CBE—Life Sciences Education
Vol. 13, 159–166, Summer 2014

Feature
Approaches to Biology Teaching and Learning

Teaching More by Grading Less (or Differently)

Jeffrey Schinske* and Kimberly Tanner†

*Department of Biology, De Anza College, Cupertino, CA 95014; †Department of Biology, San Francisco State University, San Francisco, CA 94132

INTRODUCTION

> When we consider the practically universal use in all educational institutions of a system of marks, whether numbers or letters, to indicate scholastic attainment of the pupils or students in these institutions, and when we remember how very great stress is laid by teachers and pupils alike upon these marks as real measures or indicators of attainment, we can but be astonished at the blind faith that has been felt in the reliability of the marking systems.
>
> —I. E. Finkelstein (1913)

If your current professional position involves teaching in a formal classroom setting, you are likely familiar with the process of assigning final course grades. Last time you assigned grades, did you assign an "E," "E+," or "E−" to any of your students? Likely you assigned variations on "A's," "B's," "C's," "D's," and "F's." Have you wondered what happened to the "E's" or talked with colleagues about their mysterious absence from the grading lexicon? While we often commiserate about the process of assigning grades, which may be as stressful for instructors as for students, the lack of conversation among instructors about the mysterious omission of the "E" is but one indicator of the many tacit assumptions we all make about the processes of grading in higher education. Given that the time and stress associated with grading has the potential to distract instructors from other, more meaningful aspects of teaching and learning, it is perhaps time to begin scrutinizing our tacit assumptions surrounding grading. Below, we explore a brief history of grading in higher education in the United States. This is followed by considerations of the potential purposes of grading and insights from research literature that has explored the influence of grading on teaching and learning. In particular, does grading provide feedback for students that can promote learning? How might grades motivate struggling students? What are the origins of norm-referenced grading—also known as curving? And, finally, to what extent does grading provide reliable information about student learning and mastery of concepts? We end by offering four potential adjustments to our general approach to grading in undergraduate science courses for instructors to consider.

DOI: 10.1187/cbe.CBE-14-03-0054
Address correspondence to: Jeffrey Schinske (schinskejeff@deanza.edu).

A BRIEF HISTORY OF GRADING IN HIGHER EDUCATION

> It can be easy to perceive grades as both fixed and inevitable—without origin or evolution . . . Yet grades have not always been a part of education in the United States.
>
> —Schneider and Hutt (2013)

Surprisingly, the letter grades most of us take for granted did not gain widespread popularity until the 1940s. Even as late as 1971, only 67% of primary and secondary schools in the United States used letter grades (National Education Association, 1971). It is therefore helpful to contextualize the subject to appreciate the relatively young and constantly changing nature of current systems of grading. While not an exhaustive history, the sections below describe some of the main developments leading to the current dominant grading system.

Early 19th Century and Before

The earliest forms of grading consisted of exit exams before awarding of a degree, as seen at Harvard as early as 1646 (Smallwood, 1935). Some schools also awarded medals based on competitions among students or held regular competitions to assign seats in class (Cureton, 1971). Given that universities like Yale and Harvard conducted examinations and elected valedictorians and salutatorians early in the 18th century, some scale of grading must have existed. However, the first official record of a grading system surfaces in 1785 at Yale, where seniors were graded into four categories: *Optimi*, second *Optimi*, *Inferiores*, and *Perjores* (Stiles, 1901, cited by Smallwood, 1935). By 1837, Yale was also recording

student credit for individual classes, not just at the completion of college studies, using a four-point scale. However, these "merit marks" were written in code and hidden from students (Bagg, 1871).

Harvard and other schools soon experimented with public rankings and evaluations, noting that this resulted in "increasing [student] attention to the course of studies" and encouraged "good moral conduct" (Harvard University, 1832). Concerned that such public notices would inspire competition among students, which would distract from learning, other schools used more frequent, lower-stakes "report cards" to provide feedback on achievement (Schneider and Hutt, 2013). In 1837, at least some professors at Harvard were grading using a 100-point system (Smallwood, 1935). During this same period, William and Mary placed students in categories based on attendance and conduct. The University of Michigan experimented with a variety of grading systems in the 1850s and 1860s, including various numeric and pass/fail systems (Smallwood, 1935). Still, many schools at this time kept no formal records of grades (Schneider and Hutt, 2013).

Late 19th Century and 20th Century

With schools growing rapidly in size and number and coordination between schools becoming more important, grades became one of the primary means of communication between institutions (Schneider and Hutt, 2013). This meant grades needed to have meaning not just within an institution but also to distant third parties. A record from 1883 indicates a student at Harvard received a "B," and in 1884, Mount Holyoke was grading on a system including "A," "B," "C," "D," and "E." Each letter corresponded to a range of percentage scores, with lower than 75% equating to an "E" and indicating failure. Mount Holyoke added an "F" grade (for failing) to the scale in 1898 and adjusted the percentages relating to the other letters (Smallwood, 1935). This appears to be the initial origin of the "A"–"F" system familiar to most faculty members today, albeit including an "E" grade. By 1890, the "A"–"E" system had spread to Harvard after faculty members expressed concerns regarding reliably grading students on a 100-point scale. Still, grading was not always done at schools and grading systems varied widely (Schneider and Hutt, 2013).

By the early 1900s, 100-point or percentage-based grading systems were very common (Cureton, 1971). This period also saw an increased desire for uniformity in grading, and many expressed concerns about what grades meant from one teacher or institution to the next (Weld, 1917). Numerous studies of the period sought to understand and perfect grading systems (Cureton, 1971). Grading on a 100-point scale was found to be highly unreliable, with different teachers unable to assign consistent grades on papers in English, math, and history (Starch, 1913). Researchers felt that getting away from a 100-point scale and grading into only five categories (e.g., letter grades) could increase reliability (Finkelstein, 1913, p. 18). While it is unclear exactly when and why "E" grades disappeared from the letter grade scale, it seems possible that this push to use fewer categories resulted in an "A"–"F" scale with no "E" ("F" being retained, since it so clearly stood for "fail"). Others have conjectured that "E" was removed so students would not assume "E" stood for "excellent," but whatever the reason, "E's" apparently disappeared by the 1930s (Palmer, 2010).

As research on intellectual ability appeared to show that, like other continuous biological traits, levels of aptitude in a population conformed to a normal curve, some experts felt grades should similarly be distributed according to a curve in a classroom (Finkelstein, 1913). Distributing grades according to a normal curve was therefore considered as a solution to the subjective nature of grading and a way to minimize interrater differences in grading (Guskey, 1994). Others worried that measuring aptitude was different from measuring levels of classroom performance, which might not be normally distributed (Schneider and Hutt, 2013).

Based on the above research and the pressure toward uniformity of grading systems, by the 1940s the "A"–"F" grading system was dominant, with the four-point scale and percentages still also in use (Schneider and Hutt, 2013). However, many inconsistencies remained. As one example, Yale used no less than four different grading systems from the 1960s to 1980s (Yale University, 2013).

Present Day

Grading systems remain controversial and hotly debated today (Jaschik, 2009). Some argue grades are psychologically harmful (Kohn, 1999). Others raise concerns about the integrity of the "A"–"F" system, given well-documented trends in grade inflation (Rojstaczer and Healy, 2012). One professor summed it up by saying grades do no more than "create a facade of coherence" (Jaschik, 2009). A number of colleges have abandoned numerical and categorical grading altogether, opting instead for creating contracts with students to define success or employing student self-reflection in combination with written evaluations by faculty (Jaschik, 2009). Among the Ivy League schools, Brown University does not calculate grade point averages, does not use "D's" in its grading scale, and does not record failing grades (Brown University, 2014). Even Yale, the institution that started this history of grading more than 200 yr ago, is today still considering changes to its grading system (Yale University, 2013).

Though grades were initially meant to serve various pedagogical purposes, more recent reforms have focused on "grades as useful tools in an organizational rather than pedagogical enterprise—tools that would facilitate movement, communication, and coordination" (Schneider and Hutt, 2013). So, what are the potential purposes of grading in educational settings?

PURPOSES OF GRADING—PAST AND PRESENT

Grades as Feedback on Performance—Does Grading Provide Feedback to Help Students Understand and Improve upon Their Deficiencies?

> [This] work affirms an observation that many classroom teachers have made about their students: if a paper is returned with both a grade and a comment, many students will pay attention to the grade and ignore the comment.
>
> —Brookhart (2008, p. 8)

For most faculty members, the concept of feedback has at least two applications to the concept of grading. On one hand, grading itself is a form of feedback that may be

useful to students. In addition, in the process of grading student work, faculty members sometimes provide written comments as feedback that students could use to improve their work. Because college students express a desire for feedback (Higgins *et al*., 2002), faculty members may feel pressured to grade more (rather than facilitating ungraded activities) and to provide more written feedback while grading. Especially in large classes, this can significantly increase workload on faculty (Nicol and Macfarlane-Dick, 2006; Crisp, 2007). But are grades and written comments effective forms of feedback that assist students in achieving conceptual mastery of the subject?

Feedback is generally divided into two categories: evaluative feedback and descriptive feedback. Evaluative feedback, such as a letter grade or written praise or criticism, judges student work, while descriptive feedback provides information about how a student can become more competent (Brookhart, 2008, p. 26). Butler and Nisan (1986) compared the impacts of evaluative feedback, descriptive feedback, and no feedback on student achievement in problem-solving tasks and in "quantitative" tasks (e.g., those requiring quick, timed work to produce a large number of answers). They found that students receiving descriptive feedback (but *not* grades) on an initial assignment performed significantly better on follow-up quantitative tasks and problem-solving tasks than did students receiving grades or students receiving no feedback. Students receiving grades performed better on follow-up quantitative tasks than students receiving no feedback, but did not outperform those students on problem-solving assignments. In other words, providing evaluative feedback (in this case, grades) after a task does not appear to enhance students' future performance in problem solving.

While descriptive, written feedback can enhance student performance on problem-solving tasks; reaping those benefits requires students to read, understand, and use the feedback. Anecdotal accounts, as well as some studies, indicate that many students do not read written feedback, much less use it to improve future work (MacDonald, 1991; Crisp, 2007). In one study, less than half of undergraduate medical students even chose to collect the feedback provided on their essays (Sinclair and Cleland, 2007). Other studies suggest that many students do read feedback and consider it carefully but the feedback is written in a way that students do not find useful in improving future work (Higgins *et al*., 2002). Some studies have further investigated the relationships between grading and descriptive feedback by providing students with both written feedback and grades on assignments. In these cases, the addition of written comments consistently failed to enhance student performance on follow-up tasks (Marble *et al*., 1978; Butler 1988; Pulfrey *et al*., 2011). Brookhart (2008, p. 8) concludes, "the grade 'trumps' the comment" and "comments have the best chance of being *read* as descriptive if they are not accompanied by a grade." Even when written feedback is read, there is widespread agreement that instructor feedback is very difficult for students to interpret and convert into improved future performance (Weaver, 2006).

Grading does not appear to provide effective feedback that constructively informs students' future efforts. This is particularly true for tasks involving problem solving or creativity. Even when grading comes in the form of written comments, it is unclear whether students even read such comments, much less understand and act on them.

Grades as a Motivator of Student Effort—Does Grading Motivate Students to Learn?

> Our results suggest. . .that the information routinely given in schools—that is, grades—may encourage an emphasis on quantitative aspects of learning, depress creativity, foster fear of failure, and undermine interest.
>
> —Butler and Nisan (1986)

As described in the history of grading above, our current "A"–"F" grading system was not designed with the primary intent of motivating students. Rather, it stemmed from efforts to streamline communication between institutions and diminish the impacts of unreliable evaluation of students from teacher to teacher (Grant and Green, 2013). That is not to say, however, that grades do not have an impact on student motivation and effort. At some point, every instructor has likely experienced desperate petitions from students seeking more points—a behavior that seems to speak to an underlying motivation stimulated by the grading process.

It would not be surprising to most faculty members that, rather than stimulating an interest in learning, grades primarily enhance students' motivation to avoid receiving bad grades (Butler and Nisan, 1986; Butler, 1988; Crooks, 1988; Pulfrey *et al*., 2011). Grades appear to play on students' fears of punishment or shame, or their desires to outcompete peers, as opposed to stimulating interest and enjoyment in learning tasks (Pulfrey *et al*., 2011). Grades can dampen existing intrinsic motivation, give rise to extrinsic motivation, enhance fear of failure, reduce interest, decrease enjoyment in class work, increase anxiety, hamper performance on follow-up tasks, stimulate avoidance of challenging tasks, and heighten competitiveness (Harter, 1978; Butler and Nisan, 1986; Butler, 1988; Crooks, 1988; Pulfrey *et al*., 2011). Even providing encouraging, written notes on graded work does not appear to reduce the negative impacts grading exerts on motivation (Butler, 1988). Rather than seeing low grades as an opportunity to improve themselves, students receiving low scores generally withdraw from class work (Butler, 1988; Guskey, 1994). While students often express a desire to be graded, surveys indicate they would prefer descriptive comments to grades as a form of feedback (Butler and Nisan, 1986).

High-achieving students on initial graded assignments appear somewhat sheltered from some of the negative impacts of grades, as they tend to maintain their interest in completing future assignments (presumably in anticipation of receiving additional good grades; Butler, 1988). Oettinger (2002) and Grant and Green (2013) looked specifically for positive impacts of grades as incentives for students on the threshold between grade categories in a class. They hypothesized that, for example, a student on the borderline between a "C" and a "D" in a class would be more motivated to study for a final exam than a student solidly in the middle of the "C" range. However, these studies found only minimal (Oettinger, 2002) or no (Grant and Green, 2013) evidence that grades motivated students to perform better on final exams under these conditions.

This is not to say that classroom evaluation is by definition harmful or a thing to avoid. Evaluation of students in the service of learning—generally including a mechanism for feedback without grade assignment—can serve to enhance learning and motivation (Butler and Nisan, 1986; Crooks,

1988; Kitchen *et al*., 2006). Swinton (2010) additionally found that a grading system that explicitly rewarded effort in addition to rewarding knowledge stimulated student interest in improvement. This implies that balancing accuracy-based grading with providing meaningful feedback and awarding student effort could help avoid some of the negative consequences of grading.

Rather than motivating students to learn, grading appears to, in many ways, have quite the opposite effect. Perhaps at best, grading motivates high-achieving students to continue getting high grades—regardless of whether that goal also happens to overlap with learning. At worst, grading lowers interest in learning and enhances anxiety and extrinsic motivation, especially among those students who are struggling.

Grades as a Tool for Comparing Students—Is Grading on a Curve the Fairest Way to Grade?

> You definitely compete for grades in engineering; whereas you earn grades in other disciplines . . . I have to get one point higher on the test than the next guy so I can get the higher grade.
>
> —Student quoted in Seymour and Hewitt (1997, p. 118)

The concept of grading on a curve arose from studies in the early 20th century suggesting that levels of aptitude, for example as measured by IQ, were distributed in the population according to a normal curve. Some then argued, if a classroom included a representative sample from the population, grades in the class should similarly be distributed according to a normal curve (Finkelstein, 1913). Conforming grades to a curve held the promise of addressing some of the problems surrounding grading by making the process more scientific and consistent across classrooms (Meyer, 1908). Immediately, even some proponents of curved grading recognized problems with comparing levels of aptitude in the population with levels of classroom achievement among a population of students. For a variety of reasons, a given classroom might not include a representative sample from the general population. In addition, teachers often grade based on a student's performance or accomplishment in the classroom—characteristics that differ in many ways from aptitude (Finkelstein, 1913). However, despite the reservations of some teachers and researchers, curved grading steadily gained acceptance throughout much of the 20th century (Schneider and Hutt, 2013).

Grading on a curve is by definition a type of "norm-referenced" grading, meaning student work is graded based on comparisons with other students' work (Brookhart, 2004, p. 72). One issue surrounding norm-referenced grading is that it can dissociate grades from any meaning in terms of content knowledge and learning. Bloom (1968) pointed out that, in grading on a curve "it matters not that the failures of one year performed at about the same level as the C students of another year. Nor does it matter that the A students of one school do about as well as the F students of another school." As this example demonstrates, under curved grading, grades might not communicate any information whatsoever regarding a student's mastery of course knowledge or skills.

Of even more concern, however, is the impact norm-referenced grading has on competition between students. The quote at the start of this section describes how many students respond to curve-graded classes compared with classes that do not use a grading curve. Seymour and Hewitt (1997, p. 118) explain, "Curve-grading forces students to compete with each other, whether they want to or not, because it exaggerates very fine degrees of differences in performance. Where there is little or no difference in work standards, it encourages a struggle to create it." Studies have shown that science students in competitive class environments do not learn or retain information as well as students in cooperative class environments (Humphreys *et al*., 1982). Students in cooperative environments are additionally more interested in learning and find learning more worthwhile than students in competitive environments (Humphreys *et al*., 1982). Of particular concern is that the competitive environment fostered by norm-referenced grading represents one of the factors contributing to the loss of qualified, talented, and often underrepresented college students from science fields (Seymour and Hewitt, 1997; Tobias, 1990). Disturbingly, even when a science instructor does not grade on a curve, students might, due to their past experiences, assume a curve is used and adopt a competitive stance anyway (Tobias, 1990, p. 23).

Bloom (1968, 1976) presents evidence and a theoretical framework supporting an alternate view of grading whereby most students would be expected to excel and not fall into the middle grades. He states, "If the students are normally distributed with respect to aptitude, but the kind and quality of instruction and the amount of time available for learning are made appropriate to the characteristics and needs of each student, the majority of students may be expected to achieve mastery of the subject. And, the relationship between aptitude and achievement should approach zero" (Bloom, 1968). In other words, even if we were to accept a concept of innate aptitude that is normally distributed in a classroom, that distribution should not predict classroom achievement, provided the class environment supports diverse learners in appropriate ways. This idea was a significant development, because it freed teachers from the stigma associated with awarding a larger number of high grades. Previously, an excess of higher grades was thought to arise only from either cheating by students or poor grading practices by teachers (Meyer, 1908). Bloom's model argues that, when given the proper learning environment and compared against standards of mastery in a field (rather than against one another), large numbers of students could succeed. This type of grading—where instructional goals form the basis of comparison—is called "criterion-referenced" grading (Brookhart, 2004, p. 72).

Of course, Bloom's work did not rule out the possibility that some teachers might still give high grades for undesirable reasons unrelated to standards of mastery (e.g., to be nice, to gain the admiration of students, etc.). Such practices would not be in line with Bloom's work and would lead to pernicious grade inflation. Indeed, many of those bemoaning recent trends in grade inflation in higher education (though less prevalent in the sciences) point to the abandonment of curved grading as a major factor (Rojstaczer and Healy, 2012). Such studies often promote various forms of curving—at the level of individual courses or even at the institution as a whole—to combat inflation (Johnson, 2003, chaps. 7–8). In light of the above, however, it seems strange to aspire to introduce grading systems that could further push students into competition and give rise to grades that indicate little about the mastery of knowledge or skills in a subject. The broader

distribution of grades under curve-adjusted grading could simply create the illusion of legitimacy in the grading system without any direct connection between grades and achievement of learning goals. Perhaps the more productive route is to push for stronger, criterion-referenced grading systems in which instructional goals, assessments, and course work are more intimately aligned.

In brief, curved grading creates a competitive classroom environment, alienates certain groups of talented students, and often results in grades unrelated to content mastery. Curving is therefore not the fairest way to assign grades.

Grades as an Objective Evaluation of Student Knowledge—Do Grades Provide Reliable Information about Student Learning?

> Study Critiques Schools over Subjective Grading: An Education Expert Calls for Greater Consistency in Evaluating Students' Work.
>
> —Los Angeles Times (2009)

As evidenced by the above headline, some have criticized grading as subjective and inconsistent, meaning that the same student could receive drastically different grades for the same work, depending on who is grading the work and when it is graded. The literature indeed indicates that some forms of assessment lend themselves to greater levels of grading subjectivity than others.

Scoring multiple-choice assessments does not generally require the use of professional judgment from one paper to the next, so instructors should be able to score such assessments objectively (Wainer and Thissen, 1993; Anderson, 2008, p. 451). However, despite their advantages in terms of objective grading, studies have raised concerns regarding the blanket use of multiple-choice assessments. Problems with such assessments range from their potential to falsely indicate student understanding to the possibilities that they hamper critical thinking and exhibit bias against certain groups of students (Towns and Robinson, 1993; Scouller, 1998; Rogers and Harley, 1999; Paxton, 2000; Dufresne *et al*., 2002; Zimmerman and Williams, 2003; Stanger-Hall, 2012).

Grading student writing, whether in essays, reports, or constructed-response test items, opens up greater opportunities for subjectivity. Shortly after the rise in popularity of percentage-based grading systems in the early 1900s, researchers began examining teacher consistency in marking written work by students. Starch and Elliott (1912) asked 142 teachers to grade the same English paper and found that grades on the paper varied from 50 to 98% between teachers. Because different teachers awarded scores ranging from failing to exceptional, the researchers concluded "the promotion or retardation of a pupil depends to a considerable extent upon the subjective estimate of his teacher" rather than upon the actual work produced by the student (Starch and Elliott, 1912). Even greater levels of inconsistency were found in teachers' scoring of a geometry paper showing the solution to a problem (Starch and Elliott, 1913).

Eells (1930) investigated the consistency of individual teachers' grading by asking 61 teachers to grade the same history and geography papers twice—the second time 11 wk after the first. He concluded that "variability of grading is about as great in the same individual as in groups of different individuals" and that, after analysis of reliability coefficients, assignment of scores amounted to "little better than sheer guesses" (Eells, 1930). Similar problems in marking reliability have been observed in higher education environments, although the degree of reliability varies dramatically, likely due to differences in instructor training, assessment type, grading system, and specific topic assessed (Meadows and Billington, 2005, pp. 18–20). Factors that occasionally influence an instructor's scoring of written work include the penmanship of the author (Bull and Stevens, 1979), sex of the author (Spear, 1984), ethnicity of the author (Fajardo, 1985), level of experience of the instructor (Weigle, 1999), order in which the papers are reviewed (Farrell and Gilbert, 1960; Spear, 1996), and even the attractiveness of the author (Bull and Stevens, 1979).

Designing and using rubrics to grade assignments or tests can reduce inconsistencies and make grading written work more objective. Sharing the rubrics with students can have the added benefit of enhancing learning by allowing for feedback and self-assessment (Jonsson and Svingby, 2007; Reddy and Andrade, 2010). Consistency in grading tests can also be improved by writing longer tests with more narrowly focused questions, but this would tend to limit the types of questions that could appear on an exam (Meadows and Billington, 2005).

In summary, grades often fail to provide reliable information about student learning. Grades awarded can be inconsistent both for a single instructor and among different instructors for reasons that have little to do with a students' content knowledge or learning advances. Even multiple-choice tests, which can be graded with great consistency, have the potential to provide misleading information on student knowledge.

GRADING—STRATEGIES FOR CHANGE

In part, grading practices in higher education have been driven by educational goals such as providing feedback to students, motivating students, comparing students, and measuring learning. However, much of the research literature on grading reviewed above suggests that these goals are often not being achieved with our current grading practices. Additionally, the expectations, time, and stress associated with grading may be distracting instructors from integrating other pedagogical practices that could create a more positive and effective classroom environment for learning. Below we explore several changes in approaching grading that could assist instructors in minimizing its negative influences. Kitchen *et al*. (2006) additionally provide an example of a high-enrollment college biology class that was redesigned to "maximize feedback and minimize the impact of grades."

Balancing Accuracy-Based Grading with Effort-Based Grading

Multiple research studies described above suggest that the evaluative aspect of grading may distract students from a focus on learning. While evaluation will no doubt always be key in determining course grades, the entirety of students' grades need not be based primarily on work that rewards only correct answers, such as exams and quizzes. Importantly, constructing a grading system that rewards students for participation and effort has been shown to stimulate

student interest in improvement (Swinton, 2010). One strategy for focusing students on the importance of effort and practice in learning is to provide students opportunities to earn credit in a course for simply doing the work, completing assigned tasks, and engaging with the material. Assessing effort and participation can happen in a variety of ways (Bean and Peterson, 1998; Rocca, 2010). In college biology courses, clicker questions graded on participation and not correctness of responses is one strategy. Additionally, instructors can have students turn in minute papers in response to a question posed in class and reward this effort based on submission and not scientific accuracy. Perhaps most importantly, biology instructors can assign out-of-class work—case studies, concept maps, and other written assignments—that can promote student practice and focus students' attention on key ideas, while not creating more grading work for the instructor. Those out-of-class assignments can be graded quickly (and not for accuracy) based on a simple rubric that checks whether students turned the work in on time, wrote the required minimum number of words, posed the required number of questions, and/or included a prescribed number of references. In summary, one strategy for changing grading is to balance accuracy-based grading with the awarding of some proportion of the grade based on student effort and participation. Changing grading in this way has the potential to promote student practice, incentivize in-class participation, and avoid some of the documented negative consequences of grading.

Providing Opportunities for Meaningful Feedback through Self and Peer Evaluation

Instructors often perceive grading to be a separate process from teaching and learning, yet well-crafted opportunities for evaluation can be effective tools for changing students' ideas about biology. Nicol and Macfarlane-Dick (2006) argue that, just as teaching strategies are shifting away from an instructor-centered, transmissionist approach to a more collaborative approach between instructor and students, so too should classroom feedback and grading. Because feedback traditionally has been given by the instructor and transmitted to students, Nicol and Macfarlane-Dick argue that students have been deprived of opportunities to become self-regulated learners who can detect their own errors in thinking. They advocate for incorporating techniques such as self-reflection and student dialogue into the assessment process. This, they hypothesize, would create feedback that is relevant to and understood by students and would release faculty members from some of the burden of writing descriptive feedback on student submissions. Additionally, peer review and grading practices can be the basis of in-class active-learning exercises, guided by an instructor-developed rubric. For example, students may be assigned out of class homework to construct a diagram of the flow of a carbon atom from a dead body to a coyote (Ebert-May *et al*., 2003). With the development of a simple rubric, students can self- or peer-evaluate these diagrams during the next class activity to check for the inclusion of key processes, as determined by the instructor. The use of in-class peer evaluation thus allows students to see other examples of biological thinking beyond their own and that of the instructor. In addition, self-evaluation of one's own work using the instructor's rubric can build metacognitive skills in assessing one's own confusions and making self-corrections. Such evaluations need not take much time, and they have the potential to provide feedback that is meaningful and integrated into the learning process. In summary, both self- and peer-evaluation of work are avenues for providing meaningful feedback without formal grading on correctness that can positively influence students' learning (Sadler and Good, 2006; Freeman *et al*., 2007; Freeman and Parks, 2010).

Making the Move Away from Curving

As documented in the research literature, the practice of grade curving has had unfortunate and often unintended consequences for the culture of undergraduate science classrooms, pitting students against one another as opposed to creating a collaborative learning community (Tobias, 1990; Seymour and Hewitt, 1997). As such, one simple adjustment to grading would be to abandon grading on a curve. Because the practice of curving is often assumed by students to be practiced in science courses, a move away from curving would likely necessitate explicit and repeated communication with students to convey that they are competing only against themselves and not one another. Moving away from curving sets the expectation that all students have the opportunity to achieve the highest possible grade. Perhaps most importantly, a move away from curving practices in grading may remove a key remaining impediment to building a learning community in which students are expected to rely on and support one another in the learning process. In some instances, instructors may feel the need to use a curve when a large proportion of students perform poorly on a quiz or exam. However, an alternative approach would be to identify why students performed poorly and address this more specifically. For example, if the wording of an exam question was confusing for large numbers of students, then curving would not seem to be an appropriate response. Rather, excluding that question from analysis and in computing the exam grade would appear to be a more fair approach than curving. Additionally, if large numbers of students performed poorly on particular exam questions, providing opportunities for students to revisit, revise, and resubmit those answers for some credit would likely achieve the goal of not having large numbers of students fail. This would maintain the criterion-referenced grading system and additionally promote learning of the material that was not originally mastered. In summary, abandoning curving practices in undergraduate biology courses and explicitly conveying this to students could promote greater classroom community and student collaboration, while reducing well-documented negative consequences of this grading practice (Humphreys *et al*., 1982).

Becoming Skeptical about What Grades Mean

The research literature raises significant questions about what grades really measure. However, it is likely that grades will continue to be the currency of formal teaching and learning in most higher education settings for the near future. As such, perhaps the most important consideration for instructors about grading is to simply be skeptical about what grades mean. Some instructors will refuse to write letters of recommendation for students who have not achieved grades in a particular range in their course. Yet, if grades are not a reliable reflection of learning and reflect other

factors—including language proficiency, cultural background, or skills in test taking—this would seem a deeply biased practice. One practical strategy for making grading more equitable is to grade student work anonymously when possible, just as one would score assays in the laboratory blind to the treatment of the sample. The use of rubrics can also help remove bias from grading (Allen and Tanner, 2006) by increasing grading consistency. Perhaps most importantly, sharing grading rubrics with students can support them in identifying where their thinking has gone wrong and promote learning (Jonsson and Svingby, 2007; Reddy and Andrade, 2010). Much is yet to be understood about what influences students' performance in the context of formal education, and some have suggested grades may be more of a reflection of a students' ability to understand and play the game of school than anything to do with learning (Towns and Robinson, 1993; Scouller, 1998; Stanger-Hall, 2012). In summary, using tools such as rubrics and blind scoring in grading can decrease the variability and bias in grading student work. Additionally, remembering that grades are likely an inaccurate reflection of student learning can decrease assumptions instructors make about students.

IN CONCLUSION—TEACHING MORE BY GRADING LESS (OR DIFFERENTLY)

A review of the history and research on grading practices may appear to present a bleak outlook on the process of grading and its impacts on learning. However, underlying the less encouraging news about grades are numerous opportunities for faculty members to make assessment and evaluation more productive, better aligned with student learning, and less burdensome for faculty and students. Notably, many of the practices advocated in the literature would appear to involve faculty members spending less time grading. The time and energy spent on grading has been often pinpointed as a key barrier to instructors becoming more innovative in their teaching. In some cases, the demands of grading require so much instructor attention, little time remains for reflection on the structure of a course or for aspirations of pedagogical improvement. Additionally, some instructors are hesitant to develop active-learning activities—as either in-class activities or homework assignments—for fear of the onslaught of grading resulting from these new activities. However, just because students generate work does not mean instructors need to grade that work for accuracy. In fact, we have presented evidence that accuracy-based grading may, in fact, demotivate students and impede learning. Additionally, the time-consuming process of instructors marking papers and leaving comments may achieve no gain, if comments are rarely read by students. One wonders how much more student learning might occur if instructors' time spent grading was used in different ways. What if instructors spent more time planning in-class discussions of homework and simply assigned a small number of earned points to students for completing the work? What if students themselves used rubrics to examine their peers' efforts and evaluate their own work, instead of instructors spending hours and hours commenting on papers? What if students viewed their peers as resources and collaborators, as opposed to competitors in courses that employ grade curving? Implementing small changes like those described above might allow instructors to promote more student learning by grading less or at least differently than they have before.

REFERENCES

Allen D, Tanner K (2006). Rubrics: tools for making learning goals and evaluation criteria explicit for both teachers and learners. Cell Biol Educ *5*, 197–203.

Anderson VJ (2008). Grading. In: Encyclopedia of Educational Psychology, Thousand Oaks, CA: Sage.

Bagg LH (1871). Four Years at Yale, New Haven, CT: Charles C. Chatfield.

Bean JC, Peterson D (1998). Grading classroom participation. New Direct Teach Learn *1998*(74), 33–40.

Bloom BS (1968). Learning for Mastery. Instruction and Curriculum. Regional Education Laboratory for the Carolinas and Virginia, Topical Papers and Reprints, Number 1. Eval Comment *1*(2), 1–11.

Bloom BS (1976). Human Characteristics and School Learning, New York: McGraw-Hill.

Brookhart S (2004). Grading, Upper Saddle River, NJ: Pearson Education.

Brookhart SM (2008). How to Give Effective Feedback to Your Students, Alexandria, VA: Association for Supervision and Curriculum Development.

Brown University (2014). Brown's Grading System. http://brown.edu/campus-life/support/careerlab/employers/employer-resources/browns-grading-system/browns-grading-system (accessed 19 February 2014).

Bull R, Stevens J (1979). The effects of attractiveness of writer and penmanship on essay grades. J Occup Psychol *52*, 53–59.

Butler R (1988). Enhancing and undermining intrinsic motivation: the effects of task-involving and ego-involving evaluation on interest and performance. Br J Educ Psychol *58*, 1–14.

Butler R, Nisan M (1986). Effects of no feedback, task-related comments, and grades on intrinsic motivation and performance. J Educ Psychol *78*, 210.

Crisp BR (2007). Is it worth the effort? How feedback influences students' subsequent submission of assessable work. Assess Eval High Educ *32*, 571–581.

Crooks TJ (1988). The impact of classroom evaluation practices on students. Rev Educ Res *58*, 438–481.

Cureton LW (1971). The history of grading practices. NCME Measurement in Educ *2*(4), 1–8.

Dufresne RJ, Leonard WJ, Gerace WJ (2002). Making sense of students' answers to multiple-choice questions. Phys Teach *40*, 174–180.

Ebert-May D, Batzli J, Lim H (2003). Disciplinary research strategies for assessment of learning. Bioscience *53*, 1221–1228.

Eells WC (1930). Reliability of repeated grading of essay type examinations. J Educ Psychol *21*, 48.

Fajardo DM (1985). Author race, essay quality, and reverse discrimination. J Appl Social Psychol *15*, 255–268.

Farrell MJ, Gilbert N (1960). A type of bias in marking examination scripts. Br J Educ Psychol *30*, 47–52.

Finkelstein IE (1913). The Marking System in Theory and Practice, Baltimore: Warwick & York.

Freeman S *et al.* (2007). Prescribed active learning increases performance in introductory biology. Cell Biol Educ *6*, 132–139.

Freeman S, Parks JW (2010). How accurate is peer grading? CBE Life Sci Educ *9*, 482–488.

Grant D, Green WB (2013). Grades as incentives. Empirical Econom *44*, 1563–1592.

Guskey TR (1994). Making the grade: what benefits student. Educ Leadership *52*(2), 14–20.

Harter S (1978). Pleasure derived from challenge and the effects of receiving grades on children's difficulty level choices. Child Dev *49*, 788–799.

Harvard University (1832). Annual Report of the President of Harvard University to the Overseers on the State of the University for the Academic Year 1830–1831 Cambridge, UK: E. W. Metcalf.

Higgins R, Hartley P, Skelton A (2002). The conscientious consumer: reconsidering the role of assessment feedback in student learning. Stud High Educ *27*, 53–64.

Humphreys B, Johnson RT, Johnson DW (1982). Effects of cooperative, competitive, and individualistic learning on students' achievement in science class. J Res Sci Teach *19*, 351–356.

Jaschik S (2009). Imagining College without Grades, www.insidehighered.com/news/2009/01/22/grades (accessed 20 February 2014).

Johnson V (2003). Grade Inflation: A Crisis in College Education, Secaucus, NJ: Springer.

Jonsson A, Svingby G (2007). The use of scoring rubrics: reliability, validity and educational consequences. Educ Res Rev *2*, 130–144.

Kitchen E, King SH, Robison DF, Sudweeks RR, Bradshaw WS, Bell JD (2006). Rethinking exams and letter grades: how much can teachers delegate to students? CBE Life Sci Educ *5*, 270–280.

Kohn A (1999). Punished by Rewards: The Trouble with Gold Stars, Incentive Plans, A's, Praise, and Other Bribes, New York: Houghton Mifflin Harcourt.

Los Angeles Times (2009). Study critiques schools over subjective grading. Los Angeles Times, October 4, 2009. http://articles.latimes.com/2009/oct/04/nation/na-grading-policy4 (accessed 15 April 2014).

MacDonald RB (1991). Developmental students' processing of teacher feedback in composition instruction. Rev Res Dev Educ *8*(5), 1–5.

Marble WO, Winne PH, Martin JF (1978). Science achievement as a function of method and schedule of grading. J Res Sci Teach *15*, 433–440.

Meadows M, Billington L (2005). A review of the literature on marking reliability. Unpublished AQA report produced for the National Assessment Agency. http://archive.teachfind.com/qcda/orderline.qcda.gov.uk/gempdf/184962531X/QCDA104983_review_of_the_literature_on_marking_reliability.pdf.

Meyer M (1908). The grading of students. Science *28*, 243–250.

National Education Association (1971). Reporting pupil progress to parents. Res Bulletin *49* (October), 81–83.

Nicol DJ, Macfarlane-Dick D (2006). Formative assessment and self-regulated learning: a model and seven principles of good feedback practice. Stud High Educ *31*, 199–218.

Oettinger GS (2002). The effect of nonlinear incentives on performance: evidence from "Econ 101." Rev Econ Stat *84*, 509–517.

Palmer B (2010). E Is for Fail, www.slate.com/articles/news_and_politics/explainer/2010/08/e_is_for_fail.html (accessed 19 February 2014).

Paxton M (2000). A linguistic perspective on multiple choice questioning. Assess Eval High Educ *25*, 109–119.

Pulfrey C, Buchs C, Butera F (2011). Why grades engender performance-avoidance goals: the mediating role of autonomous motivation. J Educ Psychol *103*, 683.

Reddy YM, Andrade H (2010). A review of rubric use in higher education. Assess Eval High Educ *35*, 435–448.

Rocca KA (2010). Student participation in the college classroom: an extended multidisciplinary literature review. Commun Educ *59*, 185–213.

Rogers WT, Harley D (1999). An empirical comparison of three-and four-choice items and tests: susceptibility to testwiseness and internal consistency reliability. Educ Psychol Meas *59*, 234–247.

Rojstaczer S, Healy C (2012). Where A is ordinary: the evolution of American college and university grading, 1940–2009. Teachers College Rec *114*(7), 1–23.

Sadler PM, Good E (2006). The impact of self-and peer-grading on student learning. Educ Assess *11*, 1–31.

Schneider J, Hutt E (2013). Making the grade: a history of the A–F marking scheme. J Curric Stud, 1–24.

Scouller K (1998). The influence of assessment method on students' learning approaches: multiple choice question examination versus assignment essay. High Educ *35*, 453–472.

Seymour E, Hewitt N (1997). Talking about Leaving: Why Undergraduates Leave the Sciences, Boulder, CO: Westview.

Sinclair HK, Cleland JA (2007). Undergraduate medical students: who seeks formative feedback? Med Educ *41*, 580–582.

Smallwood ML (1935). An Historical Study of Examinations and Grading Systems in Early American Universities: A Critical Study of the Original Records of Harvard, William and Mary, Yale, Mount Holyoke, and Michigan from Their Founding to 1900, vol. 24, Cambridge, MA: Harvard University Press.

Spear M (1996). The influence of halo effects upon teachers' assessments of written work. Res Educ *1996*(56), 85–86.

Spear MG (1984). The biasing influence of pupil sex in a science marking exercise. Res Sci Technol Educ *2*, 55–60.

Stanger-Hall KF (2012). Multiple-choice exams: an obstacle for higher-level thinking in introductory science classes. CBE Life Sci Educ *11*, 294–306.

Starch D (1913). Reliability and distribution of grades. Science *38*, 630–636.

Starch D, Elliott EC (1912). Reliability of the grading of high-school work in English. School Rev *20*, 442–457.

Starch D, Elliott EC (1913). Reliability of grading work in mathematics. School Rev *21*, 254–259.

Stiles E (1901). The Literary Diary of Ezra Stiles . . . President of Yale College, New York: Scribner's.

Swinton OH (2010). The effect of effort grading on learning. Econ Educ Rev *29*, 1176–1182.

Tobias S (1990). They're Not Dumb, They're Different: Stalking the Second Tier, Tucson, AZ: Research Corporation.

Towns MH, Robinson WR (1993). Student use of test-wiseness strategies in solving multiple-choice chemistry examinations. J Res Sci Teach *30*, 709–722.

Wainer H, Thissen D (1993). Combining multiple-choice and constructed-response test scores: toward a Marxist theory of test construction. Appl Measure Educ *6*, 103–118.

Weaver MR (2006). Do students value feedback? Student perceptions of tutors' written responses. Assess Eval High Educ *31*, 379–394.

Weigle SC (1999). Investigating rater/prompt interactions in writing assessment: quantitative and qualitative approaches. Assessing Writing *6*, 145–178.

Weld LD (1917). A standard of interpretation of numerical grades. School Rev *25*, 412–421.

Yale University (2013). Revised Report of the Ad Hoc Committee on Grading. http://yalecollege.yale.edu/sites/default/files/2_Report%20from%20Ad%20Hoc%20Committee%20on%20Grading%5B2%5D.pdf (accessed 15 April 2014).

Zimmerman DW, Williams RH (2003). A new look at the influence of guessing on the reliability of multiple-choice tests. Appl Psychol Measure *27*, 357–371.