if all students simply chose an answer at random, the limited-response, multiple-choice format would allow 25% to select the correct response.

Suppose we next adjust the possible responses, making the distinctions a bit more obvious:

> Who was the 17th President of the United States?
> A. George Washington
> B. Andrew Johnson
> C. Jimmy Carter
> D. Bill Clinton

Now identifying the correct response is much easier, and about 60% of students are able to answer correctly. We could probably assume that those who are still unable to identify the correct response have very limited knowledge of U.S. Presidents.

Of course, we could make a final adjustment to the possible responses in order to make the item easier still:

> Who was the 17th President of the United States?
> A. The War of 1812
> B. Andrew Johnson
> C. The Louisiana Purchase
> D. A Crazy Day for Sally

About 90% of students are able to answer this item correctly. Those who don't are usually drawn to the response "A Crazy Day for Sally" because they recognize it as the one response that doesn't belong with the others.

Some might argue that knowing who was the 17th President of the United States is a rather trivial learning outcome — and that might be true. The point is that, while each of these items assesses the same learning objective, same goal, or same achievement target, each varies greatly in its difficulty.

Suppose that there were four assessments designed to measure students' subject-area proficiency or their achievement in a high school course. Assessment 1 consisted of items of the first type described above; assessment 2 consisted of items of the second type, and so on. Those four assessment devices would present vastly different challenges to students, and the scores students attained on such assessments would undoubtedly reflect those differences. Would it be fair to set the same "proficiency" cutoff percentage for each of those four assessments?

The Challenge of Setting Appropriate Cutoffs

Focusing on a percentage correct as a

"Basically, what you're saying is I get a box of chocolate chip cookies, and the sixth-grade class gets a field trip to Tuscany?"

cutoff is seductive but very misleading because tests and assessments vary widely in how they are designed. Some assessments include items that are so challenging that students who answer a low percentage of items correctly still do very well.

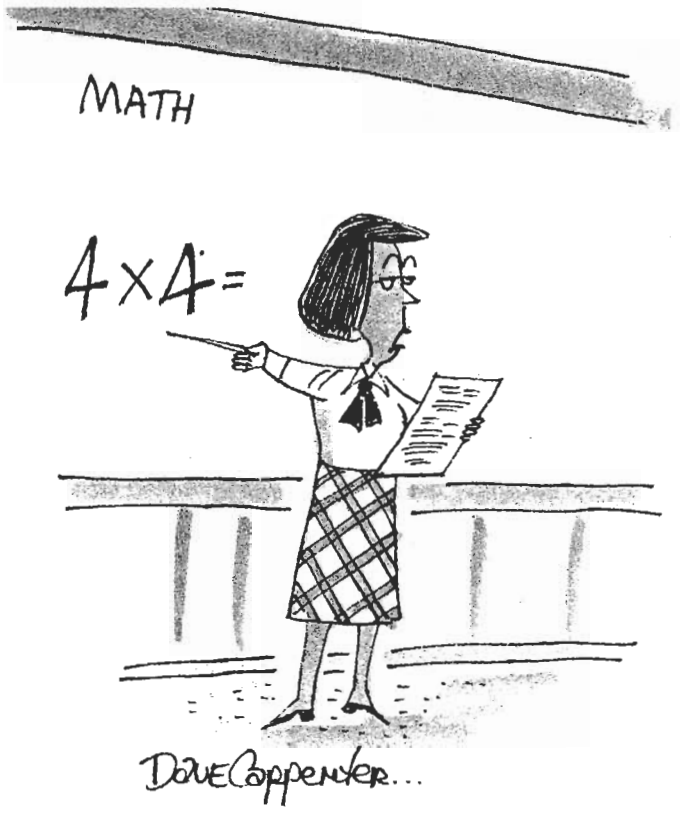
Take the Graduate Record Examinations (GRE), for example, a series of tests used to determine admission to graduate schools. Individuals who answer only 50% of the questions correctly on the GRE physics test perform better than more than 70% of those who take the test (already a highly self-selected group). For the GRE mathematics test, 50% correct would outperform approximately 60% of the individuals who take the test. And among those who take the GRE literature test, only about half get 50% correct.[2] In most classrooms, of course, students who answer only 50% correct would receive a failing grade.

Should we conclude from this information that prospective graduate students in physics, mathematics, and literature are a bunch of "failures"? Of course not. Without careful examination of the questions or tasks students are asked to address, cutoff percentages are just not that meaningful.

Researchers suggest that an appropriate approach to setting cutoffs must combine teachers' judgments of the importance of the concepts addressed and consideration of the cognitive processing skills required by the items or tasks.[3] Using this type of cutoff or grade-assignment procedure shifts teachers' thinking so that grades on classroom assessments and other demonstrations of learning reflect the quality of student thinking instead of simply the number of points attained. It incorporates the value the teacher places on successful performance and the teacher's perception of the level of thinking that students must use to answer a question or perform a task.

Sadly, this ideal is seldom realized. Rarely does such thought and consideration go into setting the cutoffs for students' performance or the grades they receive. Even in high-stakes assessment situations in which the consequences for students can be quite serious, this level of deliberative judgment is uncommon.

Making matters even more complicated is the fact that the challenge or difficulty of an assessment task is also directly related to the quality of the teaching. Students who are taught well and provided ample opportunities to practice and demonstrate what they have learned are likely to find well-aligned performance tasks or assessment questions much easier than students who are taught poorly and given few practice opportunities. Hence, a 90% cutoff might be relatively easy to meet for students who are taught well, while a 70% cutoff might prove exceptionally difficult for those students who experience poor-quality teaching.



"No, 'off-road vehicle' is incorrect."

Conclusion

The point of this discussion is not that cutoff percentages are unimportant. They are a vital and necessary consideration in any assessment of student learning. However, setting cutoffs is a more complex process than most policy makers and educators anticipate, and it is typically much more arbitrary than most imagine.[4]

What we must keep in mind is that, even when complex statistical formulas are used in setting cutoffs, their mathematical precision is not a substitute for sound professional judgment. Raising standards or increasing expectations for students' learning is not accomplished simply by raising the cutoff percentages for performance levels or different grade categories. Raising standards requires thoughtful examination of the tasks students are asked to complete and the questions they are asked to answer in order to demonstrate their learning. It might also involve taking into account the quality of the teaching students experienced prior to the assessment. Only when such judgment becomes a regular part of the assessment process will we be able to make accurate and valid decisions about the quality of students' performance.

1. This item was developed by Professor Jeffrey Smith of Rutgers University.

2. Drew H. Gitomer and Mari A. Pearlman, "Are Teacher Licensing Tests Too Easy? Are Standards Too Low?," *ETS Developments*, vol. 45, 1999, pp. 4-5.

3. Anthony J. Nitko and Boleslaw Niemierko, "Qualitative Letter Grade Standards for Teacher-Made Summative Classroom Assessments," paper presented at the annual meeting of the American Educational Research Association, Atlanta, 1993.

4. Thomas R. Guskey and Jane M. Bailey, *Developing Grading and Reporting Systems for Student Learning* (Thousand Oaks, Calif.: Corwin, 2001). K