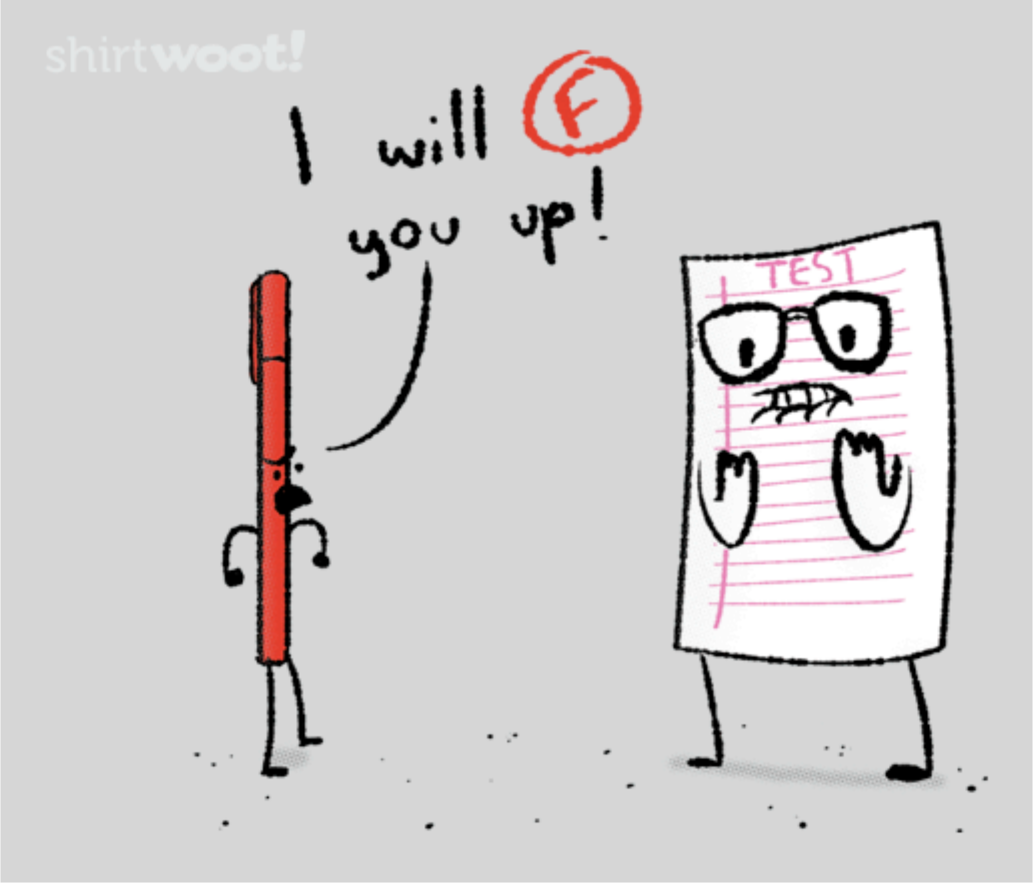

shirtwoot!
I will Ⓕ you up!
TEST

Grades Suck

Morgan C. Benton

© 2015 Morgan C. Benton. Some rights reserved.

This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License. To view a copy of this license, visit:
http://creativecommons.org/licenses/by-nc-nd/4.0/.

The foreword will go here...

Grades Suck

Chapter 1

Why Grades Suck

Use the "grandmother's roast" story here, i.e. successive generations cut off the ends of the roast, not knowing why, because that's the way that grandma did it. When you ask grandma she says it was because her pot wasn't big enough to fit the whole roast.

Chapter 2
The History of Grades

You know the good ol' days weren't always good,
and tomorrow ain't as bad as it seems.
--Billy Joel, Keeping the Faith

It is difficult for anyone who has gone through the educational system in recent decades to imagine a time when grades didn't exist. Grades have become deeply embedded into almost all aspects of the teaching and learning process. When people write about grading these days, they don't ask *if* we should be grading our students, only *how*. Their questions revolve around issues such as whether or not it is appropriate to grade on a curve, what a reasonable grade distribution should be, what is an appropriate cutoff GPA for admission, or the perennially popular topic of "grade inflation." The fact that grades must be given at all is taken for granted and never called into question. For all practical purposes, the answer to that question was settled a long time ago, and it seems almost silly to ask it now--*of course* we should be using grades to evaluate our students' learning.

It is easy to believe that grades have always been given. It doesn't seem too far-fetched to expect that some archaeologist exploring some ancient Greek ruin would uncover a scroll with Aristotle's report card on it, or find a hieroglyphic "A" carved into the wall of some as-of-yet-undiscovered chamber in one of the pyramids. We assume that even if grades didn't always exist, the people who invented the practice must have already considered all of the options, ironed out the kinks, and selected the practice that would best support learning and our educational system. There is always a tendency to look at the past through rose-colored glasses and to believe that somehow, things used to be "better."

As it turns out, however, such is not the case. The goal of this chapter is to illustrate that no "golden age" of grading ever existed. We will trace the history of the practice of grading in the United States since its origins up through the present day. Along the way we'll highlight the debates that went on among teachers and scholars about grading, and highlight that many or most of these debates have never really been resolved. Although we seldom hear of people having debates anymore about whether or not we should be grading, it is not correct to believe that the reason we don't hear them is that all the questions were answered. They never were. Rather, it seems that either we couldn't think of a better way to do things, or we just got tired of arguing about it, or perhaps both.

What Was Schooling Like Before Grades?

Schools didn't always assign grades. In Puritan colonial New England, the primary goal of education was to make it possible for every citizen to read the Bible, and as such, while most boys and girls were taught to read, less attention was paid to learning to write. The majority of schooling took place in the home, but in more populated areas, community schools were established. It was the primary responsibility of the father, or male head of household, to catechize[1] the children in the home (Moran & Vinovskis, 1985). The primary measure of success of such education was whether each student became a morally and socially upstanding citizen.

Colleges were similar. According to Rudolph (1977, p145):

> *The colonial college student was essentially ungraded and unexamined. (...) public oral examinations were gestures in public relations and therefore not designed to show up student deficiencies.*

Indeed, the role of the college professor in colonial times was more of surrogate parent--responsible for the moral, physical, and spiritual development of the young men under their care, as well as the intellectual. Public relations was only one of several reasons that colleges began a system of grading and marking that will be explored more in the next section.

However, very, very few people went to college--much less than 1% of the population. Compare that to 2009 when the U.S. Bureau of Labor Statistics reported that 70.1% of new high school graduates went on to some form of higher education (Norris, 2014). Most people would be surprised to learn that some of the most prominent of our founding fathers--George Washington, Patrick Henry, Benjamin Franklin--did not go to college. In fact, there were only nine colleges in all of colonial U.S. (Peterson, 1983). With so few people enrolled, and with such individualized attention, there was little or no reason to have grades.

Until the latter half of the 19th century, it was unnecessary for most people to have more than basic reading and writing skills. Up until

[1] "Catechizing" was a form of religious education in which students were instructed in religious doctrine via question and answer.

1790, 90% of the people in the U.S. were engaged in farming, and while the percentage of people who were farmers began to fall after that, it did not fall below 50% until around 1880, and was still around 40% in 1900 (Spielmaker, n.d.). Meanwhile the percentage of people employed in low-skilled manufacturing jobs rose from around 10% of the population in 1840 to around 30% of the population in 1900. It is important to note that the Fair Labor Standards Act, which set federal standards for child labor, was not passed until 1938, and up until that time it was quite common for children to be employed on farms and in factories from age 10 or even younger (Grossman, 1978).

Schooling was just not that important for the majority of people. Grades, which did not appear at the primary and secondary level until some time in the early to mid-1800s, were even less important. It wasn't until the 20th century that the idea of universal, free (i.e. publicly funded), high-quality primary and secondary education began to take hold. Schools focused on the basic skills that would allow people to participate in commerce and the religious lives of their communities. Why, then, have grades gone from a place of relative unimportance less than 100 years ago, to playing a central role in the lives of children and young people today?

Why Have Grades?

In the U.S., grades were first created at colleges and universities and were only later adopted at the primary and secondary level. Before discussing the origin of grading systems in these colleges, it is important to get a sense for the differences between college then and college now.

In the 17th and 18th centuries only a very, very small fraction of people even went to college. Colleges existed primarily for the purpose of training the children of the social elite to take on positions in the upper echelons of the religious and political institutions of the day, for the the propagation of religious faith, and only secondarily for the advancement of human knowledge. There was no standard age of admission, no SAT or ACT, no rigorous application process[2]. There was no standard two or four years that one would spend at a college--graduation occurred when a student's instructors determined that he[3] was ready.

Very little of what was studied would have been considered practical or useful in the lives of common people, most of whom were farmers. There was no such thing as a "major" in early colleges. All of the students followed more or less the same curriculum, which was patterned after the British model established at Oxford and Cambridge. Students learned to read, write, and speak Latin, Greek, Hebrew, and French for the purpose of being able to read and understand classic works in religion and philosophy. A great deal of the focus was on rote memorization, and examinations took the form of public recitations of great works in front of instructors and peers.

Another indication that colleges weren't considered that important in the early history of the U.S. is that there weren't very many of them. There were roughly 3.6 colleges for every million Americans in 1776, compared to roughly 13.8 colleges per million people in 2005--about four times as many. Harvard was founded in 1638 by graduates of Oxford and Cambridge who had traveled to the new world. The second college to be established, the College of William and Mary, was not founded until 1693, followed by Yale in 1701. It was another forty years before Princeton was founded in 1746, and by the time of the signing of the *Declaration of Independence*, there were only nine colleges

[2] Actually, while college admissions were originally based on subjective decisions, over time they became more rigid and based on entrance examinations that were unique to each college.

[3] Note the intentional use of "he." Women were not allowed to attend early colleges, and it wasn't until the early 1800s that colleges were established for women, notably Wheaton College (1834) and Mount Holyoke College (1837).

in existence in the U.S., each of which had no more than a few hundred students each. Now a "small" college is one with fewer than 3,000 students, while some large state schools have upwards of 50,000 students.

Examinations were used almost from the beginning to determine when a student was ready to graduate, although they were not graded as we have come to understand the term. Examinations were oral--students were asked questions or required to recite memorized works in front of a panel of instructors and/or outside examiners, and the result was a simple pass or fail. Written examinations were not commonly used until the early to mid-1800s as paper and printing were prohibitively expensive, and even then, written exams occurred at most once a year. Examinations also came to be used to make admission decisions, and were generally used to determine whether or not colleges were living up to their charters.

But just because there were examinations does not mean that there were grades, nor that people were even happy with the examinations. Josiah Quincy, who was the President of Harvard from 1829 to 1845 wrote:

> *Examinations exist in all colleges, and in Harvard they are more numerous, and, it is believed, more thorough than in any other; and in a certain degree they are very useful, considered as a stimulus to the students; but they are always general, often hurried and brief, and scarcely sufficient to enable the Committee to form a rough and often mistaken estimate of the general state of the class. They neither are, nor pretend to be, a test of the positive or relative attainments of each individual. (in Smallwood, 1935, p26)*

Keep in mind that around the time that President Quincy wrote this, there were only about 20 faculty and 400 students at Harvard (Harvard Annual Report 1830-31). Quincy knew each student, and personally kept track of their performance and behavior in each class. In a real sense, he likely thought of them as his own children.[4] He felt that these exams did not do justice to these young men. It is clear that he saw exams as a necessary evil, and even admitted they had some power to motivate students, yet they remained woefully inadequate to capture and convey the uniqueness of each student. Why is it, though, that exams were considered necessary at all, if they were only able to provide an "often mistaken estimate" of how colleges were performing?

Early colleges suffered from a perception problem. Recall that very, very few people went to college, and most people probably had almost no idea what went on there. Colleges were self-conscious about being seen as lax, and afraid that standards would fall because of academic inbreeding. This led to having examinations conducted and evaluated by external examiners. In the words of another Harvard president:

> *It would be a great gain if all subsequent college examinations could be as impartially conducted by competent examiners brought from without the college and paid for their services. When the teacher examines his class, there is no effective examination of the teacher. If the examinations for the scientific, theological, medical, and dental degrees were conducted by independent boards of examiners, appointed by professional bodies of dignity and influence, the significance of these degrees would be greatly enhanced. (from the Inaugural Address of President Eliot of Harvard, in Smallwood, 1935, p22)*

In other words, the teachers at the colleges were not fully trusted to be fair and impartial judges of their own students' performance. This mistrust was probably not unwarranted given the close personal nature of the teacher-student relationship--it would be like expecting a parent to be a fair and impartial judge of their own child's performance. As such, colleges labored to bring in outside examiners to conduct public examinations of their students.

[4] As a "father" to the students, however, Quincy was likely perceived as a harsh disciplinarian. The Harvard website cites his "rough touch" as being responsible for destructive rebellion amongst the students.

This practice was not without complications, however. There just weren't that many people even qualified to serve as examiners, and these people had to be brought in from long distances (on horseback) at great expense. Consider that the number of people in the colonies fluent in Latin, Greek, Hebrew, and French was low to begin with. Further consider that these people were probably very busy in their clerical, governmental, or business endeavors. It is easy to understand that getting these people to take time off and travel to the colleges one or more times a year was difficult.

Another perception problem of colleges had to do with the social class of the students. It was customary for colleges to award honors to students at the commencement ceremonies held when students graduated. These ceremonies consisted of various recitations and oratory in the classical languages. Some of these oratories were more prestigious than others, and as such, selection of which students were to deliver them was a significant undertaking. Ostensibly, the selections were based on merit and the actual academic performance of the students, but in practice, honors were often accorded to students because they came from families of high social status, and were chosen strategically so as not to insult those families. This, more than anything, may have been the original impetus for grading--coming up with a clearly and publicly fair and impartial strategy for ranking students that did not (primarily) depend upon social class.

Since examinations were used to determine whether or not students were qualified for graduation, and since they were also used to award honors at commencement, it is reasonable to believe that there always existed some form of ranking or grading system. However, in her research, Smallwood found no mention of a formalized system until it appeared in a footnote of the diary of Ezra Stiles, President of Yale, in 1785--nearly 150 years after the founding of the first college in the colonies. It appears that the first grading systems consisted of adjectives used to divide students into four groups (from highest to lowest): *optimi,* second *optimi, inferiores,* and *pejores.* The first mention of a numeric scale also appears at Yale in 1813 when examinations were scored on a scale of 1 to 4 in each subject, and the average of these scores were recorded. Although there is no direct evidence of this, it is possible that the numeric system was seen as superior to the category system since decimals could be used to make a more nuanced ranking of the students. However, there was no standardization of grading systems and different colleges elected to use both the numeric and categorical systems throughout the 1800s.

There was a great deal of experimentation and variation in the numeric scales that were developed in the 1800s. The first record of a numeric grading system at Harvard appears in 1830 and describes a 20-point scale. Interestingly, students appear to be roughly evenly distributed amongst the 20 levels. By 1837 there was evidence at Harvard of the first 100-point scale (decimals were not used), which gave teachers the ability to provide finer distinctions between students, but in the written reports, students were still grouped into those that received 75-100, 51-75, 26-50, and under 26. In 1842, Yale recorded that it used a scale from 0 to 4, with 2 being the average, and that students not receiving at least a 2 in all subjects would not be allowed to continue with their class until they were able to pass tests in all subjects with a 2 or better. The first evidence of a weighted average appears at Yale in 1846, when it was decided that for freshmen, translation would only count as one-tenth of the average for all classes.

From the outset it appears that college faculty were unhappy with the various numeric formulations of grades. This dissatisfaction led to frequent changes. At times some colleges abandoned numeric systems altogether, returning to using adjectives to describe student accomplishment. Yale seems to have remained consistently with a 4-point scale (although at one point, apparently dabbled with a 9-point scale).

Numeric scales including 4, 8, 24, and 100 points were all tried, with frequent revision and adjustment. Throughout this time, it appears that there was little doubt that the system of examinations was satisfactory for determining students' worthiness for promotion and graduation, but the translation of those examinations into numeric values was a constant bone of contention. The dissatisfaction of faculty with numeric grading systems is perhaps best captured in this 1890 report related by Smallwood:

> *The marking system ... has been overhauled and reduced to the* ***least obnoxious condition****. Formerly, the maximum rank for any recitation was eight; the students were ranked for the year on a scale of 100, but, though the scale was the same,* ***no two instructors agreed in their use of it****. ... some frankly admitted that it was impossible to get within five or ten percent of absolute exactness; others were so delicately constituted that they could distinguish between fractions of one percent. One instructor was popularly supposed to possess a marking "machine"; another sometimes assigned marks less than zero... (my emphasis, in Smallwood, 1935, p53)*

Note the sarcastic tone of the author of this report. The point to be made here is that grading systems were always "obnoxious," and the inconsistency in the application of grading scales described here has always existed. Little, if anything, has changed since then.

To make this point more concrete, please excuse a small interruption in the history that will illustrate that things are no better now than they were then. The first study to call attention to is one done by Paul Diederich in 1966. Diederich was the Director of Research in English at Educational Testing Service (ETS)--the company that produces the SAT, GRE, AP and other standardized tests. In this study, 300 college essays were each graded by 60 evaluators from a variety of disciplines and sorted into nine categories from "worst" to "best." In the words of the investigators:

> *The result was nearly chaos. Of the 300 papers, 101 received all nine grades, 111 received eight, 70 received seven, and no paper received less than five. (p442)*

In other words, 94% of the papers were put into seven or more different quality categories--there was very little agreement about what constitutes a "good" essay. While Diederich and his colleagues then went on to devise ways to make evaluations of student writing more reliable, there is little evidence that these practices were ever adopted on a wide scale.

Several more studies will demonstrate that grades are as unreliable and arbitrary now as they have ever been. In 1976, Goldman and Widawski found systematic evidence that students with "lower ability" were being graded more leniently than those with "higher ability." Despite their call for a more equitable system, little seems to have changed by 1988 when Elliot and Strenta again demonstrated a high degree of variability amongst grading standards between college departments. A study of 8,454 high school students performed by Willingham, Pollack, and Lewis (again from ETS) in 2002 found that "grading variation was a major source of discrepancy between grades and test scores" (p1). In 2003, Smith explained the unreliability of classroom grading in terms of sufficiency (actually *in*sufficiency) of information about student performance. Finally, a 2012 study by Duckworth, Quinn, and Tsukayama explained the discrepancies between test scores and GPA, and the relative inability for one to predict the other, in terms of IQ, which better predicts standardized test scores, and self-control, which better predicts classroom grades. The bottom line is that from 1890 to 2012 grades have become no less obnoxious--we're still struggling to figure out the "best" way to reduce the dynamic and multifaceted performance of students to a small set of letters and numbers.

So why should we have grades at all? While the history shows that the original motivation to grade students was to improve the general public's opinion and trust of what went on at U.S. colleges, and also to have a more equitable ranking system that didn't rely solely upon the social class of the students being evaluated, by the late 1800's the question of *should* we grade was no longer being asked. Already the existence of grades was a foregone conclusion and debates had shifted to the question of *how* to grade. Arguably, given the rise in the number and prominence of U.S. universities, grades fulfilled their purpose. It is probably impossible to know to what degree grades can be credited for the rise of education in this country. It may have been inevitable.

Regardless of the reason why we no longer ask whether or not it is appropriate or useful to grade our students, grades are clearly well entrenched into our modern society. The first mention of the use of the letter-grade system that we use now didn't occur until 1883 at Harvard where a student was noted to have received a "B" in a course. The first formal description of the full scale, from A to F, appeared at Mount Holyoke in 1898. It is unclear how the practice of grading spread from these colleges to other educational settings, but it seems plausible that a great number of college graduates entered the teaching profession and brought grades with them. By 1913, the practice was so widespread, it prompted this comment from the editor of Finkelstein's analysis of grading systems:

> *When we consider* ***the practically universal use in all educational systems 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 system. (p1, emphasis added)*

This "blind faith" is still very much in evidence today, over 100 years later.

Not everyone has always taken the mechanics, validity, reliability, or even appropriateness of grades on faith, however. The next section will summarize the major debates and discussions that have occurred about grades and grading systems over the last 200 years in the U.S.

Historical Debates About Grading

While the vast majority of teachers, students, and parents rarely talk about it, there has always been a small subset of scholars who have debated both *if* and *how* grading should be done. John Laska and Tina Juarez compiled a collection of essays written between 1840 and 1988 documenting these ongoing debates that will be summarized here. The goals of this section are to show first, that fierce debates about grades and grading have always existed, and second, what the general themes of these debates have been. Many of these themes will be addressed in further depth later in this book.

Comparative versus Mastery Grading

There are two major schools of thought about how to grade students that can be described as *comparative* and *mastery*. Both approaches encode deeply held cultural beliefs about the fundamental purpose and nature of education and learning. The debate about whether or not students should be graded in comparison to other students, or graded based on whether or not each individual has personally mastered a topic reflects our society's underlying argument about whether life on this planet is fundamentally competitive or cooperative.

Comparative grading reflects the belief that life is fundamentally competitive, and therefore, the most appropriate and caring way to educate our children is to prepare them to compete, to weather the vicissitudes of life, and to understand how they measure up against other people. Students are graded either against an objective standard, i.e. how much of the material in a particular unit was learned, or against their peers, i.e. where do the students rank in comparison to the other students in their class--frequently referred to as grading on the curve. Implicit in this view is the idea that not everyone can or should be good at everything.

Comparative graders recognize that being evaluated under such a system will be frequently unpleasant for the students. Alfie Kohn (2014) sums up this philosophy as follows:

> *Let's take a step back and ponder the phrase "subject kids to unpleasant experiences" in more general terms. We often hear an argument that runs as follows: If adults allow (or perhaps even require) children to play a game in which the point is to slam a ball at someone before he or she can get out of the way, or hand out zeroes to underscore a child's academic failure, or demand that most young athletes go home without even a consolation prize (in order to impress upon them the difference between them and the winners), well, sure, they might feel lousy--about themselves, about the people around them, and about life itself--but* ***that's the point****. It's a dog-eat-dog world out there, and the sooner they learn that, the better they'll be at dealing with it.*
>
> *The corollary claim is that if we intervene to relieve the pain, if we celebrate all the players for their effort, then we'd just be coddling them and giving them false hopes. A little thanks-for-playing trophy might allow them to forget, or avoid truly absorbing, the fact that they* ***lost****. Then they might overestimate their own competence and fall apart later in life when they learn the truth about themselves (or about the harshness of life). We do them no favors by sheltering them from the fact of their own inadequacy or from the cruelty that awaits them when they're older. (p86)*

Kohn sums up this attitude with the acronym BGUTI (rhymes with "duty")--Better Get Used To It. Proponents of this style of grading highlight its realism and its pragmatism.

Mastery grading, on the other hand, reflects the belief that life is fundamentally cooperative, and therefore, the most appropriate and caring way to educate our children is to recognize their uniqueness, acknowledge that people learn at different paces and in different ways, and that given enough time and support, almost all people are capable of mastering any given subject. Everyone has something to contribute to society and to the world, and nobody benefits by forcing students to study arbitrary amounts of information according to arbitrary timeframes. If a given topic is worth spending time on, why on earth would we not give every student ample time to master it? Learning is not a race or a competition and the goal is not to "win."

One of the obvious problems with mastery grading is that since each student is given as much time as he or she needs to master each topic, there's no way to predict how long it is going to take for all of the students in a class to complete one subject before moving on to the next. It becomes more difficult to manage classrooms if all of the students are in a different place. It becomes more difficult to assess the students as a group if they are not all following the same schedule. Proponents of mastery learning would argue that this is precisely the point: if each student truly is a unique individual, why would we want to assess them as a single unit? Why would we want to promote a system that makes it easier for a teacher to administer a classroom at the expense of students' learning?

The next several sections will describe the most important debates that have raged over the last 200 years with regard to grading.

How Many Categories?

The first theme of debates that have raged for centuries has to do with how many categories to put into the grading scheme and how to represent them. For proponents of mastery grading, this question is simple--there are only two categories: done and not done. However, for proponents of comparative grading there seems to be no end to the discussions of the various ways to slice and dice student performance.

Chapter 3
What do Grades Mean?

THIS CHAPTER IS about interpretation--it is about how people make sense of the world. In particular, this chapter is about how people interpret the meaning of grades. For example, when your son or daughter brings home a physics test and has received an 83%, what does this mean--83% of what? When you take a class in music appreciation and get an A-, what does the minus mean--that you are not quite perfectly able to appreciate music? When an employer or college admissions office reviews a transcript and sees that a student got a C in AP American Literature, what does that mean--what kind of impact should that grade have on the decision to employ a person or accept them into a college? How should you, as a parent, react to the various grades that your child brings home from school? How should you, as a student, use the information you get from grades to adjust your study habits, your self-image, your plans for the future? How should you, as an employer interpret the grades that come on the transcripts from a wide variety of schools? Grades have real, significant impacts on the lives of people. We are constantly using grades to make decisions about how we should act, and how we should treat people. So it makes sense to ask the question, what do grades really mean?

Let's explore the possible meanings a grade might have through an example. Imagine that John has written a five-page essay entitled "*The Impact of Social Media on the Political Process in US Presidential Elections*," and that when he gets the paper back, his teacher has given him a grade of B-. How should we interpret that? There are many ways. Some people believe that a C should be awarded for work that is of average quality and basically satisfactory--in that case a B- is slightly above average. More commonly these days, anything less than an A is seen to be as significantly flawed work. It is possible that the teacher knows that John didn't start on the essay until the night before it was due and is penalizing him for procrastination. On the other hand, it is possible that the teacher knows that John struggles as a writer and has really, really worked hard on this essay, submitting multiple drafts for feedback and suggestions for revision. In both of these cases, the B- is partially the result of an ongoing relationship between John and his teacher. Perhaps the teacher is a "hard grader" and none of the students in the class received anything higher than a B, in which case, John's B- represents excellent work. Perhaps the teacher delegated the task of grading to a teaching assistant, who doesn't know John at all, and marked his essay based on a rubric provided by the teacher. Perhaps the essay was extremely well-written and defended, but took a position that the teacher doesn't like or support. Perhaps John's essay was just one of 28 essays that the teacher had to grade over a weekend when the teacher would rather have gone hiking, and so it didn't get the teacher's full attention (28 essays at 5 pages each is 140 pages--sound like an exciting way to spend a weekend?). What does a B- say about John and the quality of his essay, about the quality of his ability to gather and analyze information, about his level of effort, about his ability to organize and express ideas effectively, about whether or not John has a future with technology or political science? Is a B- good or bad? Does B- mean the same thing to John and the teacher and his parents? How should John use this information to adapt his future efforts? How should the teacher use this information to interact with John in the future? How should John's parents react? I hope you'll agree with me that the meaning of John's B- is not very clear.

There are many more questions we could ask about John's B-. Did the teacher provide guidelines beforehand as to how the grades would be determined? Did the teacher use a rubric, or provide constructive comments on the paper when it was returned? What does an "objective" evaluation of such an essay look like anyway? Assuming both John and his teacher understand the true meaning of the B-, how will other people outside of their immediate relationship understand it? Who else will read the essay besides John's teacher--is this essay destined to help other people learn more about technology and politics, or will it just end up in the garbage after John gets it back? Did John enjoy the task of writing the essay? Did he learn anything about himself? Did the experience make him more thirsty to understand the way the world works, or did he find it to be a boring waste of time? What else was going on in John's life while he was writing this essay? Perhaps he was studying hard for a physics test at the same time and couldn't devote full effort to his essay. Perhaps he was training for a marathon, or just got dumped by his girlfriend and wasn't up to doing his best. Does the teacher know about these other things going on in John's life and take account of them in the final grade? In light of these questions, what does John's B- really mean? It is very tempting to say that the B- can mean lots of different things, or even to go as far as to say that it is meaningless. In any case, figuring out the answer to this question can be extremely difficult.

Let me be clear: *grades have meaning*. They do mean something, in fact, many things, and in practice it is almost impossible to know the precise meaning without understanding the full context of the assignment. In scientific language, we say that grades have "no reliably meaningful interpretation." This is a huge problem. Since grades are a part of the every day life of our young people, and since they also set up the patterns and expectations of how people will think, learn, and act for the rest of their lives, it is a problem that really needs to be solved. If grades are to be useful, we must know how to interpret them.

This chapter will first look at the question of whether or not grades are a form of measurement, and if so, what they (attempt to) measure. Next this chapter will dive into the world of psychometrics, a field of study that was created to make the measurement of human ability more scientific. After that we will study the average, perhaps the most important statistic used in grading. You'll see how examination of psychometrics and the average only make it harder to interpret grades. Even still, we will end the chapter by asking the question of whether any of this really matters--so grades aren't perfect, but aren't they good enough?

I really do hate this chapter. I hate it for a couple of reasons. First of all, this chapter is all about math and statistics, but that's not why I hate it--math and stats are cool. I hate it because while it contains some of the strongest arguments against the use of grades, it is also the least accessible to most people. I'm including it in the book because the egghead academic in me knows that I have to, especially if I'm ever to convince other egghead academics that we're doing it wrong. However, I admit that I've never really found a way to convey the ideas here in a way that is clear, convincing, and won't instantly put people to sleep. Many times I have had the experience of watching people's eyes glaze over as I've tried to explain the things I'm about to explain. My inability to make the gobsmacking ludicrousness of the stuff in this chapter plain to everyone has been very frustrating.

So, dear reader, part of me wants to tell you that it's okay to skip this chapter. If math is "not your thing," or if you never really "got" statistics, I'd love to be able to tell you that this chapter is not really that important, and that you'll get just as much out of the book without getting into it. But I don't feel that way. I really do want you to read it, and re-read it, or find someone to explain it to you, until you get it because the logical fallacies that have led us to continue to use grades are really quite staggering.

I've probably taken at least seven formal courses in statistics and experimental design at both the graduate and undergraduate level. Statistics don't come easily to me, and I've never had the same sort of intuitive understanding of them that many of my academic colleagues seem to have. However, given that even the people who teach statistics continue to misuse them with such wanton abandon, and that nobody seems to have realized the bitter irony of this, leads me to believe that maybe I'm not alone in falling short of having a full grasp of the subject matter.

There are a lot of people out there who "should have known better" and yet have failed to make the realizations that I've made here. It is my sincere hope that reading this chapter will provoke in you the same kind of reaction it provoked in me: your jaw will hit the floor, you'll slap your forehead, and you'll say something like "Oh my God! I can't believe we've been doing it this way for so long!!!" On the other hand, I guess we could look at this chapter as a cautionary tale, and a testament to the power of people's faith in numbers and calculations. We all share this faith on some level, and there's good reason to--many, if not most, of the technological and scientific advances we've made over the past several centuries have come as a result of our adept wielding of the tool that statistical analysis is. My arguments here are emphatically NOT designed to shake people's faith in the power and appropriateness of using statistics, but they are meant to challenge you to really understand what kind of insights and information they can and cannot provide.

So I ask your forgiveness in advance for what is likely to be a difficult chapter for many people to get through. I've done my best to make the concepts as clear and as easy to understand as possible, but the fact that this is not easy to do should also give the reader some insight into perhaps why we all could have dropped the ball on this for so long. Please persevere and I think you'll find that this chapter may be one of the most rewarding of the book.

Do grades measure anything? This is an important question to ask. While I hope that the answer to this question is obviously "yes," the easiest way to derail the rest of the arguments in this chapter is to say that grades are not really measurement. If grades were not measurement, then it would not be necessary for grades to live up to the standards of other forms of measurement. It would not be necessary for them to be reliable or valid, since they would really just be the opinions of the teacher assigning the grade. It would not be necessary for us to have a standard interpretation of what the grades mean. Teachers would be free to assign any grades that they wanted without having to worry about the psychometric or statistical properties of the numbers and letters that they assign. Employers and colleges would be able to interpret the grades in any way that suited them to make the decisions they want to make. Concepts like fairness and objectivity would become irrelevant. If grades don't measure anything, then it would really be just up to each student to game the system as well as possible to achieve the best possible outcome. In the absence of any rules, like those imposed by psychometrics or statistics, it would be justifiable to use any means necessary gain access to the benefits that good grades bestow--good jobs, places at prestigious universities, scholarships. It is arguable that many people already act as if this is the case. Clearly though, the vast majority of us believe that grades do, in fact, measure something, and should be awarded based on principles like objectivity, fairness, reliability, and validity. But what do grades measure?

The most obvious answer to this question is that grades measure learning. Grades provide evidence that a person has learned to solve differential equations, or write a sonnet, or identify major events in European history, or explain the importance of photosynthesis. Grades, theoretically, provide a measurement of how well a learner was able to do all of these things. Taken more abstractly, grades measure how good a person is at learning in general. Grades allow us to say that this person is good at science and math, or that person is good at art and music. Grades measure our past performance and serve as an indicator of our future potential. But is that all they measure?

What about the "A for effort?" Grades also in part measure how hard a person is trying to learn, or how persistent they are at carrying through with learning tasks. Extra credit is awarded to students who go above and beyond and do additional work. Partial credit is awarded for showing the steps you followed to arrive at a solution, even if the final solution is incorrect. Teachers certainly look with favor on the students who seem to be trying the hardest. Demonstrating that you are a hard and persistent worker is certainly a thing of value in our society. In some part, grades are an indication of your work ethic, and we certainly use grades to try to convince people that hiring us is a good idea. But grades measure even more than that.

Grades also measure how much our teachers like us and believe in us. Although we don't like to admit it, we know that subjectivity and bias creep into the grading process. Imagine that there is a student who has a grade that is on the borderline between an A and a B. The teacher could decide to go either way and make a reasonable justification for it. If the student is always attentive, hard-working, cheerful and polite, the decision is likely to go in the A direction. However, if the student is rude, lazy, and unpleasant to be around in class, the teacher is going to be much less likely to give them the "benefit of the doubt." This kind of bias is just human nature.

Decades of research on prejudice and discrimination have shown us that people frequently are unaware that they discriminate based on race, gender, or just whether or not they like a person. Stereotypes are powerful and creep into the learning environment in very subtle ways. One interesting study showed just how powerful stereotypes can be by playing off two opposing stereotypes. On the one hand, women are stereotyped to be less capable at mathematics, and women that are good at math are seen as somehow less feminine. On the other hand, asians are stereotyped to be better at math than other race groups. In the study, a group of asian women was randomly assigned to two groups, both of which took exactly the same math test. Statistically, there should have been no difference whatsoever between the final scores achieved by the two groups. However, one group was reminded of their being female prior to the test, and the other group was reminded of their being asian. The group that was reminded of being asian scored significantly higher on the test. The women themselves were unaware that they possessed these ingrained beliefs about themselves. Teachers can unwittingly trigger these stereotypes both when giving assignments, and also when grading them (Shih et al., 1999).

So in practice, grades measure some combination of performance, effort, and how our teachers feel about us. But what *should* they really measure? Most people would agree that the teacher's personal feelings or prejudices should have no place in the grade, but what about effort? As it turns out, the American Educational Research Association (AERA), National Council of Mathematics Educators (NCME), and the American Psychological Association (APA)--three organizations that represent the most current and rigorous standards in educational assessment--got together to write a manual for how to perform assessments of academic ability. They clearly state that in order for a measurement to be valid, effort should absolutely not be a part of the calculation (Joint American Educational Research Association, 1999). In short, a grade should be based solely, and entirely, on an objective measurement of a student's actual performance on an assignment or test. In other words, at the end of the day it doesn't matter how hard you try, either you can do a task, or you can't.

The reasoning behind the AERA/NCME/APA standard is pretty easy to understand. If a grade is made up of some combination of performance, effort, and bias, how are we to know how much of each one is used to determine the grade? How can we reliably interpret grades if we don't know if the grade was awarded because the teacher liked the student, the student worked hard, or the student actually performed well? The goal of the standard is to resolve this quandary. By assigning grades based only upon actual performance, we will, at the very least, have a valid and reliable understanding of what individuals can do.

That being said, anyone who has been involved in our educational system recently will know that this standard is rarely, if ever, applied to grades. Despite that, for the rest of this chapter we will pretend like teachers are unwaveringly fair, and unwaveringly objective in how they assign grades. We will assume that every grade that is given is based solely on the actual performance of the learner, regardless of race, gender, level of effort, or any other distinction. We will show that even in this ideal case, it is still extraordinarily difficult, if not impossible, for grades to be meaningful measures of learning. The path to get to this realization is through psychometrics.

Psychometrics is a term that comes from the prefix "psycho-" which connotes things like knowledge, abilities, attitudes, feelings, and other psychological properties, and the suffix "-metric" which means "to measure." In other words, psychometrics is the study and development of ways to measure things like knowledge, ability, attitude, and feeling. It is the foundation of the field of educational measurement.

Assessment and measurement are vital parts of the learning process. The results of assessment provide crucial feedback to both teachers and learners and can help guide students to correct their misconceptions and guide teachers in finding appropriate ways to help students achieve their goals. Dieters measure their progress with a bathroom scale. Runners measure their progress with a stopwatch. Businesspeople measure their progress by keeping careful accounts and watching the bottom line. Learners measure their progress through assessment and measurement. An educational system that didn't engage in frequent assessment of learning would have a difficult time maintaining its effectiveness.

Unlike weight or speed, however, there is no physical device that can measure learning. The brain is not measurably heavier after an hour of lecture. There is no thermometer that you can insert into a student to see how many degrees of ability have been gained. In scientific terms, we say that learning is not directly observable. Even if we had the time, energy, and resources to put all of our students through a brain scan on a regular basis, we still don't know anywhere near enough about the brain to determine what, if anything, a person has learned. This is why the field of psychometrics was developed.

Instead of bathroom scales, stopwatches, and thermometers, psychometricians use tests to measure learning. Although this is not a direct observation in the traditional sense, the thinking is relatively straightforward. If I want to determine if a student has learned how to multiply, I give the student a multiplication problem to do. If the student gets the problem correct, then most likely that skill has been learned. The process seems simple enough, but it rapidly gets more complicated than that.

Say for example, a student gets the first multiplication problem correct, but then gets the second one incorrect. Perhaps the student got lucky and guessed correctly on the first one, or maybe a careless mistake was made on the second one. No problem, just give the student a bunch of problems to solve, say ten or twenty, and look for a pattern. If most of the problems were answered correctly, then it is highly likely that the skill has been learned. If most of the problems were answered incorrectly, then it's likely that the student needs more work on this skill. Or is it?

When we analyze the situation further, more complications arise. We quickly realize that the student is getting problems correct because he or she memorized the correct answers, as most of us were made to do at some point with our multiplication tables. We discover that our student is merely regurgitating the correct answers from memory but still doesn't clearly understand the basic mechanism of multiplication. We can tell this because the student does well on single or double-digit multiplication problems, but ability breaks down for multiplication of numbers with three, four, or more digits. Now we are unsure whether we are measuring our student's memory or cognitive ability.

Another complication may have nothing to do with cognitive ability at all. We may find that the student does well on our multiplication test on some days and not so well on others. When we dig deeper and talk to the student we find that on some days the student oversleeps and misses breakfast, or that the student is facing a difficult situation at home. Now we begin to suspect that the student's performance on our tests may not be completely related to whether or not multiplication has been mastered or not. Or we may find that the student does well on oral tests, but not written ones, again an indication that the student may, in fact, understand multiplication and still not do well on our tests.

This is one of the core problems of psychometrics: we can never really truly know what people understand or can do--we can only make educated guesses based on how they perform on tests. This begins to get even more difficult as the level of difficulty of the abilities being tested increases. What happens when we introduce multiplication of fractions, or quadratic equations, or integral calculus? And while mathematics seems straightforward enough, how do we begin to assess more abstract abilities, such as the ability to infer the meaning of a new vocabulary word based on its context in a sentence or passage? We begin to understand that it is extremely important what questions we ask, how we ask them, and how we interpret the results. This realization led to the development of something called classical test theory (CTT).

What is a "good" question?

When psychometricians discovered that the quality of the questions being asked makes a difference, they naturally began to search for ways to measure the quality of questions. Classical test theory answered that question by coming up with the concept of discrimination. A good question can discriminate between students who understand a concept and those who don't. This concept is very useful because if a teacher can discover which students don't understand a concept yet, those students can be given additional practice or instruction until they do.

CTT developed a formula for determining the discriminatory power of a test question as follows:

$$D = \frac{U - L}{U}$$

In this formula, U represents the number of students in the upper quartile of students who got a particular question correct (the quarter of the students who got the top scores on the test), L represents the number of students in the lower quartile that answered correctly, and D represents the discriminatory power of a particular question. Assume that twenty students take a test. U is calculated based on the five students who got the highest scores, and L is calculated based on the five students who got the lowest scores. So if all of the "top" students (U) got a question correct, and all of the "bottom" students (L) got a question incorrect, the formula would look like:

$$D = \frac{5-0}{5} = 1$$

The discriminatory power of the question (D) would equal one because that question was perfectly able to discriminate between the "high ability" and "low ability" students. If all of the students in both the U group and L group got a question correct the formula would look like:

$$D = \frac{5-5}{5} = 0$$

The discriminatory power of this question would be zero since regardless of ability level, all of the students got the right answer. While it may be reassuring to know that everyone seems to understand the concept, such questions don't provide much guidance in future instruction and probably should be thrown out because they are too easy.

It is also possible that L could be bigger than U. For example, if all of the "bottom" students got a question correct and all of the "top" students missed it, then the formula would look like:

$$D = \frac{0-5}{5} = -1$$

In this case, a classical test theorist would probably conclude that the question was a "trick" question or otherwise misleading since the D score was negative. This question would likely be thrown out and not used again.

The CTT model of discrimination has some important strengths. First of all, it is very easy to calculate D for any set of questions. Second, it is relatively easy for anyone giving a test to understand what the discrimination value means for any particular question. Third, interpretation of the D score is relatively straightforward. In general, it is unlikely that the scores will ever be zero or one, but questions having a D score closer to one are better for figuring out what skills or abilities are causing students the most trouble. Unfortunately, the weaknesses of this model surfaced fairly quickly.

The first problem is that the value of D is dependent on the specific group of students that takes the test. Imagine giving a test to a group of students who are all at about the same level of mastery. In truth, there is no real difference between the U group and the L group--they all understand the concepts at about the same level. Because of this, for any given question, the number of students in both U and L who got a correct answer is going to be about the same, and therefore all of the questions will have relatively low D scores. However, if the same exact questions are given to a different set of students who have a lot more variability in their ability levels, the same questions will on average have higher D scores. As such, D scores depend on the group of students being tested.

The second problem is that the values of D depend on what other questions are asked on the same test. Since the mix of questions on a test will determine which students are in the U group and the L group, the same question can have a higher or lower D score depending on whether or not the rest of the questions on the test are relatively "harder" or "easier," again with respect to the particular set of students taking the test.

The problem of fluctuating D scores in many ways defeated the purpose of CTT. The goal of CTT was to determine which questions were good questions, so that those questions could be re-used on future tests with different students. However, if D changes depending on the set of students or the set of questions that gets chosen for the test, then it means that we can never reliably predict if a particular question is going to be a good discriminator ahead of time.

On the other hand, as a result of their efforts with CTT, psychometricians clarified what they were looking for. Their goal now became to find a way to measure the quality of questions, defined as the question's ability to discriminate between students who have and have not mastered a particular skill or ability, in a way that did not depend on the particular group of students, or group of questions selected for the test. In other words, they needed a way to predict how a question would perform before it was administered. This led to the development of item response theory (IRT).

Item Response Theory (IRT)

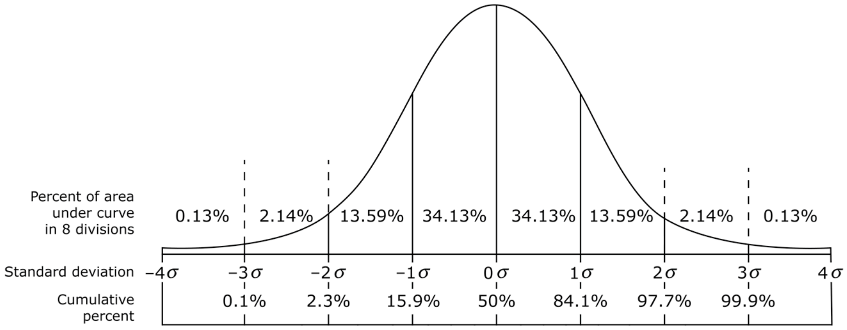
Item response theory is the current standard for determining question quality in the professional standardized testing industry. It is employed for determining the quality of questions on "high stakes" tests such as the SAT, GRE, and many or most of the other standardized tests that are administered regularly to school children of all ages. From a mathematical perspective, IRT models are relatively complex, but I will try to explain one of the more common models in a way that is easy for most anyone to understand.

The model I will explain is known as the 3-parameter model and is named so for its three parameters *a*, *b*, and *c*. The model is an equation that looks like this:

$$p_i(\theta) = c_i + \frac{1 - c_i}{1 + e^{-a_i(\theta - b_i)}}$$

Before I try to explain how this equation works, let me explain some of the new concepts involved in psychometric theory. The first concept is θ (theta), which is the theoretical, unobservable ability level of a person. Remember how we said you can't directly observe the knowledge and ability that a person has? Just because we can't see it directly, doesn't mean it's not there. We know that people have a certain ability level, say for math or language, and that is what we're trying to figure out. The Greek letter theta, θ, is the symbol that psychometricians use to represent ability level. Although we can make a pretty good guess at a person's ability level based on their answers to test questions, we can never ever truly know it's exact value.

For most types of abilities that they study, psychometricians assume that ability level is normally distributed throughout society. You might have heard of normally distributed things referred to as representing a bell curve.



The image above depicts the standard normal distribution of some ability within the population of all people. To make this concept a little more concrete, let's use an example. Pretend that I got 10,000 high school seniors to come to my basketball court and try to make fifty free throws. Most people would probably make a few of the free throws, a few people might make all fifty, and a few people wouldn't make any. If I counted up all of the people according to how many free throws they made and stacked them in categories, the resulting graph would probably look something like the bell curve above, with a hump in the middle (the average) that gradually tapered off at each end. This is not too difficult to understand. The majority of high school seniors have some ability to make free throws. Some of them are extremely good, and some of them are extremely bad. Theoretically, if you tested everyone in the society, you could get a bell curve to describe pretty much any ability from math, to cooking, to carpentry. The hump might be tall and narrow if there is not a lot of difference between the highest and lowest ability levels, or it might be flat and spread out if there is a wide variation of ability levels.

Statisticians have come up with a handy way to describe how far away from the average (i.e. the top of the hump) a person's ability level is. Called the "standard deviation," it is denoted by the Greek letter sigma σ. Those are the vertical lines in the bell curve above. About two-thirds of the population is within one standard deviation of the average ($\pm1\sigma$). If you extend the range to two standard deviations ($\pm2\sigma$), it accounts for about 95% of the population, and if you extend the range to three standard deviations ($\pm3\sigma$), it accounts for 99.8% of the population. In psychometrics, ability levels, denoted by Greek letter theta (θ), are typically described in terms of standard deviations from the average. So if your θ level was +2, you would have a higher ability than 97.7% of the population. The goal of each question is to be able to estimate your θ level based on whether or not you got the right answer.

The second major concept to understand is probability, or the likelihood that something is true. For each test taker there is a probability that they will get a correct answer on each question. As we discussed, just because they get the answer right, doesn't necessarily mean they understand the concept or have the ability, and likewise, just because they missed the question doesn't mean they don't have the ability.

However, a correct answer means there is a higher probability that the person has the skill, and an incorrect answer means there is a lower probability that they have the skill. To flip it around, it stands to reason that if a person has a high ability level (θ) then the probability level of answering any given question correctly goes up. Therefore, the left hand side of the 3-parameter model, P(θ), can be read "the probability of a correct answer if the person has ability level θ."

With these two concepts we can now introduce the item characteristic curve (ICC). (In psychometrics, a question is usually referred to as an "item.") Here is a sample ICC:

[Insert ICC graphic]

The ICC represents a single question (or item). On the horizontal axis in this graph is the test taker's θ level, the value we're trying to figure out. You'll notice that this axis has values from –3 to 3 representing the standard deviations away from the average ability level for this skill. On the vertical axis is P(θ), or the probability that the test taker will answer this question correctly. The values on the vertical axis go from 0 to 1 indicating from a 0% chance to a 100% chance. The line on the graph slopes up from the left to the right, which indicates that as ability level (θ) rises, so does the probability that the person will answer correctly. You'll notice that the line never reaches 100% (the top of the graph) because regardless of how high a person's ability level, there's always a chance that they might make a careless mistake. You'll also notice that the line intersects the vertical axis above zero. That's because regardless of how low a person's ability level, there's always a small chance that they could guess the right answer.

Now we can begin to understand the three parameters in the 3-parameter model. In this case, parameter is just a number that helps describe the relationship between a person's true ability and the likelihood or probability that they will answer a particular question correctly. The three parameters are called *a*, *b*, and *c*. Parameter *a* is the discrimination parameter. It determines the slope of the line in the graph. If the slope is steep, then it means there is a clear cutoff point between the ability levels at which people are likely to get a correct answer. The closer to vertical the slope of the ICC, the better able the question is to discriminate between ability levels. Below is an ICC with two different lines drawn on it. The steeper slope represents a good discriminator. The more gradual slope means that it's harder to know exactly where a person's ability level is based on their answer to the question.

The second parameter, *b*, is the difficulty parameter and it determines how far to the left or right the graph is shifted. If the graph is shifted to the left it means that the question is easier since test takers with lower ability levels have a higher probability of a correct answer. If the line is shifted to the right it means the question is harder since test takers have to have a higher ability level to have a good chance of getting the question correct. The graph below shows several curves that are shifted to the left, right, and middle of the graph. The only difference between these curves is their *b* parameter.

The third parameter, *c*, is the guessing parameter and it describes the place at which the ICC intersects the vertical axis. This intersect represents the probability that a test taker can guess the right answer to the question even if they have a very, very low ability level. The lower the *c* level, the lower the probability that a question can be guessed correctly. This parameter is particularly important for multiple choice questions, since there is always some non-zero chance that anyone can guess the right answer. The graph below illustrates several curves with different *c* values.

The beautiful thing about the 3-parameter model is that once we know the *a*, *b*, and *c* parameters for a question, we know how it will perform on any test with any set of questions, regardless of who the people are taking the test. It overcomes all of the shortcomings of classical test theory. It makes it possible for two people to take two versions of the SAT years apart with totally different questions, but still allow us to make meaningful comparisons between the scores. Or it makes it possible for the same person to take the SAT multiple times, with different questions and still be able to compare the scores. With IRT we can have confidence that even though we may use different questions to assess ability, we're still testing for the same abilities. If you're an educational assessment geek like I am, IRT and the 3-parameter model are nothing short of astounding.

Of course, nothing comes for free. The 3-parameter model has some major drawbacks. The major one is that for any given question, there is no way to determine *a*, *b*, and *c* ahead of time. The three parameters have to be estimated using a relatively advanced method in statistics called maximum likelihood estimation (MLE). In order to do MLE you have to have 500 to 1000 people answer your question first. Only then can you estimate the parameters and draw an ICC. If you've taken the SAT any time recently you may have noticed a section of the test marked "experimental." Those questions are not used to determine your actual score because they are brand new questions for which no ICC has been created yet. Since thousands of people take the SAT every year, ETS, the company that produces the SAT, has many, many people that they can use to try out their new questions before they actually get used for scoring.

I had the good fortune to work for ETS as an intern one summer when I was a graduate student, and I saw the process that ETS uses to generate questions from the inside. Because of the potential for cheating, ETS has to constantly be creating new questions. They employ dozens of test question writers in every subject that they create tests for. Each question has to be tested on one of the "experimental" sections before it can be used. As it turns out, it is extremely difficult to write a good question and quite a lot of the questions are thrown out because they do not have attractive ICC curves, and this is even with a great number of employees with a PhD level degree in psychometrics.

This is another major drawback of the 3-parameter model--the level of statistical and psychometric understanding needed to produce and evaluate test questions puts it out of the range of the majority of teachers. Few teachers have this level of understanding. Even fewer have access to the hundreds of students necessary to test questions out before we use them on an actual exam.

In summary, the good news about IRT is that we finally have a way to assess students' understanding that is objective and based solely on performance and contains no bias with respect to race, class, gender or any other personal characteristics. In other words, we have a basis for assigning grades that would have a reliably meaningful interpretation, that wouldn't have any of the messy questions involved in it that came up when evaluating the B- that John got on his essay. The bad news is that the level of technical understanding necessary to produce such questions, and the amount of resources needed to validate the questions is far beyond the practical reach of most teachers.

What other methods do psychometricians use?

Clearly, there are other ways to assess people's learning than just by using test questions. While the IRT models described above are typically used for multiple-choice questions (or other questions that have a definitive right or wrong answer), psychometricians have devised other ways to assess abilities that don't rely on test questions. This section will describe expert judging, which is used to assess ability in areas like writing or visual arts, and which is closest to what actual classroom teachers do.

How do you judge the quality of an essay or a painting? Unlike test questions, there is not necessarily a "right" or "wrong" answer here. The basic approach of expert judging comes from the principle that "I'll know it when I see it." While multiple-choice questions can be scored by a computer, to evaluate writing or art, a human being has to actually read or look at the work and make a decision about its quality. In classrooms, the teacher is the judge of quality. In standardized testing situations, the company doing the testing typically hires a panel of human expert judges. These people actually read the essays or look at the art and render a judgment.

A good question to ask is how do we know whether or not, when two people read an essay, they are judging it by the same criteria and applying the same weight to those criteria? The short answer is, we don't. If you'll recall the study described in Chapter 2, Paul Diederich, the Director of Research in English at Educational Testing Service (ETS), ran a study in which 300 college essays were each graded by 60 evaluators from a variety of disciplines and sorted into nine categories from "worst" to "best." In the words of the investigators:

> *The result was nearly chaos. Of the 300 papers, 101 received all nine grades, 111 received eight, 70 received seven, and no paper received less than five. (p442)*

In other words, 94% of the papers were put into seven or more different quality categories--there was very little agreement about what constitutes a "good" essay. If you are a professional psychometrician, this is a serious problem, and a good deal of the work Diederich and others like him have done over the years has been to fix this. The two components of the solution are rubrics and expert panels.

Rubrics

The first component of the solution to the interpretation problem is the rubric. A rubric is a tool designed to allow both teachers and learners to understand and describe the quality of various components of a complex piece of work, like an essay. For example, the 6+1 Trait® Rubric for grades 3-12 breaks the evaluation of writing into seven major categories, each with subcategories, for a total of between thirty-one and thirty-six categories:

1. **Ideas**: Main idea, Details and support, Reasoning/thinking, Evidence selection and acknowledgement, Awareness/engagement of reader
2. **Organization**: Lead and conclusion, Transitions, Sequencing, Pacing, Purpose/text structure, Title (optional)
3. **Voice**: Engagement with reader, Individual expression, Tone, Commitment, Fit with audience/purpose
4. **Word Choice**: Meaning, Quality, Usage, Grammar
5. **Sentence Fluency**: Structure, Sense and rhythm, Variety, Connecting sentences
6. **Conventions**: Spelling, Punctuation, Capitalization, Grammar/usage, Editing needed, Bibliography (optional)
7. **Presentation**: Font style/size, White space, Text features (optional), Visuals and Graphics (optional), Handwriting (optional)

In each category, a piece of writing can be scored on the following 6-point scale, from worst to best:

1. Beginning
2. Emerging
3. Developing
4. Capable
5. Experienced
6. Exceptional

The first three categories are grouped as "Not Proficient," and the second three categories are labeled "Proficient." Furthermore, every category-score combination is described in a way that facilitates grading. A small segment of the rubric table is reproduced here for illustration:

	Not Proficient	Proficient
	3. Developing	4. Capable
Main Idea	Suggests a main idea, but the direction of the piece is still unclear	Has a clear, focused main idea

On the plus side, rubrics take something that is large and complex, like a piece of writing, and break it up into smaller, more manageable elements about which useful feedback can be provided. They can serve as a vehicle for communication between teachers and students (and their parents). They also provide a structure for organizing instruction about a particular topic. Furthermore, rubrics are very flexible and can be designed for use with most any subject area, from science projects to interpretive dance. Particularly for more prominent subject areas, like writing, there are already a number of widely-available, high quality rubrics that are the result of extensive research and development.

On the minus side, for other subject areas developing good rubrics is very challenging and time-consuming. Grading with a rubric can also be very time-consuming. (Imagine grading 30 student essays with the above rubric that has over thirty categories!) Frequently, effective use of rubrics requires formal training for the instructor. Finally, some critics argue that the use of rubrics can lead to an overly-formulaic style.

This last point was made humorously in a 2004 article in *The Atlantic* called "Would Shakespeare get into Swarthmore?" which used the then-new rubric for the new SAT writing test to evaluate the writings of famous authors like Ernest Hemingway, Gertrude Stein, William Shakespeare, and infamously Ted Kaczynski (aka the Unabomber). It perhaps will not be surprising to find out that of these, the Unabomber scored the highest on his writing.

In all seriousness, however, rubrics have proven to be an excellent tool for making meaningful assessments of student work, and conveying meaningful feedback to the students themselves. That being said, converting rubric-based assessments into grades is still not a straightforward task. In professional psychometrics, this is accomplished with the second component

Chapter 4
Measurement

Chapter 5
Motivation

Chapter 6
Positivity and Mindfulness

Chapter 7

Your Brain on Grades

Chapter 8

"Grade Inflation"

Chapter 9

Accountability

Chapter 10

Education Reframed

Chapter 11

For Students

Chapter 12
For Teachers

Chapter 13
For Parents

Chapter 14

For Administrators

Chapter 15
Conclusion

References

ASTIN, A. W. (2012). *Assessment for excellence: The philosophy and practice of assessment and evaluation in higher education*. Rowman & Littlefield Publishers.

Bridgeman, B., McCamley-Jenkins, L., & Ervin, N. (2000). Predictions of Freshman Grade-Point Average from the Revised and Recentered SAT® I: Reasoning Test. *ETS Research Report Series*,2000(1), i-16.

Diederich, P. B. (1966). How to measure growth in writing ability. English Journal, 435-449.

Elliott, R., & Strenta, A. C. (1988). Effects of improving the reliability of the GPA on prediction generally and on comparative predictions for gender and race particularly. *Journal of Educational Measurement*, 25(4), 333-347.

Finkelstein, I. E. (1913). *The marking system in theory and practice (No. 10)*. Baltimore: Warwick & York.

Goldman, R. D., & Widawski, M. H. (1976). A within-subjects technique for comparing college grading standards: Implications in the validity of the evaluation of college achievement. *Educational and Psychological Measurement*, 36(2), 381-390.

Grossman, J. (1978). Fair Labor Standards Act of 1938: Maximum struggle for a minimum wage. Monthly Labor Review, 22-30.

Harvard University. Annual report of the President of Harvard University to the Overseers on the state of the university for the academic year 1830-31. Harvard University Archives. Accessed 4/9/15: http://nrs.harvard.edu/urn-3:hul.arch:15000

Harvard University. History of the Presidency of Josiah Quincy. Accessed 4/9/15: http://www.harvard.edu/history/presidents/quincy

Joint American Educational Research Association. (1999). *The standards for educational and psychological testing*. American Educational Research Association, Washington, DC.

Kohn, A. (2014). *The myth of the spoiled child: Challenging the conventional wisdom about children and parenting*. Da Capo Press.

Laska, J. A., & Juarez, T. (1992). *Grading and marking in American schools: Two centuries of debate*. Charles C Thomas Pub Limited.

Moran, G. F., & Vinovskis, M. A. (1985). The great care of godly parents: Early childhood in Puritan New England. *Monographs of the Society for Research in Child Development*, 24-37.

Norris, F. (2014) Fewer U.S. Graduates Opt for College After High School, *The New York Times*, April 25, 2014. Accessed 3/31/15, http://nyti.ms/1k1JtDS

Peterson, R. A. (1983) Education in Colonial America. *The Freeman: Ideas on Liberty*, 33, September 1983.

Rudolph, F. (1977). *Curriculum. A History of the American Undergraduate Course of Study Since 1636*.

Shih, M., Pittinsky, T. L., & Ambady, N. (1999). Stereotype susceptibility: Identity salience and shifts in quantitative performance. Psychological science,10(1), 80-83.

Smallwood, M. L. (1969). *An historical study of examinations and grading systems in early American universities*. Harvard University Press.

Smith, J. K. (2003). Reconsidering reliability in classroom assessment and grading. *Educational Measurement: Issues and Practice*, 22(4), 26-33.

Spielmaker, D. (n.d.). Historical Timeline -- Farmers & the Land. Retrieved March 31, 2015, from https://www.agclassroom.org/gan/timeline/farmers_land.htm

Wilbrink, B. (1997). Assessment in historical perspective. *Studies in Educational Evaluation*, 23(1), 31-48.

Willingham, W. W., Pollack, J. M., & Lewis, C. (2002). Grades and test scores: Accounting for observed differences. *Journal of Educational Measurement*,39(1), 1-37.