CHAPTER 2

THE EVIDENCE FOR GRADING REFORM

ARCHWAY PUBLISHING

- ❖ Evaluating Educational Research
- ❖ A Brief History of Grading
- ❖ A Century of Research on Grading
- ❖ Mathematical Certainties
- ❖ The Supremacy of Local Evidence

I HAVE OFTEN ASKED TEACHERS AND EDUCATIONAL LEADERS WHY they distrust educational research. After all, the only alternative to research is personal opinion and conjecture. Nevertheless, there is a predictable set of explanations for why research is the subject of suspicion and ridicule:

- It doesn't apply to me because the research was in another school with different characteristics.
- The researchers had a conflict of interest because they were paid by the producers of the program they are trying to sell.

- The characteristics of the students are different from my students.
- The teachers in the study had more resources, a different union contract, better leadership, and more time than we have.
- The sample size was too small.
- Studies contradict one another, so no one knows what to believe.
- You can always find a researcher to say anything.

The list goes on, but the conclusion is consistent—research can't be trusted. Imagine applying this logic to medical research. It is true that medical research is often subject to conflicts of interest; has small sample sizes; and is conducted in hospitals and communities far away from the reader in conditions with different resources, schedules, and unions. Moreover, there is contrasting research, with one study affirming the efficacy of a drug or surgical procedure and another study challenging those conclusions. Indeed, it is through the process of contention—different researchers coming to different conclusions, followed by careful testing of alternative hypotheses—that scientific truth emerges. Despite the cacophony of different findings in medical research, practitioners look at the weight of the evidence and make the best decisions based on the best available evidence. It's not perfect, but it is better than witchcraft or the physicians' personal preferences.

Grading reform is plagued by personal preferences and experiences that supplant evidence. Therefore, this chapter begins with a consideration of how to evaluate research claims. We then consider the history of grading through almost four hundred years of schooling in the U.S. and a much longer history in other nations. There is no lack of research in grading, and thanks to extraordinary efforts of scholars like Susan Brookhart, Tom Guskey, and others who will be cited in the following pages, we can synthesize more than a century of research in grading reform. But not every question about grading is a matter of field research. Sometimes the evidence is a matter of mathematical certainty. We need not commission a federally funded

ARCHWAY PUBLISHING

study, for example, to determine that the area of a square is one side to the second power. In the case of grading reform, we need only the simplest number operations to demonstrate the inaccuracy of some grading systems. The chapter concludes with the supremacy of local evidence, the single best way to demonstrate that external research has direct applicability to local schools.

EVALUATING EDUCATIONAL RESEARCH

There are five levels of evidence in educational research: belief, personal experience, collective experience, systematic comparison, and preponderance of the evidence (Reeves, 2020a). Let us consider each of these levels of research.

Level 1. Belief:

I believe it! The First Amendment to the U.S. Constitution enshrines the right of freedom of speech, assembly, and religion, allowing us to believe and speak our own truths. With few exceptions, such as inciting violence or defamation, these freedoms protect speech and beliefs ranging from the well-reasoned to the banal, from evidence-based conclusions to mysticism. But as the late Senator Daniel Patrick Moynihan reminded us, "Everyone is entitled to his own beliefs, but not to his own facts." (Moynihan Estate, 2010).

There are people of intelligence and good will who believe that grading as punishment is an effective method of behavioral modification for students, just as there are people with neither intelligence nor good will who believe that beating children will improve their behavior. Whatever protections these beliefs enjoy in the fora of public discourse, they have little place in determining educational policies.

Level 2. Personal Experience:

We are all informed by our personal experiences. A single mishap near a window or in a creek can lead to a lifelong fear of heights or

ARCHWAY PUBLISHING

water. Success in conquering a mountaintop or victory in a swim meet might convince us that mountaineering or water sports are ideal for everyone. When we judge from personal experience, we are making the fundamental attribution error, a pervasive challenge to which all of us are susceptible (McLeod, 2018). The fundamental attribution error occurs when we overemphasize data that agree with our theories about life and carefully avoid including data that do not conform to our theories: "I was terrified of low grades and point deductions, and that is why I became a successful college student and a distinguished educator. For the past thirty years I've used low grades and point deductions to motivate my students, but this year I have more late work than ever!" No one doubts how meaningful personal experience is and how it shapes sincerely held beliefs.

The problem is when we take our personal experience—a sample size of one—and attempt to generalize that observation to the entire world. The fundamental attribution error is what gives rise to the legions of leadership books that are based on the author's personal experience but systematically ignore the data that do not conform to the writer's theories. For example, former General Electric Chief Executive Jack Welch was lionized for his business acumen, published best-selling books (Welch, 2003), and provided six-figure lectures sharing his wisdom. But his eager buyers and listeners ignored the fact that his much-ballyhooed forced ranking system was later repudiated not only by other businesses but also by General Electric itself. More importantly, many of Welch's "straight from the gut" decisions, supported more by hubris than evidence, were unsuccessful, leaving the shareholders and employees of General Electric worse off after Welch had retired (Gelles, 2022).

Level 3. Collective Experience:

One rejoinder to the suggestion that one is over-inferring from one's personal experience is the claim that the purveyor of the fundamental attribution error is joined by other colleagues. "It's not just me—it's the entire math department!" I have learned to fact-check assertions such as, "All the teachers are saying…." and "Everybody agrees with me on

gutter

ARCHWAY PUBLISHING

this." Most people don't like conflict, and on emotional issues such as grading reform, it is frequently the case that those who are embracing practices that lead to greater student success and accompanying higher grades just keep their heads down, get the job done, and stay clear of the controversy swirling around the school. Their professional and personal modesty leads them to avoid any victory laps when they and their students succeed, preferring the quiet anonymity of success to a public declaration of success that might put them in the eye of the storm. Nevertheless, even if 100% of the faculty believed with confidence that Boyle's Law was wrong, I'd still side with the old scientist every time. Even against the noise of the crowd, educational policy must be based on fact rather than opinion.

Level 4. Systematic Comparison:

When teachers engage in systematic comparison, they have as close to a perfect experiment as we find in education. Consider the teacher who decides halfway through the year to change grading systems, perhaps by getting rid of the average and grading performance rather than compliance. All the conditions are the same—the same students from the same home environments, the same per-pupil funding, the same school building and classroom, the same bargaining agreement, the same behavior and attendance, the same curriculum, and the same teacher. The only thing that changed was the grading practice. When the second half of the year shows a lower failure rate, better student engagement, fewer late assignments, and an improved classroom climate and culture, then one can assert with confidence that it was the change in grading practice that caused these results. It couldn't have been a change in students, home environment, attendance, or any other factors commonly associated with academic outcomes, because all of those variables were the same.

In my experience, this is the single best way to implement change, because we are not asking teachers to "trust the research" and we are not subject to the common complaints about educational research that were reviewed earlier in this chapter. It was the teachers and their professional practices that demonstrably caused the change in

ARCHWAY PUBLISHING

student results. Plenty of people will not believe a writer or researcher, including your faithful author, but they will believe their colleague down the hall whose evidence from systematic comparison of the same students to the same students show clearly the impact of effective grading.

Level 5. Preponderance of the Evidence:

A common criticism of educational research is that one can find a researcher to say anything, and that contradictory findings in research mean that we should not believe anything. A more thoughtful approach is to consider a wide variety of research studies, including different methods; different researchers; different student populations; different settings (urban, rural, and suburban); and different cultural, linguistic, and national contexts. If we find consistency, then there is a strong probability that the finding has empirical support. Consider the examples of efficacy and effective feedback as very strong factors that contribute to student success. Quantitative studies of more than 2,000 school plans placed efficacy as the most salient variable related to gains in student achievement in elementary, middle, and high schools—including a variety of school settings and student populations (Reeves, 2013). Case studies and qualitative methods came to similar conclusions, as did meta-analyses (studies of studies) and syntheses of meta-analyses (Donahoo, Hattie, & Fells, 2018). While agreement on the power of efficacy may not be universal, this critical factor in teaching and leadership certainly passes the test of the preponderance of the evidence.

There are legitimate concerns about educational research and there are, to be sure, many deeply flawed articles and books that claim to be evidence-based. But the answer to these challenges is not to supplant research with myth and personal preference but to be discerning about the research that informs our policy and practice. Any individual research study is a pebble on the mountain of research. When we are guided by the standard of the preponderance of the evidence, then we are focused on the mountain, not the pebble.

ARCHWAY PUBLISHING

gutter

A BRIEF HISTORY OF GRADING

Grades have been around for a very long time. Though we have no records from the Lyceum in which Aristotle taught in 335 BCE, the conversations of his student Plato leave little doubt that students were evaluated and sometimes praised, and other times found wanting. In the U.S., Harvard graded students in 1646, and Yale standardized the practice in 1785 (Schinske & Tanne, 2014). Over the centuries, grades have been in Latin and English, letters and numbers—typically with a range of three or four levels of attainment. The 100-point scale is a relatively recent innovation, arising in the late 1800s and early 1900s, coincident with the rise of intelligence tests, also using the 100-point scale. Thus, teachers from the time of Aristotle to the present age have found it useful to evaluate the work of their students and report on their progress.

A CENTURY OF RESEARCH ON GRADING

In a review of more than a century of studies on grading, a nuanced picture of grades emerges (Brookhart & Guskey, 2019). On the one hand, high school grades can be a better predictor of college success than standardized tests. On the other hand, there are consistent findings that grades are inconsistent and unreliable. The same student work could receive widely varying grades from different teachers. This is consistent with the research discussed in the first chapter and is at the heart of the persistent violation of the value of fairness. While the teachers and administrators whom I interviewed were consistent in claiming to know that variation in grades was due to variations in student performance, attendance, prior knowledge, and organization—the consistent research finding is that variation in grades is due to idiosyncratic grading policies and practices of teachers.

In order to improve the reliability of grading, researchers found that two key practices are essential. First, teachers must be clear and specific about the quality of student work that they expect.

ARCHWAY PUBLISHING

Whether they are describing a complex performance task or simply an instruction to "show your work" in math, the expectations of teachers vary widely. Second, the fewer the marks that teachers can use, the more likely they are to be consistent—and the more marks they have available to assign to students, the less likely they are to be consistent. Thus, when the only grades available are A, B, C, D, and F—or 4, 3, 2, 1, and zero—then teachers are far more likely to be consistent in their evaluation of student work than when they grade on the 100-point scale. While the 100-point scale offers the illusion of precision, it provides the opposite, as the same student work can receive widely varying grades. Even on supposedly objective measures, such as multiple choice items in which there is theoretically only one correct answer, the fact is that every measurement includes both the estimated value and error. Even the most skillfully designed test items created by highly trained psychometricians, such as those on the SAT, ACT, and similar exams, have an element of error (Jaschik, 2019). Assessments that are subjective—such as essays or the requirements of students to provide graphs, charts, or written explanations—can be highly reliable, provided that the scoring criteria are so clear that different scorers almost always come to the same conclusion about the rating of the student's work.

In order for teachers to improve the consistency, and hence the fairness, of their scoring, they must agree on the criteria by which student work will be judged. First, they must agree that it is the student work, not the individual student, who is being judged. This leaves no room for leniency, "After all, this is Randall, and he tried really hard," or harshness, "Juanita was absent half the time and didn't pay attention when she was in class, so I don't think she deserves a break on this test." The only thing that matters is the work, and not the personality, work habits, attendance, behavior, or any other characteristics of the students outside of their actual work. I have heard teachers and administrators say that they were not comfortable giving passing grades in a social studies class to students who, despite earning B-grades on their tests, failed to submit homework and did not seem motivated. The reluctance of teachers to award passing

ARCHWAY PUBLISHING

gutter

grades to these students might be justified if the line on the report card reads "homework" or "engagement". But if the line reads "social studies" and the student earned a B, then B is the grade that should be recorded.

The best practice for focusing on student work rather than the individual student is for teachers to practice scoring anonymous work. This removes any consideration of the individual characteristics of the student and forces the scorers to consider only the work. When teachers practice scoring anonymous work, there are two beneficial effects. First, in experiments I have conducted with California educators, the consistency increases remarkably from 20% to 92%. Second, the speed of scoring increases dramatically, from 45 minutes to 9 minutes (Reeves, 2020b).

MATHEMATICAL CERTAINTIES

Not every controversy in grading requires field research. Sometimes the way to resolve an argument is not a study; rather, it is a mathematical calculation. For example, the combination of the 100-point scale and the use of the average to calculate final grades incorporates two glaring mathematical errors. The first error is inconsistent intervals. As we previously discussed, when we calculate grade-point-averages, then A = 4; B = 3; C = 2; D = 1; and F = 0, there is a one-point difference between each grade and the intervals between the letters of A, B, C, D, and F are equal to the intervals between the numbers of 4, 3, 2, 1, and 0. But when teachers use the 100-point scale, where A = 90; B = 80; C = 70; D = 60; and F is anything lower than 60; then the interval between the grades of A, B, C, and D is ten points—but the interval from D to F can be 60—as if failing to turn work in was six times worse than turning in wretched work that deserved a grade of D.

Another grading error that is discoverable by simple mathematical analysis is the use of the average to calculate final scores. The average, or arithmetic mean, is calculated by dividing the sum of all grades

ARCHWAY PUBLISHING

earned by the number of grades earned. Consider the following example:

Homework #1	0
Homework #2	0
Homework #3	0
Homework #4	0
Homework #5	0
Homework #6	0
Homework #7	0
Test #1	80
Test #2	80

gutter

ARCHWAY PUBLISHING

The sum of all scores is 160, and 160 divided by 9—the total number of grades awarded—is 17.8, a solid F. What if the tests are weighted more heavily than homework? Perhaps the test should count four times as much as homework. So the sum of scores is now 4 times 160, or 640. This sum divided by 9 equals 71, or a grade of C. But whatever the weighting, both of these examples ignore the evidence that this student is doing B work. We do not punish prior failures when, for example, students take the driver's test after seven failures. We only consider their performance on the eighth test—the one that they pass. The same is true for pilots and brain surgeons who, we hope, make lots of mistakes in *simulated* flights and surgeries before they actually fly a plane with passengers or operate on a live human.

When this interval error is combined with the use of the average to calculate the final grade, then the zero becomes the academic death penalty. No matter how hard students work, no matter how much they persevere, no matter how resilient they are, no matter how much they use teacher feedback to improve, they will fail. This explains why so

many students in the last month of any school year simply give up and, in many cases, stop coming to school. They can see that all the rhetoric of hard work, resilience, and social and emotional learning was without substance, because they will be punished for their failures three months ago. As an example of the homework and test score results above, a student in a school in the Rocky Mountains completed the semester with two tests earning a B. But because the student had missed seven homework assignments, each of which earned a zero, the average of those zeroes overwhelmed the final performance, and the final average was an F. Because the student was clearly proficient at the end of the semester, teachers and administrators argued whether the student should perhaps receive a passing grade of D. Here is what they did not discuss:

1. If a student can achieve a B on the tests at the end of the semester without having turned in any homework, what does that say about the value and relevance of the homework?
2. If a student can achieve a B on the tests at the end of the semester without studying or practicing, then what does that say about the rigor of those tests?

In this district, those questions were never considered, because the focus was on the slovenly habits of the student and the desire of the teachers to punish them.

ARCHWAY PUBLISHING

THE SUPREMACY OF LOCAL EVIDENCE

No matter how well-founded the research and mathematical logic, external evidence alone does not necessarily persuade reluctant teachers and administrators. The best evidence is local evidence—that is, compelling testimony from colleagues that show that grading reform is not only associated with lower failure rates but also with better discipline, attendance, classroom climate, and school culture. A veteran teacher can say with authority, "This really works; I have fewer

gutter

failures, and that means fewer sections of repeaters, more electives, and more advanced classes. And on top of all that, I have the least incidences of discipline problems and office referrals in the past decade. This works with our schedule, our culture, and our bargaining agreement." This testimony is worth a thousand speeches from outside experts and dozens of research studies that were conducted elsewhere. In order for teachers to engage in their own field research, I have found it most useful to keep the format very simple—no dissertations, no PowerPoint, no glitter—just a simple format that every teacher can display on a three-panel board similar to those used by students in a science fair.

Left-Hand Panel – Challenge
It might be the number of failures in ninth grade math, the percentage of students with chronic absenteeism, or the number of parents failing to respond to teacher communication. It is important to allow teachers to select a challenge that is directly relevant to them, provided that the challenge is expressed in clear and measurable terms.

Middle Panel – Practice
This describes exactly what the teacher did. It might be, for example, calculating the final grade based on student performance at the end of the semester rather than the average; it might be the change from assigning homework to getting practice done during class periods, or it might be the use of a menu system so that students can select different ways to show proficiency in the class. The key is that the professional practice is directly linked to the challenge and the practice can be monitored and demonstrated to colleagues.

Right-Hand Panel – Results
The results must be directly related to the challenge in the left-hand panel and should be displayed in a clear and easily understood manner. For example, if the issue is student failure,

ARCHWAY PUBLISHING

gutter

a simple two-bar chart will show the number of failures before the teacher's changes in professional practice and the number of failures after that change. Sometimes the results will be qualitative, such as comparing writing samples from students before a system of multi-disciplinary writing and clear-scoring rubrics to writing samples from the same students after those improvements were implemented.

While the challenges teachers face may be widely variable, the format of the science fair presentations should be consistent: Challenge, Practice, and Results. We use the label "science fair" because that is precisely what it looks like. Teachers or teacher teams share three-panel displays. While the content is unique, the format is consistent.

- The left panel, "Challenge", shows the initial problems the teachers are facing: low third grade reading scores, too many failures in ninth-grade math, or inadequate parent engagement.
- The middle panel shows the professional "Practice": changing the way practice is conducted from homework to in-class, changing the grading calculation from the average to final scores, changing parent communication from e-mails to Vimeos in the parents' language.
- The right column shows "Results": third-grade reading achievement rose from 33% proficient to 75% achievement, ninth-grade math failures declined from 52% to 9%, parent engagement tripled compared to the previous year.

* * *

In this chapter we considered the reasons why so many practitioners doubt educational research, and we found that the best research is that which is conducted locally. We also reviewed the role that mathematical precision plays in discussions of grading reform. The illusion of mathematical certainty comes from the 100-point scale,

ARCHWAY PUBLISHING

which in fact, leads to greater errors rather than precision in grading. Finally, we considered the role of local research by which teachers can share their results with colleagues to make sustainable changes in grading practices and policies.

In the next chapter we will consider the essence of effective grading practices. The changes need not be complex or overwhelming. Small changes, we will learn, can lead to big results for both students and faculty.

QUESTIONS FOR REFLECTION:

1. Why do you and many of your colleagues distrust educational research?
2. What are local examples of improved grading practices that might be shared more widely?
3. What are the barriers in your school and district to sharing best practices and holding a science fair?

ARCHWAY PUBLISHING

gutter