EL EDUCATIONAL LEADERSHIP

April 2016 | Volume **73** | Number **7**

Looking at Student Work | Pages 50-54

Grading: Why You Should Trust Your Judgment

Thomas R. Guskey and Lee Ann Jung

Although computerized grading programs have advantages, teachers' judgment has been shown to be more reliable.

Ask teachers today to describe a student's learning progress, and most will begin by opening their computerized gradebook. The teacher will look over the student's scores, and then skip to the summary grade. The gradebook typically allows the teacher to attach various weights to different assignments and assessments in calculating the summary grade, and it may also sort scores according to specific learning targets or standards. The teacher will explain how the grading program precisely computes the summary grade in the same way for all students and records that grade on a report card that is shared with parents at the end of the marking period.

Computerized grading programs are ubiquitous in modern education. They rank among the best-selling computer software in education, with more than 40 programs currently available.[1] They appeal to teachers because they simplify record keeping and seem to objectify grading. Their data management capabilities make it easy for teachers to enter and precisely tally large amounts of numerical information on students' performance. They are particularly well-suited to the needs of middle and high school teachers, who generally record data on the learning progress of more than 100 students weekly (Guskey, 2002).

Despite their many advantages, however, computerized grading programs also have drawbacks. In particular, their pervasive use has caused teachers to doubt their professional judgment. Instead of looking carefully at the array of data on students' performance and making thoughtful decisions about what grade best describes what students have achieved, teachers rely on the grading program's statistical algorithms to calculate grades. In teachers' minds, these dispassionate mathematical calculations make grades fairer and more objective. Explaining grades to students, parents, or school leaders involves simply "doing the math." Doubting their own professional judgment, teachers often believe that grades calculated from statistical algorithms are more accurate and more reliable.

Computer-Generated Grades: More Accurate?

But *are* the grades that are determined by computerized grading programs fairer? Are they truly more objective than those based on teachers' professional judgment? Are they more accurate and reliable?

We frequently test this idea by asking groups of teachers to consider the data in Figure 1. These data represent a particular student's scores from six assessments of learning during a grading period. The top row shows the date of each assessment, and the bottom row shows the student's scores on the assessments (derived from a well-designed rubric). A score of 1 represents the lowest level of performance; 4 represents the highest.

To determine what summary grade to record, teachers generally combine scores from multiple sources of evidence gathered over time. So in our research, we ask groups of teachers, Given the scores shown here and Gloria's pattern of performance, what summary grade should she receive for this learning target? We ask them to *first* answer this question by using their professional judgment—simply looking at this pattern of scores and deciding whether Gloria deserves a grade of 1, 2, 3, or 4—before turning to a statistical algorithm.

Typically, 80 percent or more of the teachers at all grade levels agree that Gloria should receive a summary grade of 4. Although she struggled during the first part of the grading period, Gloria's recent performance clearly reflects that she has mastered this learning target.

Next, we show teachers the summary grade that would result using a computer-generated algorithm. Computerized grading programs typically offer the choice of several statistical algorithms for determining a student's summary grade. The most common algorithms include the mean (the average score); median (the middle score); mode (the most frequently appearing score); and the trend score (a score pattern analysis). Although each option computes the summary grade in an impersonal, objective way, the choice of which algorithm to use is highly subjective and could yield widely divergent results.

The default algorithm in most computerized grading programs is the mean, or average, score. If the teacher chooses this method, Gloria will receive a summary grade of 2. If the teacher selects the median or mode score, Gloria's grade will be 1; if the teacher chooses the trend score, it will be 2.7 or, rounded up, 3 (Marzano, 2000). So depending on the statistical algorithm chosen, Gloria could receive a summary grade of 1, 2, or 3. No algorithm would result in a grade of 4.

What About Reliability?

Reliability is an index of consistency in measures or responses and a necessary prerequisite for validity. Unreliable measures can never be valid. Calculations of reliability range in value from 0.0 to 1.0. Researchers generally consider .8 as a minimal level of reliability in measures that have important consequences for students, such as grades.

Researchers have several ways of computing reliability. In situations like calculating a student's grade, researchers would be most concerned with *inter-rater reliability*, the degree to which equally knowledgeable and competent judges or raters—in this case, teachers—can look at the same evidence and consistently make the same decision regarding a summary grade. If raters consistently come to the same decision, the summary grade would be considered a reliable measure.

Consider several teachers looking at the data in Figure 1. If all teachers used the same statistical algorithm, all would assign Gloria the same summary grade, and the grade would be considered highly reliable. But if teachers varied in their choice of statistical algorithms, the resulting summary grades would vary, with some 1s, some 2s, and some 3s. Because of this variability, the summary grade would be considered an unreliable indicator of Gloria's true performance. So even though each algorithm would yield a precise grade, differences among teachers in their choice of algorithm would make that grade unreliable.

Figure 1. A student's scores from six assessments of a learning target.

Student	Learning Target #1						
	9/9	9/14	9/22	9/27	10/3	10/6	Summary Grade
Gloria	1	1	1	1	4	4	?

Consider the Purpose

Let's suppose that instead of relying on a computerized grading program's statistical algorithm, teachers got together and determined the purpose of a summary grade. And suppose that after considering different points of view, they reached consensus that the purpose of the summary grade is, "To best describe the student's level of proficiency regarding the learning target at this time."

We have asked the same group of teachers to determine what summary grade Gloria should receive for the learning target, using their judgment and keeping this purpose in mind. We remind them that they know nothing about the subject area involved, the grade level, the learning target, the nature of the assessments—or Gloria. They have only numbers, which is all their computerized grading program has. In every instance, more than 90 percent of teachers conclude that Gloria's summary grade should be a 4. Researchers would consider this an exceptionally high level of inter-rater reliability.

A More Complex Example

Some might argue that the case of Gloria is obvious and simplistic. They might justifiably question whether the same high level of reliability in teachers' professional judgment would be obtained in situations in which the patterns of students' performance were less consistent.

To explore this question, we next invite our groups of teachers to consider the scores of the five fictitious students shown in Figure 2. In each case, we ask teachers to use their professional judgment to determine each student's summary grade, keeping in mind the stated purpose: To best describe the student's level of proficiency regarding the learning target at this time. In every instance, teachers are remarkably consistent in determining students' summary grades—when they ignore the math and rely on their professional judgment.

Figure 2. Five students' scores from six assessments of a learning target.

Student	Learning Target #1						
	9/9	9/14	9/22	9/27	10/3	10/6	Summary Grade
Gloria	1	1	1	1	4	4	?
Ralph	2	1	2	3	3	3	?
Alice	2	2	4	4	4	3	?
David	3	1	3	2	3	1	?
Ellen	2	3	2	3	4	4	?

In the case of Ralph, for example, all teachers note his consistent performance on the three most recent assessments and assign a summary grade of 3. Alice poses an anomaly: On three assessments, she scored at the highest level, but she dropped to a 3 on the most recent assessment. After some discussion, most teachers conclude that something unusual may have affected Alice's performance on that last assessment. Perhaps an event outside of school - such as a distressing family issue – influenced her score. Being reluctant to give Alice a lower grade because of this single, anomalous score, most teachers give her a 4.

David presents the most inconsistent data. On the first assessment of the grading period, David received the highest score in the group; on the final assessment, he received the lowest. Even given this erratic pattern, however, teachers are remarkably consistent. Few say David deserves a 1 (his most recent score), and no one assigns a 4. Generally teachers are evenly divided between a summary grade of 2 or 3.

Ellen's scores fluctuated between 2s and 3s early in the grading period, but she received 4s on the two most recent assessments. Almost all teachers conclude that Ellen should receive a summary grade of 4.

After teachers complete this task, we show them the summary grades these students would have received if their teachers had relied on one of the statistical algorithms offered by their computerized grading program. We include an option available in many programs that allows teachers to base the summary grade on the most recent score. Figure 3 shows these grades, along with the grade chosen by the overwhelming majority of teachers involved in this experimental grading session. The summary grades determined by algorithms that differ from those chosen through teachers' professional judgment are in bold.

Figure 3. Algorithms yield summary grades different from grades derived by teachers' professional judgment.

Student	**Algorithm Used to Calculate Grade**					
	Mean (Average)	**Median**	**Mode**	**Trend**	**Most Recent Score**	**Teachers' Professional Judgment**
Gloria	**2**	**1**	**1**	**2.7**	4	4
Ralph	**2**	**2.5**	3	2.7	3	3
Alice	**3**	**3.5**	4	**3.5**	**3**	4
David	2	2.5	3	2.3	**1**	2 or 3
Ellen	**3**	**3**	**—**	**3.2**	4	4

Numbers in the rows represent the summary grade given to each student using the data shown in Figure 2. Numbers in bold indicate a summary grade determined by an algorithm that differs from the summary grade for that student determined by teachers' judgment. (There is no mode score for Ellen because her three scores all occur with the same frequency.)

In more than half the cases, the summary grade determined by a statistical algorithm differs from the summary grade teachers chose using their professional judgment. In Gloria's case, it can differ by as much as three grade categories. No algorithm would yield the same grade as teachers' professional judgment in every case.

If teachers chose these five algorithms with equal frequency (an unlikely scenario), the resulting reliability would be only about .6. Researchers would consider this a dubious level of reliability. When teachers use their judgment, however, the reliability is always .9 or greater. And we can assume that if teachers had knowledge of the students, their grade levels, the subject area, and the assessments as they considered these scores, their professional judgments would be even more consistent.

Trust Your Mind, Not Your Machine

As these examples reflect, teachers' thoughtful and informed professional judgments yield greater consistency in determining students' grades than do varied statistical algorithms. The takeaway message for teachers is, trust your mind instead of your machine (Jung, 2014). Teachers at every level *must* be able to defend the grades they assign and *must* have evidence to support their decisions. To serve as meaningful communication, grades must be fair, accurate, and reliable. They are more likely to be so when thoughtful professionals concur on the purpose of grades, look at the evidence they have, and then decide the grade that best summarizes that evidence.

Computers use only numbers. They know nothing of the individual students who produced those numbers, the learning environment, or the nature and quality of the assessments. Can having such knowledge sometimes result in teacher judgments being biased positively or negatively? Of course. But our experience indicates that this broader knowledge more often leads teachers to fair, accurate, and meaningful judgments.

For some students and some purposes, a grade based on a statistical algorithm may be fair and accurate. But rigidly applying the same algorithm to determine grades for all students in all classes distorts as often as it clarifies. Some computerized grading programs allow teachers to use different statistical algorithms in different classes. But no program allows teachers to vary the algorithm used from student to student within a class. Only by relying on professional judgment based on a clearly defined purpose can teachers appropriately individualize the grading process.

Grading is more a challenge of effective communication than a simple documentation of achievement (Guskey & Bailey, 2010). Teachers who trust their own minds—knowing that informed colleagues would likely make the same judgment—offer grades that communicate meaningful, reliable information to all.

This article is part of a Special Section on Examining Current Assessment Practices in the April 2016 issue of *Educational Leadership.*

- **Pre-Assessment: Promises and Cautions,** by Thomas R. Guskey and Jay McTighe
- **Standardized Tests: Purpose Is the Point,** by W. James Popham
- **Grading: Why You Should Trust Your Judgment,** by Thomas R. Guskey and Lee Ann Jung

References

Guskey, T. R. (2002). Computerized gradebooks and the myth of objectivity. *Phi Delta Kappan*, *83*(10), 775–780.

Guskey, T. R., & Bailey, J. M. (2010). *Developing standards-based report cards*. Thousand Oaks, CA: Corwin Press.

Jung, L. A. (2014, October). *Leading efforts to implement standards-based grading*. Presentation at the Conference on Educational Leadership, ASCD, Orlando, FL.

Marzano, R. J. (2000). *Transforming classroom grading*. Alexandria, VA: ASCD.

Endnote

[1] For a list of some of the many available computerized grading programs, see www.educationworld.com/a_tech/tech/tech031.shtml or www.capterra.com/gradebook-software.

Thomas R. Guskey is professor in the Department of Educational, School, and Counseling Psychology and **Lee Ann Jung** is professor in the Department of Special Education and director of international school partnerships in the College of Education at the University of Kentucky in Lexington.