Educational Leadership

December 2006/January 2007 | Volume **64** | Number **4**
Science in the Spotlight Pages 36-42

December 2006/
January 2007

Improving the Way We Grade Science

Standards-based grading systems can improve how we communicate learning expectations to students.

Jacqueline B. Clymer and Dylan Wiliam

Imagine, for a moment, a school that has an eight-week marking period, with students receiving a grade each week. Lesley starts out with four *A*s but ends up with four *C*s. Overall, of course, she gets a *B*. Chris, on the other hand, starts out with four *C*s but ends up with four *A*s. He gets a *B* too.

But who has learned more? In terms of overall achievement, Chris, with his four final *A*s, seems to have mastered the content for the marking period and really deserves an *A*. Conversely, Lesley, with her four final *C*s, seems far from mastering the content, but she gets a *B* because of her good start. The fact is that our current grading practices don't do the one thing they are meant to do, which is to provide an accurate indication of student achievement.

What the Research Shows

Within the last few years, research studies from around the world have shown that assessment can help students learn science, as well as measure how much science they have learned (Black & Wiliam, 1998; Wiliam, Lee, Harrison, & Black, 2004). Research has also suggested that the use of formative assessment, or assessment *for* learning, can double the rate of student learning (Wiliam & Thompson, in press). Perhaps more remarkable, such improvements have occurred even when achievement is measured using standardized tests.

To be effective, however, assessment *for* learning must be integrated into assessment *of* learning systems. In the United States, this means that in addition to taking into account fluctuations in student learning, teachers must still assign grades. Consequently, educators need to develop and implement a system that supports both the formative and summative functions of assessment—formative, in that teachers can use evidence of student achievement to adjust instruction to better meet student learning needs; and summative, in that teachers can amass the information to provide a final grade for a marking period.

Some years back, Terry Crooks (1988) reviewed more than 240 studies of the effect of assessment practices on students. He concluded that using assessment for grading purposes had completely overshadowed using assessment to support student learning. A more recent review of studies conducted between 1988 and 1997 found that nothing had improved (Black & Wiliam,

1998). Indeed, considerable evidence showed that many common grading practices actually lowered student performance.

The Meaning of Feedback

In reviewing more than 3,000 research reports on the effects of feedback in schools, colleges, and workplaces, Kluger and DeNisi (1996) found that only 131 of these studies were carried out with enough rigor and reported on in enough detail to be reliable. In 50 of these 131 studies, providing feedback actually made people's performance worse. In other words, in almost two of every five carefully controlled scientific studies, performance would have been better had the feedback *not* been given. When the researchers looked to see what kinds of feedback caused this decline in performance, they found that it was feedback that focused on the person, rather than on the task. When feedback focused on what the person needed to improve and on how he or she could go about making such improvements, learning improved considerably.

In fact, Ramaprasad (1983) reminds us that what teachers often refer to as "feedback" is not something an engineer would recognize as such. He pointed out that a feedback system (a room thermostat, for example) has four essential features. It provides

- A way to identify the current value of some system parameter.
- A way to set the desired value of some system parameter.
- A way to compare these two values, to see whether they are the same.
- A way to change the current value of the parameter to bring it closer to the desired value in the event the values differ.

In the case of the room thermostat, there is a device for measuring the current room temperature, a setting for specifying the desired temperature, a mechanism for comparing these two, and, of course, some wires that lead to the air conditioner or the furnace so that if there is a mismatch, we can do something about it. Few students experience comparable feedback systems that help "change the current value of the parameter" (their current level of achievement) and move them toward "the desired value" (mastery of content standards).

Given the evidence about the negative effects of grading practices in U.S. schools, it is hardly surprising that there have been regular calls for abandoning grading completely (see Kohn, 1999). We do not believe that the evidence supports such extreme action for two reasons. First, if teachers do not provide some indication of their students' achievement, then school systems are likely to resort to mechanisms like timed written examinations to do the job. Second, we believe that appropriately designed grading systems can help identify where students are in their understanding and what they need to do to improve.

Assessment that Supports Learning

Black and Wiliam (2004) contended that the starting point for any integrated assessment system must be the formative purpose. Teachers can always aggregate fine-scale data on student achievement to provide a grade or other summary of achievement, but they cannot work out what the student needs to do next on the basis of a grade or score.

The first requirement is a standards-based record-keeping system. For the record to serve as more than merely a justification for a final report card grade, the information that we collect on student performance must be instructionally meaningful. Knowing that a student got a *B* on an assignment is not instructionally meaningful. Knowing that the student understands what protons, electrons, and neutrons are but is confused about the distinction between atomic number and atomic mass *is* meaningful. This information tells the teacher where to begin instruction.

The second requirement of an assessment system that supports learning is that it should be dynamic rather than static. Grades based on the accumulation of points over time are counter-productive for several reasons. First, this approach encourages shallow learning. In most classrooms, if students forget something that they have previously been assessed on, they get to keep the grade. When students understand that it's what they know by the *end* of the marking period that counts, they are forced to engage with the material at a much deeper level. Second, not altering grades in light of new evidence of learning sends the message that the assessment is really a measure of aptitude rather than achievement. Students who think they will do well will engage in the assessments to prove how smart they are, whereas students who think that they are likely to fail will disengage. When assessment is dynamic, however, all students can improve. They come to see ability as incremental instead of fixed; they learn that smart is not something you are—it's something you become.

Assessing 8th Graders in Science

During the 2005–2006 school year, we conducted a pilot study on grading in an 8th grade physical science class in Quakertown Community School District. This suburban district located 30 miles from Philadelphia enrolls approximately 5,450 students. Strayer Middle School, where the pilot study took place, is a Title I school, with 22 percent of students receiving free or reduced-price lunch.

The school's academic year is divided into five marking periods that are each seven weeks long. Each marking period focuses on 10 content standards, which are derived from the state standards. For example, the first marking period in 8th grade physical science focuses on (1) the appropriate use of laboratory equipment; (2) metric unit conversion and labeling; (3) calculating density; (4) applying density (floating, sinking, layering, thermal expansion); (5) density as a characteristic property; (6) the phases of matter (at a molecular level); (7) gas laws; (8) communication (graphing); (9) communication (lab reports); and (10) inquiry skills.

For each content standard, the teacher identifies an evidence base that will verify the student's level of proficiency. For the standard relating to laboratory equipment, for example, the teacher may ask students to do the “clay boat” task, in which they explore buoyancy by comparing the mass, volume, and density of a ball of clay with the mass, volume, and density of a boat made of clay. Students might be required to complete a lab about the density of soda, in which they learn why some cans of soda float whereas others sink, and to pass a lab quiz in which they must measure, without their lab partners, the mass, volume, and density of solids and liquids. For the standard relating to metric unit conversion and labeling, the teacher might ask students to take a quiz on metric conversion, observe the use of units in the lab, and show mastery on relevant test items.

On the basis of this evidence, the teacher assesses each student's performance on each content standard using the following "traffic lights":

- *Green (Mastery): The student consistently meets and often exceeds the content standard*. The student, with relative ease, grasps, applies, and extends key concepts, processes, and skills for the grade level.
- *Yellow (Developing): The student regularly meets the content standard*. The student, with limited errors, grasps and applies key concepts, processes, and skills for the grade level.
- *Red (Beginning or below basic): The student is beginning to, and occasionally does, meet the content standards, or the student is not meeting them*. The student is beginning to grasp and apply key concepts, processes, and skills for the grade level but produces work that contains many errors.

The final grade for the marking period is based on the aggregate level of proficiency displayed in the 10 content standards. "Green lights" are worth 2 points, "yellow lights" are worth 1 point, and "red lights" are worth 0 points. Consequently, the highest score for the marking period is 20 points (10 content standards × 2 points), or 100 percent.

To receive an *A*, students need to master at least 90 percent of the required content, earning a minimum of 18 points. A student can achieve this with 10 greens (20 points), 9 greens and 1 yellow (19 points), 9 greens and 1 red (18 points), or 8 greens and 2 yellows (18 points). A grade of *B* reflects 80 percent mastery (a minimum of 16 points), and a *C* reflects 70 percent mastery (a minimum of 14 points). Students can achieve these points through various configurations of "lights."

At the end of the unit, students take a test to verify their level of mastery in each identified content/skill area. If students do better than expected, the teacher updates their achievement profile with this "latest and best" evidence. If students fail to show mastery of previously mastered content, the teacher interviews them to provide them with additional opportunities to "show what they know" about the topic. For example, if the student misses a question about Charles's Law, which describes the direct relationship between temperature and volume of a gas, the teacher might ask the student to explain the gas law and provide a real-life example. If the interview reveals that the student has not mastered this material, the teacher provides the student with additional practice and multiple opportunities to learn it.

If a student does not master the content by the end of the marking period, the grade for that marking period reflects this lack of mastery. But if the student masters the content by the end of the school year, then that increase is reflected in the end-of-year grade. The end-of-year grade is the percentage of mastery on all 50 standards and *not* an average of the five marking period grades.

A crucial feature of this assessment system is that no grade is final until the end of the marking period. While students are learning, the teacher maintains a record of the current evidence of achievement. One particularly effective way to keep this record is by using an electronic spreadsheet (see fig. 1). Each row represents a student's performance on the content standards for that unit, and each column represents one of the 10 content standards. The teacher enters 0,

1, or 2 in each cell to indicate the student's existing level of mastery, calculating the final score and grade as previously described.

Figure 1. Screenshot of a Portion of a Teacher Grade Book

	A	B	C	D	E	F	G	H	I	J	K	L	M
		Use of Lab Equipment	Metric Unit Conversion	Density Calculation	Density Application	Density as a Characteristic Property	Phases of Matter	Gas Laws	Communication (Graphing)	Communication (Lab Reports)	Inquiry Skills	Score	Grade
	Period	1	1	1	1	1	1	1	1	1	1		
	Standard	1	2	3	4	5	6	7	8	9	10		
1	Georgie	2	2	2	2	2	2	1	2	2	2	95	A
2	Kirsty	1	1	1	1	2	2	1	2	1	2	70	C
3	Victoria	2	1	2	2	2	1	2	2	1	2	85	B
4	Jennifer	2	1	2	1	2	1	1	1	2	2	75	C
5	Andrew	2	1	1	1	2	1	1	2	1	2	70	C
6	Jonathan	0	2	2	2	1	1	2	1	1	2	70	C
7	Charlotte	1	1	2	1	1	1	2	1	2	2	70	C
8	Scott	0	0	2	2	1	1	1	1	2	2	60	D
9	Amy	2	1	2	2	2	1	2	1	1	2	80	B
10	Grace	1	0	2	2	1	1	1	1	2	2	65	D
11	Lee	2	2	2	1	2	1	2	2	2	2	90	A
12	Peter	2	1	2	2	0	2	1	2	1	2	75	C
13	Thomas	1	1	2	1	0	1	2	2	2	2	70	C
	Average %	69	54	92	77	69	62	73	77	77	100		

Read horizontally to determine a student's mastery of each of 10 standards. Green (2) shows mastery, yellow (1) shows developing skills, and red (0) shows beginning or below basic skills. Read vertically to determine a group of students' mastery of a given standard. The percentages at the bottom indicate the group's mastery of each of 10 standards.

One particularly useful feature in Excel that facilitates data analysis is the option to color cells differently. The teacher can give cells different traffic light colors, providing an immediate display of student achievement. The rows indicate the skills and topics that individual students have and have not mastered. The columns show how well the entire class has or has not mastered a specific content area. This information is formative because it helps the teacher adjust instruction to better meet student needs. The data shown in Figure 1 suggest that this group of students has a good grasp of standards 3 and 10, whereas standards 1, 2, 5, and 6 merit further attention.

In the pilot study, the teacher kept the students apprised of their traffic light ratings through weekly progress reports. Individual assignments were not allocated a single color or grade. Instead, the teacher reviewed student work for evidence of mastery on one or more of the standards and recorded this information in a grade book that designated one page for each standard. In a lab report on the density of pennies, for example, the student would receive a traffic light score for the graphing standard, for the inquiry-skill standard, for the calculation of density standard, and so on.

In addition, the teacher commented on individual students' work to coach the students to higher achievement. Students were encouraged to act on the feedback by providing additional evidence of mastery or by revising their work to improve both their understanding and ratings.

Student Reaction

In May 2006, we asked a sample of 19 students to explain the new grading approach. All but one student understood that their achievement at the end of the marking period was more important than their achievement when a topic was first introduced. They understood that they were expected to improve as a result of instruction and not expected to arrive at school already knowing the content.

Many students shifted from a performance orientation to their work, in which the goal is to get the highest grade, to a mastery orientation, in which the goal is understanding (Dweck, 2000). Students said that they understood more, focused more on learning important concepts, and were more relaxed because the teacher judged their performance on the basis of their understanding. One immediate, if unanticipated, outcome was the change in classroom atmosphere. Students became more engaged in monitoring their own learning. They repeatedly asked for clarification, from their peers and from the teacher, to ensure their understanding.

A majority of students preferred this system to the previous one. Several students indicated that the provisional, rather than final, nature of the grading system was an important benefit, noting that it enabled them not only to revise their grades as they improved their understanding but also to see what they were good at and where they needed to improve. Other students noticed that the new assessment system focused more on learning than on performance. Earning points and percentages became less important than understanding the content.

Responding to a question about teacher feedback, several students mentioned how helpful it was to receive feedback that not only indicated what was incorrect but also provided some idea of how to improve or correct it. This greatly helped students revise their work. One student appeared to have intuitively adopted Ramaprasad's notion of feedback, because for him the most helpful part was “just having feedback at all.” He noted that comments like “Good job!” which he had received from other teachers in the past, were not really feedback because they didn't provide information about how to improve.

The pilot study also revealed some other notable reactions. Under the new assessment system, two-thirds of the students saw the teacher as a coach, one-fifth saw the teacher as both a coach and a judge, and only one in 10 students saw the teacher solely as a judge. Three-fourths of the students noted that they prepared for tests, and half of these students indicated that this was a change from the previous year. Half the students thought they were doing better in science that year, and half thought they were doing the same. No students believed that they were doing worse in science than the previous year, despite the possibility that their scores could be revised downward as a result of new evidence.

The Effect on Student Achievement

Because this was a pilot study with just a single classroom, any quantitative information about

the achievement of this class relative to other classes must be regarded as merely indicative. However, the average score of these students on the final examination in May 2006 was 79 points, with a standard deviation of approximately 14 points. The average score achieved by the equivalent class in 2004–2005 was 73, suggesting that the 2005–2006 class outperformed the 2004–2005 class by 0.4 standard deviations.

A comparison of the two classes' scores on ACT's Explore test established the similarities between the two classes. Although the 2005–2006 class averaged 2 percentage points higher than the 2004–2005 class on the science component of the Explore test, the 2005–2006 class outscored the 2004–2005 class by 3 points on the final exam (after controlling for prior achievement in science), equivalent to a standardized effect size of 0.27 standard deviations. Because of the small size of the sample, this result was not statistically significant ($p = 0.1785$), but it is comparable to the effect sizes found in larger studies that have examined the effect of assessment for learning (Black & Wiliam, 1998; Hayes, 2003; Wiliam, Lee, Harrison, & Black, 2004). Analysis of the students' scores suggests that although the new grading system was, on average, better for all students, it was particularly beneficial for the highest- and lowest-achieving students.

A Caveat for Special Ed

The increased student engagement we saw in the classroom, together with evidence of increased achievement, suggests that this kind of assessment system has considerable potential for enabling teachers to integrate the summative and formative functions of assessment. Several teachers at Quakertown have implemented similar approaches in middle school math, high school social studies, and AP Statistics with considerable success.

Taking this work forward, we need to pay careful attention to assessing students with special education needs. In a genuine standards-based assessment system, teachers need to assess and record what a student can actually do. Many students with special needs may fail to progress beyond the “basic” level; applying a standards-based assessment system would lead to these students getting mostly *D*s and *F*s. As a result, to mark these students' modest progress, some teachers may be inclined to give them credit for mastering a content standard that they have not truly mastered.

We think this is a mistake. The record-keeping system must provide accurate information about what the student can and can't do. It might make sense to “disapply” some standards as inappropriate for some students; consequently some cells in their rows would be empty. But when a cell includes a number, it should mean the same thing for all students. If teachers want to incorporate other information, such as effort or improvement, they can best do this by reporting effort and improvement scores separately so that the information about the student's achievement is clear and interpretable.

Some school districts require teachers to aggregate all information into a single grade that combines information about achievement with a range of other, less objective factors. To maintain the integrity of the standards-based record, however, the teacher needs to separate the assessment of achievement from other factors. For example, a teacher might specify that 50 percent of the points are for achievement, 25 percent are for effort, and 25 percent are for

improvement. A student who masters all the standards specified for a marking period gets at least 50 percent, but he or she does not get even a *C* without demonstrating both effort and improvement from the last marking period. This approach is far from perfect, but at least the achievement record accurately records the student's achievement.

Standards-based assessment systems are a significant improvement over the grading practices prevalent in U.S. schools today. They communicate standards for success, helping students see what they need to improve. They reposition the teacher as coach rather than judge, leading to less confrontational classroom environments. Most important, they support the teacher in using assessment to improve learning rather than just to measure it.

References

Black, P. J., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy, and Practice, 5*(1), 7–73.

Black, P. J., & Wiliam, D. (2004). The formative purpose: Assessment must first promote learning. In M. Wilson (Ed.), *Towards coherence between classroom assessment and accountability: 103rd yearbook of the National Society for the Study of Education (part 2)* (pp. 20–50). Chicago: University of Chicago Press.

Crooks, T. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research, 58*, 438–481.

Dweck, C. S. (2000). *Self-theories: Their role in motivation, personality, and development*. Philadelphia: Psychology Press.

Hayes, V. P. (2003). *Using pupil self-evaluation within the formative assessment paradigm as a pedagogical tool*. Unpublished thesis: King's College, University of London.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*(2), 254–284.

Kohn, A. (1999). From degrading to degrading. *High School Magazine, 6*(5), 38–43.

Ramaprasad, A. (1983). On the definition of feedback. *Behavioral Science, 28*(1), 4–13.

Wiliam, D., Lee, C., Harrison, C., & Black, P. J. (2004). Teachers developing assessment for learning: Impact on student achievement. *Assessment in Education: Principles, Policy, and Practice, 11*(1), 49–65.

Wiliam, D., & Thompson, M. (in press). Integrating assessment with instruction: What will it take to make it work? In C. A. Dwyer (Ed.), *The future of assessment: Shaping teaching and learning*. Mahwah, NJ: Erlbaum.

Jacqueline B. Clymer is an 8th grade science teacher and lead teacher for K–12 science in Quakertown Community

School District, Quakertown, Pennsylvania; jclymer@qcsd.org. **Dylan Wiliam** is Deputy Director of the Institute of Education, University of London, United Kingdom; dylanwiliam@mac.com.
