Avoiding Unintended Consequences in Grading Reform

Dylan Wiliam

Before we enact grading reforms, let's understand why the policies we're scuttling were adopted in the first place.

In 1929, British author G. K. Chesterton put forward the principle that changes in policies should not be made until the reasons behind the previous policies were understood:

> In the matter of reforming things, as distinct from deforming them, there is one plain and simple principle; a principle which will probably be called a paradox. There exists in such a case a certain institution or law; let us say, for the sake of simplicity, a fence or gate erected across a road. The more modern type of reformer goes gaily up to it and says, "I don't see the use of this; let us clear it away." To which the more intelligent type of reformer will do well to answer: "If you don't see the use of it, I certainly won't let you clear it away. Go away and think. Then, when you can come back and tell me that you do see the use of it, I may allow you to destroy it." (1929/1990, p. 157)

As schools and districts across the United States consider changes to long-established grading practices, the idea that we need to understand the reasons for those policies before we sweep them away could, I believe, help us reduce the number and severity of unintended consequences of grading reform. Moreover, such a perspective helps us understand that all policy decisions—especially those involving grading—involve trade-offs, and it's better to make such trade-offs explicitly, rather than have them emerge as unintended consequences of other decisions. Let's examine these points by looking at three controversial issues in grading.

Giving Zeroes for Missing Work

In the United States, the practice of awarding zeroes for missing work is still widespread, although hotly debated. At first sight, this seems like an unjustifiable policy. It has been conventional in the United States to regard 70 percent as a passing score for at least the last 120 years (Smallwood, 1935), even though tests tend to be more reliable when passing scores are set around 50 percent (Hambleton, Swaminathan, & Rogers, 1991). When the passing score is 70 percent, giving a student a zero for missing work seems like a disproportionate penalty. With, let's say, eight grades in a marking period, a zero for one missing piece of work means that student would have to get a *B* on every other assignment in the marking period to get a barely passing grade.

So it's hardly surprising that almost all books of advice about grading practices are unyielding in their condemnation of the practice of giving zeroes for missing work. But if we take Chesterton's advice, before we ban the practice of assigning zeroes, we should understand why the practice arose in the first place.

Giving zeroes for missing work means that the grade a student is awarded is unlikely to describe the quality of a student's work or learning. In other words, on an *evidential* basis, it makes no sense. But the disproportionate nature of the penalty creates a strong incentive for the student to hand in *something*. So to the extent that such a policy can be defended, it is justified by its *consequences*. And of course, because of the disproportionate nature of the penalty, the policy works best when it never needs to be enforced.

Those who criticize awarding zeroes for missing work are placing greater emphasis on the *meaning* of the grade, while those who defend the policy are emphasizing the *consequences*. These are really two ends of a continuum rather than two different categories. For example, if our prime concern was that the grade represented the quality of the student's work, instead of giving any missing pieces of work a zero, we could substitute the average score of all the pieces of work that *were* submitted. Many educators object to this, saying this would mean a student would "get away with" not submitting the work. But that is, in effect, an argument about consequences rather than evidence. As a compromise

between these two extremes, missing work could be scored as 50 percent, which lessens—but doesn't entirely dispense with—the penalty for non-submission.

There are two important lessons here. First, the perspective afforded by Chesterton's fence forces us to consider why practices of which we disapprove have been adopted. If administrators or school boards are going to prevent teachers from using sanctions that encouraged students to submit work, then they should consider what alternative incentives to submit work they're going to offer students. Second, by looking at the trade-off between emphasizing meaning and emphasizing consequences, we might find better compromises, or at least compromises we're happier with, such as the idea of giving a notional score above zero for missing work.

Similar arguments apply to the common, but often-criticized, practice of deducting points for tardy submission of assigned work. From an evidential perspective, this makes no sense at all: "The work you submitted suggests that you understand the gas laws well, but you submitted it late, so in fact you don't." But from a consequential perspective, it does make some sense. If school districts are going to mandate policies that prevent teachers from using this particular sanction for late submission of work, they should provide teachers with some alternatives.

Using Percentage Scores Rather Than Grades

Another continuing debate is whether student achievement should be reported on a 100-point scale or by using the familiar letter grades, *A, B, C, D*, and *F*. As many authors, notably Guskey (2014), have pointed out, percentage scales make little sense in practical terms. After all, the idea that a score of 82 is somehow better than a score of 81 is silly. The error of measure on a typical teacher-produced test is likely to be at least five percentage points. This means that on any single test, approximately one-third of the students in a class will get a score that differs by 5 points from the score they would have gotten with a *perfectly* reliable test. Worse, one student in a class of 25 will get a score that differs from the score she or he would have received on a perfectly reliable test by more than 10 points. Unfortunately, we won't know which student it is—or whether the score she or he got was higher or lower than the score with a perfectly reliable test. So, given this inaccuracy, why do so many teachers insist on using percentage scores?

One reason is that using grades instead of percentage scores involves discarding information. Imagine an idealized student who, if she took 10 perfectly reliable assessments over a quarter, would score a 91 on each of them. Because the assessments she actually takes *aren't* perfectly reliable, she gets the following scores (assuming a value of 0.80 for the reliability of the assessment):

90, 92, 94, 89, 91, 86, 98, 87, 89, 91

Teacher A's policy is that each score is converted into a grade and the grades are averaged at the end of the quarter. So the student gets *A, A, A, B, A, B, A, B, B, A*, and her grade point average for the quarter is 3.6.

Teacher B's policy is to average the scores over the quarter and then convert that average into a grade at the quarter's end. The average of the 10 scores is 91, so the grade the student gets for the quarter is a straight *A*and the GPA is 4.0.

The grades the two teachers award the student differ because Teacher A, by converting each score into a grade, is not including key information, which prevents the variation in scores caused by unreliability from averaging out. This seems to penalize the student unfairly.

A second reason giving letter grades may not be ideal is that there seems to be a tendency, among students, teachers, and parents, to over-interpret the difference between grades, to think of an *A* as qualitatively different from a *B*, even though the difference between the two may just be half a point. Scores do suffer from spurious precision, but grades often suffer from spurious accuracy.

Yet another argument in favor of percentage scores over letter grades is that scores make it easier to report errors of measurement in an understandable way. No measurement, in education or elsewhere, is perfectly reliable. Informed assessment use requires that users of assessment information know how big the measurement error is. In the example above, the standard error of measurement was five points, so for most of the students, the score they get will be within

five points of the score they would've gotten with a perfectly reliable test. Given the widespread use of margins of error in reporting the results of opinion polls, parents and caregivers are likely to understand the idea of a score of 67 plus or minus five points much more easily than they'll grasp a *C* plus or minus half a grade.

As a side note, while using percentage scores makes reporting measurement errors easier, discussing measurement error in this way raises problems for teachers—largely because we've often pretended our assessments are perfect. We need to be more honest with students and parents. I would welcome the following kinds of exchange between a parent and a teacher:

PARENT: What's the passing score on this test?

TEACHER: 70

PARENT: What score did my child get?

TEACHER: 67, plus or minus 5 points.

PARENT: So did my child pass?

TEACHER: Probably not on this test. But it's possible he knew the material adequately—although barely. It's hard to know from one test score.

PARENT: Why don't you know?

TEACHER: Because no test is perfectly reliable.

PARENT: Can't you make the test more reliable?

TEACHER: We could, but reducing the error of measurement to, say, two points would mean making the test seven times longer. We want to make better use of our time in the classroom.

There are two important points here that everyone involved in education should understand. First, error is an inevitable part of measurement; pretending it doesn't exist is dishonest. Second—and perhaps more important—current levels of measurement error may well be optimal. When people first understand how large errors of measurement in educational assessment are in practice, they tend to urge that something must be done to reduce these errors, but, again, there's always a trade-off here. We *can* reduce errors of measurement, but we'd need large increases in total testing time to do so. We're better off spending less time testing, understanding the degree of error in our results, and making sure we don't place more weight on the results than is warranted by their accuracy.

Averaging Grades Over a Marking Period

Consider two students who, over an eight-week marking period, get the following grades in math:

Jeanne: *C C C C A A A A*

Angela: *A A A A C C C C*

Using the traditional method of combining grades—simple averaging—both students would get a *B* overall for the marking period—which makes little sense. After all, Jeanne, after a slow start, handed in consistently high-quality work. Particularly in a cumulative subject like math, it could be argued that the most appropriate grade for this student would be an *A*. Angela, on the other hand, has clearly failed to master the material for this marking period. At least in math, the most appropriate grade for her may be a *C*.

Arguments like this have led many assessment experts to suggest that grades should reflect the learning achieved by the *end* of the marking period, rather than the average quality of the work across the whole period. There are many ways this could be done, including placing greater weight on assessments completed toward the end of the marking period, especially when the assessment involves "capstone" projects that integrate the students' learning across the whole marking period. So why do some teachers insist on a straight average of all the grades?

One answer is that there are some students who, knowing they can redeem poor quality work in the first half of the marking period with better work in the second half, may decide not to put much effort into the earlier assignments. Students often overestimate their ability to "pull out all the stops" toward the end of the marking period and leave themselves with too much to do at the end.

However, there is a more important reason why teachers might want to ensure that students work steadily over the whole marking period, one related to what we're learning about how human memory works. In memory research, *performance* describes how well a learner completes a learning task, while *learning* is reserved for the long-term changes in capability that result (Kirschner, Sweller, & Clark, 2006). If students satisfactorily complete a learning task and can do what the task was designed to teach them today, but not in two weeks' time, then no learning has taken place. And researchers have discovered that the quality of performance in a learning task is no guarantee of long-term learning. Indeed, Robert and Elizabeth Bjork (1992) have shown that there is often an inverse relationship: When learners encounter "desirable difficulties" while performing the learning task, they remember more of what the task was intended to teach.

This is why what's called "distributed practice" is preferable. If students have a math test on Friday, they could prepare for it by studying for an hour on Thursday evening (what's called "massed" or "blocked" practice). Alternatively, they could prepare by studying for 15 minutes on Monday evening, 15 minutes on Tuesday evening, 15 minutes on Wednesday evening, and so on ("distributed" practice). Both patterns involve a total of 60 minutes of practice, but because students will have forgotten some of what they learned (or relearned) on Monday by the time they review this material (and newer concepts) on Tuesday, "desirable difficulties" will arise for students in retrieving what they did the previous day—which will result in greater long-term learning.

So if the way grades are combined over a marking period allows students to leave everything to the last moment, they may get a high grade for their work. But, because their studying has been concentrated toward the end of the marking period, they're likely to remember less of the content over the long-term. In contrast, if every assignment students complete "counts," they have to distribute their study time over the whole marking period, resulting in better long-term learning. On an evidential basis, it may make sense to place greater weight on evidence from the end of the marking period, but in terms of consequences, learning is likely to be enhanced if work over the whole quarter or semester contributes to the final grade.

No Simple Process

At first sight, grading seems like a simple process. Students complete assignments, we look at what they did, and we assign a number or a letter to that work. From such a perspective, what matters is whether the grade fairly represents that work. But grades also send messages and have social consequences—for students, teachers, and families. The *consequences* of grading practices are often at least as important as their meanings—and sometimes more important.

There will never be a perfect grading system. Any system will be the result of a messy series of trade-offs, between how well the grades describe the students' work and how people react to those grades, between precision and accuracy, between the short-term and the long-term, and so on. I have discussed three such issues here, but the more I study assessment, the more issues I find. Indeed, I've realized that if I cannot identify the trade-offs, I need to look harder, because they are definitely there. The important thing is that we are aware of these trade-offs and make them in a principled way.

Heeding Chesterton's advice, we should also not be too quick to dismiss common practices that we think ill-advised. If they've been in widespread use for considerable periods of time, we should try to understand why they are so prevalent before we rush to abandon them. If we start from what philosophers call a "principle of charity"—assuming the best of people, their actions, and their motives—then we'll make smarter decisions about grading—and much else in education.

Reflect & Discuss

➛ Did this article make you think differently about grading reform initiatives? Why or why not?

➛ Has your school or district done a sufficient job in considering trade-offs that come with grading (or other) changes?

➛ In light of possible unintended consequences, how should grading-reform advocates proceed?

References

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. F. Healy, S. M. Kosslyn, & R. M. Shiffrin (Eds.), *From learning processes to cognitive processes: Essays in honor of William K. Estes* (vol. 2, pp. 35–67). Hillsdale, NJ: Lawrence Erlbaum Associates.

Chesterton, G. K. (1929/1990). *The drift from domesticity. The collected works of G. K. Chesterton* (vol. 3, pp. 157–164). San Francisco, CA: Ignatius Press.

Guskey, T. R. (2014). *On your mark: Challenging the conventions of grading and reporting.* Bloomington, IN: Solution Tree.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist, 41*(2), 75–86.

Smallwood, M. L. (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*. Cambridge, MA: Harvard University Press.

Dylan Wiliam is Emeritus Professor of Educational Assessment at University College of London. He is the author of many books, including *Creating the Schools Our Children Need* (Learning Sciences International, 2018).