The Case Against

It's time to abandon grading scales that distort the accuracy, objectivity, and reliability of students' grades.

Thomas R. Guskey

DAVE CUTLLER © IMAGES.COM/CORBIS

Assessment and grading have become a major focus in education reform. But one basic component of most present-day grading systems stands as a major impediment to making grades fairer, more accurate, and more meaningful. That component is percentage grades.

Percentage grades are the foundation of many state grading policies. Nearly every online grading program available to educators calculates percentage grades. Yet despite their popularity, percentage grades are difficult to defend from a procedural, practical, or ethical perspective.

A Brief History

Before 1850, grading and reporting were virtually unknown in U.S. schools. Most schools grouped students of all ages and backgrounds together with one teacher in a one-room schoolhouse, and few students went beyond the elementary level. The teacher commonly reported students' learning progress orally to parents during visits to students' homes.

As enrollments increased in the late 1800s, however, schools began to group students in grade levels according to age (Edwards & Richey, 1947) and to use formal progress evaluations. In most cases, these were narrative reports in which teachers described the skills each student had mastered and those on which additional work was needed. The main purpose of such reports was to inform students when they had demonstrated mastery of the current performance level and were ready to move on to the next level.

With the passage of compulsory school attendance laws in the late 19th and early 20th centuries, high school enrollments increased rapidly. Between 1870 and 1910, the number of public high schools in the United States rose from 500 to 10,000 (Gutek, 1986). Subject-area instruction became increasingly specific, and student populations became more diverse. Although elementary teachers continued to use narrative reports to document student learning, high school teachers began using percentages and other similar markings to certify accomplishment in different subject areas (Kirschenbaum, Simon, & Napier, 1971).

The shift to percentage grades was gradual, and few U.S. educators questioned it. The practice seemed a natural result of the increased demands on high school teachers, who now served growing numbers of students.

But in 1912, a study by two Wisconsin researchers seriously challenged the reliability and accuracy of percentage

Percentage Grades

grades. Daniel Starch and Edward Charles Elliott found that 147 high school English teachers in different schools assigned widely different percentage grades to two identical student papers. Scores on the first paper ranged from 64 to 98, and scores on the second paper ranged from 50 to 97. One paper was given a failing mark by 15 percent of the teachers and a grade of over 90 by 12 percent of the teachers. Some teachers focused on elements of grammar, style, neatness, spelling, and punctuation, whereas others considered only how well the paper communicated its message. With more than 30 different percentage grades assigned to a single paper and a range of more than 40 points, it is easy to see why this study created a stir among educators.

Starch and Elliott's study was immediately criticized by those who claimed that judging good writing is, after all, highly subjective. But when the researchers repeated their study using geometry papers graded by 128 math teachers, they found even greater variation. Scores assigned by teachers to one of the math papers ranged from 28 to 95 percent. Some of the teachers deducted points only for a wrong answer. Others gave students varying amounts of partial credit for their work. Still others considered neatness, form, and spelling in the grades they assigned (Starch & Elliott, 1913).

These demonstrations of wide variation in grading practices among teachers led to a gradual move away from percentage grades to scales that had fewer and larger categories. One was a three-point scale that employed the categories *Excellent*, *Average*, and *Poor*. Another was the familiar five-point scale of *Excellent*, *Good*, *Average*, *Poor*, and *Failing*, or *A, B, C, D,* and *F* (Johnson, 1918; Rugg, 1918). This decrease in the number of score categories led to greater consistency across teachers in the grades assigned to student performance.

The resurgence of percentage grades **appears to come** mainly **from the increased use of technology.**

A Modern Resurgence

Percentage grades continued to be relatively rare in U.S. schools until the early 1990s, when grading software and online grade books began to gain popularity among educators. Today, schools can choose from more than 50 electronic grading software programs (see www.gradebooks4teachers.com). Because these programs are developed primarily by computer technicians and software engineers rather than educators, they incorporate scales that appeal to technicians—specifically, percentages.

Like monetary systems based on the dollar, percentages have 100 levels that are easy to divide into increments of halves, quarters, and tenths. Percentages are also easy to calculate and easy for most people to understand. Thus, the resurgence of percentage grades appears to come mainly from the increased use of technology and the partialities of computer technicians, not from the desire of educators for alternative grading scales or from research about better grading practice.

Modern percentage grading scales differ significantly, however, from those that were used in the past. The 100-point scale that teachers employed in the early 20th century was based on an average grade of 50, and grades above 75 or below 25 were rare (Smallwood, 1935). In contrast, most modern applications of percentage grades set the average grade at 75 (which translates to a letter grade of *C*) and establish 60 or 65 as the minimum threshold for passing. This practice dramatically increases the likelihood of a negatively skewed grade distribution that is "heavily gamed against the student" (Carey & Carifio, 2012, p. 201).

Ironically, neither this narrower grade distribution nor a century of research and experience in scoring students' writing seems to have improved the reliability of the percentage grades assigned by teachers. Recently, Hunter Brimi (2011) replicated Starch and Elliott's 1912 study and attained almost identical results. Brimi asked 90 high school teachers—who had received nearly 20 hours of training in a writing assessment program—to grade the same student paper on a 100-point percentage scale. Among the 73 teachers who responded, scores

ranged from 50 to 96. And that's among teachers who received specific professional development in writing assessment!

So even if one accepts the idea that there are truly 100 discernible levels of student writing performance, it's clear that even well-trained teachers cannot distinguish among those different levels with much accuracy or consistency.

Problems with Percentage Grades

Logistics

From the perspective of simple logic, percentage grading scales make little sense. As noted earlier, teachers who use percentage grades typically set the minimum passing grade at 60 or 65. The result is a scale that identifies 60 or more distinct levels of failure and only 40 levels of success. In other words, nearly two-thirds of the percentage grading scale describes levels of failure! What message does that communicate to students?

And distinguishing 60 different levels of failure is hardly helpful. Does any teacher consider percentage grades in the 50s to denote modest failure and those in the teens or 20s to represent extreme failure? Are unsuccessful students concerned about which of the 60 different levels of failure they achieved?

Some teachers counter that no one really uses those 60 different levels of failure. But if that is the case, then why have them? Why not use a 50-point grading scale and designate ten levels of failure rather than the 100-point percentage grading scale with 60 levels of failure? After all, the choice of 100 is quite arbitrary.

A grading scale in which two-thirds of the designated levels describe failure also implies that degrees of failure can be more finely distinguished than degrees of success. Should the focus of educators be to determine more minutely different levels of failure than those of learning success?

Accuracy

The accuracy of any measure depends on the precision of the measurement instrument. A sophisticated stopwatch, for example, can very accurately measure the time an individual takes to run a 100-meter race. The instruments we use to measure student learning, however, are far less accurate and precise.

Measurement experts identify precision by calculating the *standard error of measurement.* This statistic describes the amount by which a measure might vary from one occasion to the next using the same device to measure the same trait. For example, suppose the standard error on a 20-item assessment of student learning is plus or minus two items. That may not seem like much, but using a percentage grading scale, that would be a range of 20 percentage points—a difference in most cases of at least two letter grades.

Many educators assume that because the percentage grading scale has 100 classification levels—or categories—it is more precise than a scale with just a few levels (such as *Excellent, Average,* and *Poor*). But in the absence of a truly accurate measuring device, adding more gradations to the measurement scale offers only the illusion of precision. When assigning students to grade categories, statistical error relates to the number of misclassifications. Setting more cutoff boundaries (levels or categories) in a distribution of scores means that more cases will be vulnerable to fluctuations across those boundaries and, hence, to more statistical error (Dwyer, 1996). A student is statistically much more likely to be misclassified as performing at the 85-percent level when his true achievement is at the 90-percent level (a difference of five percentage categories) than he is of being misclassified as scoring at an *Average* level when his true achievement is at an *Excellent* level. In other words, with more levels, more students are likely to be misclassified in terms of their performance on a particular assessment.

Overall, the large number of grade categories in the percentage grading scale and the fine discrimination required in determining the differences among categories allow for the greater influence of subjectivity, more error, and diminished reliability. The increased precision of percentage grades is truly far more imaginary than real.

Percentage Grades vs. Percentage Correct

Percentage grades are often directly derived from the percentage of items a student answers correctly on an assessment; this, in turn, is assumed to reflect the percentage of the content the student has learned or the percentage of the skills the student has mastered. Because assessments of student performance vary widely in their design, however, this assumption is rarely true. Some assessments include items or problems that are so challenging that even students who have mastered the essential content and skills still answer a low percentage of the items correctly.

Take, for example, the Graduate Record Examinations (GRE), a series of assessments used to determine admission to many graduate schools. Individuals who answer only 50 percent of the questions correctly

on the GRE physics exam perform better than more than 70 percent of all individuals who take the exam. For the GRE mathematics exam, a person answering 50 percent correctly would outperform approximately 60 percent of the individuals who take the exam. And among those who take the GRE literature exam, only about half get 50 percent correct (Gitomer & Pearlman, 1999). In most classrooms, of course, students who answer only 50 percent correctly would receive a failing grade.

Should we conclude from this information that majorities of prospective graduate students in physics, mathematics, and literature are "failures"? Does it mean that most of those interested in doing advanced graduate work in these subjects have learned little of the important content and skills in their respective disciplines? Of course not. Percentage grades derived solely from the percentage correct, without careful examination of the items or tasks students are asked to address, are just not all 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 assessment items or tasks (Nitko & Niemierko, 1993). Sadly, this ideal is seldom realized. Even in high-stakes assessment situations, where the consequences for students can be quite serious, policymakers rarely put this level of deliberative judgment into setting the cutoff scores for student performance.

Further, the challenge or difficulty of an assessment is also related to the quality of the teaching that students experience. Students who are taught well and provided ample opportunities to practice and demonstrate what they have learned typically find well-aligned performance tasks or assessment questions much easier than do students who are taught poorly and given few practice opportunities. Hence, a percentage score of 90 might be easy for well-taught students to attain, whereas attaining a score of 70 might prove exceptionally difficult for poorly taught students. Multiple factors influence students' performance, many lying outside students' control (Guskey & Bailey, 2001).

Distinguishing 60 different levels of failure is hardly helpful.

The Distortion of the Zero

In recent years, much ado has been made about legislation passed in several states that bars school districts from stipulating that the lowest percentage grade teachers can assign to students is 50 rather than zero (Montgomery, 2009; Peters, 2009; Richmond, 2008). School districts that enact these minimum-grade policies have no intention of giving students credit when no credit is due. A percentage grade of 50 is still a failing grade in nearly every school. In addition, although some have suggested that minimum-grade policies promote grade inflation and social promotion in schools, well-designed, longitudinal studies show this is not the case (Carey & Carifio, 2012; Carifio & Carey, 2010). Rather, school districts implement minimum-grade policies simply to eliminate the confounding effects of a zero in a percentage grading system.

When combined with the common practice of grade averaging, a single zero can have a devastating effect on a student's percentage grade. The student's overall course grade is unfairly skewed by that one, atypical low score. To recover from a single zero in a percentage grade system, a student must achieve a perfect score on a minimum of nine other assignments. Attaining that level of performance would challenge the most talented students and may be impossible for struggling learners. A single zero can doom a student to failure, regardless of what dedicated effort or level of performance might follow (Guskey, 2004).

Certainly, students need to know that there are consequences for what they do and do not do in school. Irresponsible actions and malingering should be penalized. But should the penalty be so severe that students have virtually no chance of recovery?

The true culprit in this matter, however, is not minimum grades or the zero—it's the percentage grading system. In a percentage grading system, a zero is the most extreme score a teacher can assign. To move from a *B* to an *A* in most schools that use percentage grades requires improving only 10 percentage points at most—say, from 84 to 94 percent. But to move from a zero to a minimum passing grade requires six or seven times that improvement, usually from zero to 60 or 65.

If the purpose of grading is to communicate information about how well students have learned and what they have accomplished in school, the grading system should not punish students in ways that make recovery from failure impossible. In a percentage grading system, assigning a grade of zero does exactly that.

What's the Alternative?

Rather than argue about minimum grades or zeros, an easy solution to this dilemma is to do away with percentage grades and use an integer grading system of 0–4 instead. In such a system, improving from a failing grade to a passing grade means moving from 0 to 1, not from 0 to 60 or 65. An integer system makes recovery

possible for students. It also helps make grades more accurate reflections of what students have learned and accomplished in school.

Educators at all levels are familiar with integer grades. The majority of colleges and universities in the United States use integer grading systems, and most high schools use integer grades when they compute students' grade-point averages (GPAs). In fact, using 0–4 integer grades would eliminate the problems that many high schools experience in trying to convert percentage grades to four-point or five-point GPAs. And integer grading scales align with the levels used to classify student achievement in most state assessment programs (for example, *Below Basic, Basic, Proficient,* and *Advanced*) and with the four-point rubrics that many teachers use in judging students' performance on classroom assessments.

The use of integer grading systems will result in grades that are more meaningful and reliable. With modest training and experience, different teachers considering a specific collection of evidence of student learning can generally reach consensus about the 0–4 integer grade that evidence represents. Integer grades do not necessarily make grading easier; they simply make the process more accurate and honest.

No Substitute for Professional Judgment

Percentage grading systems that attempt to identify 100 distinct levels of performance distort the precision, objectivity, and reliability of grades. They also create unsolvable methodological and logistical problems for teachers. Limiting the number of grade categories to four or five through an integer grading system allows educators to offer more honest, sensible, and reliable evaluations of students' performance. Combining the grade with supplemental narrative descriptions or standards checklists describing the learning criteria used to determine the grade further enhances its communicative value.

A single zero can have a devastating effect on a student's percentage grade.

Assigning fair and meaningful grades to students will continue to challenge educators at every level. The process requires thoughtful and informed professional judgment, an abiding concern for what best serves the interests of students and their families, and careful examination of the tasks students are asked to complete and the questions they are asked to answer to demonstrate their learning. Only when such examination and reasoned judgment become a regular part of the grading process can we make accurate and valid decisions about the quality of students' performance. EL

References

Brimi, H. M. (2011). Reliability of grading high school work in English. *Practical Assessment, Research and Evaluation*, *16*(17), 1–12.

Carey, T., & Carifio, J. (2012). The minimum grading controversy: Results of a quantitative study of seven years of grading data from an urban high school. *Educational Researcher*, *41*(6), 201–208.

Carifio, J., & Carey, T. (2010). Do minimum grading practices lower academic standards and produce social promotion? *Educational Horizons*, *88*(4), 219–230.

Dwyer, C. A. (1996). Cut scores and testing: Statistics, judgment, truth, and error. *Psychological Assessment*, *8*(4), 360–362.

Edwards, N., & Richey, H. G. (1947). *The school in the American social order*. Cambridge, MA: Houghton Mifflin.

Gitomer, D. H., & Pearlman, M. A. (1999). Are teacher licensing tests too easy? Are standards too low? *ETS Developments*, *45*(1), 4–5.

Guskey, T. R. (2004). Zero alternatives. *Principal Leadership*, *5*(2), 49–53.

Guskey, T. R., & Bailey, J. M. (2001). *Developing grading and reporting systems for student learning*. Thousand Oaks, CA: Corwin.

Gutek, G. L. (1986). *Education in the United States: An historical perspective*. Englewood Cliffs, NJ: Prentice Hall.

Johnson, R. H. (1918). Educational research and statistics: The coefficient marking system. *School and Society*, *7*(181), 714–716.

Kirschenbaum, H., Simon, S. B., & Napier, R. W. (1971). *Wad-ja-get? The grading game in American education*. New York: Hart.

Montgomery, D. (2009, November 19). Half-dozen districts sue over Texas law prohibiting minimum grades. *Dallas Star Telegram*.

Nitko, A. J., & Niemierko, B. (1993, April). *Qualitative letter grade standards for teacher-made summative classroom assessments*. Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.

Peters, E. (2009, October 20). Minimum grades no longer allowed. *Abilene Reporter-News,* Retrieved from www .reporternews.com/news/2009/oct/20/minimum-grades

Richmond, E. (2008, February 4). A floor for failing grades: Parents, educators debate whether kids should get 50 points for doing nothing. *Las Vegas Sun*.

Rugg, H. O. (1918). Teachers' marks and the reconstruction of the marking system. *Elementary School Journal*, *18*(9), 701–719.

Smallwood, M. L. (1935). *An historical study of examinations and grading systems in early American universities*. Cambridge, MA: Harvard University Press.

Starch, D., & Elliott, E. C. (1912). Reliability of the grading of high school work in English. *School Review*, *20*, 442–457.

Starch, D., & Elliott, E. C. (1913). Reliability of the grading of high school work in mathematics. *School Review*, *21*, 254–259.

Thomas R. Guskey (Guskey@uky.edu) is professor, Department of Educational, School, and Counseling Psychology, College of Education, University of Kentucky, Lexington.