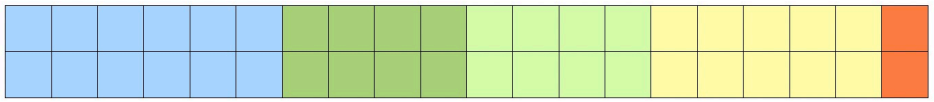
Better Grades



a proposal for Arlington Public Schools

by Miles Townes

Better Grades

A proposal for Arlington Public Schools

by Miles Townes
July 2026

Contents

INTRODUCTION

Although we tend to think of equity in terms of marginalized students, it is in fact a question of how we treat all students. Our grading system harms most students, but it harms marginalized students most of all.

This proposal represents several years of analysis and evaluation of the grading system used by Arlington Public Schools. When I began this project, my goal as a teacher was to ensure I was measuring my students' success as accurately and fairly as possible. In what follows, I show that our grading system is neither accurate nor fair.

We owe our students better. We owe them **equity**. It is widely understood that equity means more than mere equality. While equality assumes everyone has the same choices, equity requires us to pay more attention to those for whom some choices are not available or possible. This is most important with respect to marginalized students — Black students, immigrant students, disabled students, and others — but equity bears on how we treat all our students, whether or not they belong to marginalized populations.

Consider that our students are required to participate in our education system; the only students who can avoid compulsory public education are those whose families can afford private school or have the capacity to homeschool. As good as our schools are — and we can be proud of their excellence — every student in every school is not allowed to make a better choice.

The only way we can justify this imposition is by ensuring our schools expand students' possibilities. In a democracy, compulsory public education must be an equalizer to every extent possible, in terms of the choices available to students during school and the choices available after. While we at the district level cannot control the state standards nor the

resources county taxes provide, we are yet obliged to do our best for every student in every aspect within our control.

Our grading system fails any reasonable test of both equality and equity. It does not grade students equally with respect to their effort and achievement, and it especially harms low-performing students by limiting their choices after graduation. It is important to recognize that this is not merely a problem of math, but fundamentally a problem of measurement, which requires methodological and ethical choices as to the most accurate and fairest ways to measure. The measurements we use are neither accurate nor fair, because those were never considerations in the origins of our grading system.

In my previous career I trained as a social science researcher, with extensive graduate work in quantitative and qualitative analysis. That is: I was trained how to measure social phenomena and outcomes, using a wide variety of tools and techniques. This document includes a lot of mathematical analysis, but I have included many charts and tables to make it clearer (in fact, this project began life as a slide deck, now over 230 slides long). If I am over-explaining at times, I apologize: I am being thorough to ensure that everyone — including secondary students — can understand the problems with our grading system.

As we will see, that system *increases* inequality — sometimes drastically. It exaggerates the very differences public education is meant to bridge. To the extent our students recognize or suspect this at some level, it will inevitably lead to disengagement and disruption.

This is especially true for students who enter our schools with significant disadvantages. Because we know those students are already likely to perform lower than their peers, sustaining a system that deprecates their performance creates a structural harm that multiplies their disadvantages.

We know inequity is an issue in Arlington Public Schools. The table below shows graduation rates for APS as reported by Virginia's Department of Education.[1] Note the significant disparity for Hispanic, Economically Disadvantaged, English Learners, and Students with Disabilities.

APS Graduation	Current Rate
All Students	89%
Asian	93%
Black	89%
Hispanic	75%
White	96%
Multiple Races	97%
Economically Disadvantaged	81%
English Learners	75%
Students with Disabilities	76%
Homeless	83%

These disparities have real consequences for students' life choices and outcomes, to the point that they may well indicate marginalization within and by our school system.

1 "Federal Graduation Indicator." at http://schoolquality.virginia.gov/divisions/arlington-county-public-schools#desktopTabs-8 (12/1/2023)

The disparities our students experience open the door to a Title VI claim against the district. The Supreme Court's decision in *Hazelwood* affirms that "Where gross statistical disparities can be shown, they alone may in a proper case constitute prima facie proof of a pattern or practice of discrimination" with respect to Title VII (i.e. employment discrimination) claims. While Title VI (which covers education) does not include 'pattern or practice' language, the Justice Department notes that principles derived from Title VII claims can "inform the investigation and analysis of Title VI claims".[2]

If graduation rates are the pattern, grading is a major part of the practice. We could not possibly justify our grading system in court. Our grading system is indefensible by any math or methodology. In fact, after I began exploring some of these ideas as a part-time teacher with Arlington, I began certification to be a full-time teacher. In my education program, I searched for a rationale for traditional grading systems. I found no real defense of those practices, but I did find many teachers, scholars, and critics who agree that they are harmful. This proposal builds on and extends their work.

I propose we replace our traditional system with one that is relatively easy to implement, gives students more agency in their education, and builds on the strengths of our teachers. I call this approach 'outcome gradation', because it is based on the educational outcomes in Bloom's taxonomy. It means less no-value work for low-performing students, more meaningful work for high-performing students, and less grading work for teachers. Our grades will be better — that is, more accurate and more equitable.

This proposal starts with Chapter I — **Grades**, which discusses why we use grades, what they should do, and some of the criticisms of traditional grading. The next three chapters cover the functions that grades perform in our school system:

---

2 U.S. Department of Justice (nd). "Title VI Legal Manual (updated)." At https://www.justice.gov/crt/fcs/T6Manual6#AH (12/2/2023)

II. Measure, **III. Report**, **IV. Compare**. In these chapters, we see how the math in our grading system makes our grades less accurate, increases their bias, and reduces student motivation.

In Chapter V — **Work** — I show how our grading system requires students and teachers alike to do a tremendous amount of work that provides no usable information about students' progress. Chapter VI introduces and explains my solution — **Gradating** — and shows how it solves almost all the problems our grading system creates.

I believe APS has a sincere and meaningful commitment to equity. This document is not at all a critique of that commitment, but a proposal for how we can better realize it. I appreciate your consideration, and I look forward to working with you towards better grades for all Arlington students.

I. GRADES

Grades are vital to our educational system. Any grading system must have a clear purpose. We use grades to measure students' achievement, to report that achievement, and to compare students. There is significant and growing criticism that points to problems in traditional grading.

Given the scope and ambition of public education as a social enterprise, some indicator of progress is necessary. We call that a 'grade', and for most students, our grading system is the only constant in their journey from sixth grade through graduation. Grades show not just the effort and progress of our students, but also reflect on our teachers, schools, and our district as a whole. We can hardly imagine an educational system without grades — though many students wish we could!

APS Policy I-7.2.3.34 "Reporting Student Progress and Grades"[3] describes the goals of our grading system:

> Students' grades shall accurately reflect students' knowledge and skills mastery to date. Grading practices should be accurate, bias-resistant, and motivational and recognize that students develop and demonstrate knowledge and skill mastery in different forms and at different rates. Grading shall use calculations that are mathematically sound, easy to understand, and correctly describe a student's level of academic performance through their mastery of grade level/ course standards. Grades shall not reflect student behavior, how a student compares to another student, and/or a teacher's implicit bias.

3 This document was very difficult to find and has a massive URL. I shortened it: https://tinyurl.com/5n93pwv6 (12/14/2025)

By this standard, our grading system fails our students, teachers, and families alike. Our grades do not accurately measure, are not bias-resistant nor motivational, and do not use mathematically-sound calculations. This last point has eluded many critics of traditional grading, and it is the concern I will spend the most time on. In math and measurement terms, our grading system is a mess that promises to make grades easy to understand, but then grossly misrepresents the performance of the vast majority of our students.

We should first step back and look at the **purposes** grades serve.[4] First, they **measure** student achievement. Grades are not, in practice, a single system of measurement, but rather an amalgamation of all the various measurements teachers take in the course of a student's career. Sometimes this process obscures massive differences in teachers' individual approaches, or differences from school to school.

Second, our grades serve to **report** — to inform students and others of their progress. Parents often want to know how well their students are doing, but the most important reporting function is to keep students apprised of progress towards their own educational goals, whether that is graduation, college, a scholarship, or something else. Grades also inform other schools of students' achievement: when a student matriculates from middle to high school or applies to college, grades report their educational progress to date.

Third, grades **compare** students. It is important to understand that grades should not be based on comparison to other students (this is called curving, and is addressed later), but grades will inevitably used to compare students. Again, the most important comparison is a student's progress from quarter to quarter, year to year. Grades also allow us to compare students within a class, classes in a school, or schools in the district. We know there may be differences in grading from one

4 The discussion of purposes of grades is derived from Guskey, T.R. (2015). *On Your Mark*. (Bloomington, IN; Solution Tree Press).

class to the next, but still tend to believe our system provides a reliable distinction between, say, an A student and a C student.

As central as grades are to public education, it is no surprise grading practices and policies draw criticism. One common critique is that grades are prone to inflation. Recently, American College Testing published a study pointing to "a decade of dramatic grade inflation"[5] — but it is worth noting that doubt about the reliability of grades is a major selling point for standardized tests. And the ACT study likely casts too narrow a window; use of the phrase 'grade inflation' blew up in the 1970s (see below), and has been a perennial concern ever since:[6]

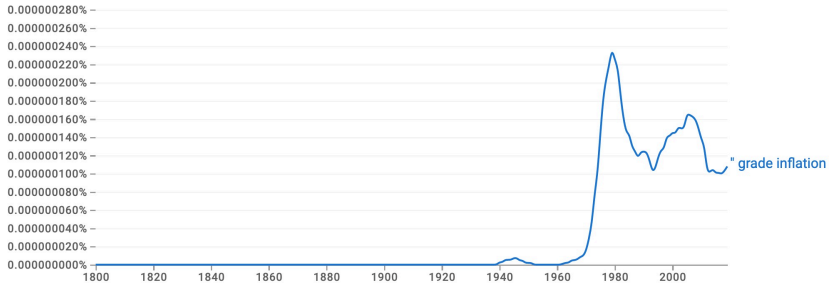


Figure 1.1 Ngram for "grade inflation"

More recently, however, **criticism** of traditional grading has focused on whether that system serves our students and schools. Thomas R. Guskey's *On Your Mark* questions nearly every aspect of traditional grading, including percentage grades, bell curves, class rank, and mathematical algorithms. He

---

5 Hess, F. (2023). "Grade Inflation Is Not a Victimless Crime". *Forbes,* 9/5/2023 at https://www.forbes.com/sites/frederickhess/2023/09/05/grade-inflation-is-not-a-victimless-crime/?sh=5da69fd117b2. The study itself is available at https://www.act.org/content/dam/act/secured/documents/Evidence-of-Grade-Inflation-in-English-Math-Social-Studies-and-Science.pdf (12/1/2023). Note that the author does not describe the effect of increased AP/IB/DE enrollment.

6 https://books.google.com/ngrams/

demonstrates that "most current policies and practices are bound more by tradition than by evidence of effectiveness."[7]

Alfie Kohn offers another angle of attack on grading systems in *Punished By Rewards,* arguing "grades in particular undermine intrinsic motivation and learning, which only serves to increase our reliance on them".[8] Kohn points to a tremendous body of research to make this point. Not only do grades hurt lower performing students, but they also encourage top students to be less creative, more cautious, and more superficial in their work.

Joe Feldman, author of *Grading for Equity,*[9] echoes and extends the critiques offered by the previous authors, with a focus on equity. He argues that a proper grading system must be 1) accurate, 2) bias-resistant, and 3) motivational. APS clearly has adopted his language for our "Reporting Student Progress and Grades" policy, which is a good choice.

However, this choice is completely undone by the 100-point grading scale promulgated in Policy Implementation Procedure I-7.2.3.34 PIP-2 (next page).[10] There we see our commitment to accuracy, bias-resistance, and motivation falter and fail.

Part of the oversight here may be due to Feldman's too-brief treatment of the problem of accuracy. While he does mention the mathematical issues in traditional grading scales, I believe he misses their deeper dysfunction. He also has a very traditional mindset when it comes to assessment design; I will demonstrate how we can resolve a significant share of the problems in grading by being more deliberate about assessment design — in effect, front-loading our efforts in a way that Feldman does not consider.

---

7 Guskey, T.R. (2015). *On Your Mark*. (Bloomington, IN; Solution Tree Pr.) p. 109.

8 Kohn, A. (2018). *Punished by Rewards* (New York; Mariner). p. 203

9 Feldman, J. (2018). *Grading for Equity.* (Thousand Oaks, CA; Corwin)

10 Also hard to find, large URL: https://tinyurl.com/yc6786p2 (12/14/2025)

Reporting Grade and Percentage Equivalency	Grade Descriptor
A 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100	**Excellent Achievement** Demonstrates accurate and thorough understanding of course content; Creatively applies and extends knowledge and skills in content area; Consistently meets course standards; Overall performance and student products reflect a deep level of analysis of the content and critical thinking
B+ 87, 88, 89	**Good Achievement** Demonstrates an accurate understanding of course content; Applies knowledge and skills in content area; Consistently meets course standards; Overall performance and student products reflect a consistent understanding of the content and critical thinking
B 80, 81, 82, 83, 84, 85, 86	
C+ 77, 78, 79	**Satisfactory Achievement** Demonstrates acceptable understanding of course content; Consistently demonstrates basic knowledge and skills in content area; Meets course standards; Overall performance and student products reflect a consistent understanding of most of the content and reflect some degree of critical thinking
C 70, 71, 72, 73, 74, 75, 76	
D+ 67, 68, 69	**Unsatisfactory Achievement (Passing)** Demonstrates limited understanding of course content at this time; Inconsistently demonstrates knowledge and skills in content area around some standards; Inconsistently meets course standards; Overall performance and student products reflect an inconsistent understanding of the content
D 60, 61, 62, 63, 64, 65, 66	
E 0 – 59	**Unsatisfactory Achievement (Failing)** Demonstrates minimal understanding of grade level content at this time; Inconsistently demonstrates knowledge and skills in content area around most standards; Overall performance and student products reflect minimal understanding of the content

Figure 1.2 APS Grading Scale

These three books are just a fraction of the literature on traditional grading. Although they capture many problems with traditional grading — the system APS currently uses — these authors significantly understate the measurement problems. In what follows, I address those problems with more rigorous, mathematical analysis, and describe a solution I have used in my own teaching and grading.

II. MEASURE

We use grades to measure students' achievement. Equitable grading means those measurements have to be accurate, but our grading system is prone to inaccuracy:

- **We assign arbitrary numerical values to letter grades.**
- **Grades focus on low-level outcomes, not mastery.**
- **Assessments are almost never validated.**
- **Multiple choice tests are often too 'sensitive'.[11]**
- **Rubrics are implemented inconsistently.**
- **0 = 50 is masking tape over serious problems.**

The first function of grades is to measure students' progress and achievement. Before we can choose how to measure, we have to decide *what* to measure. What do achievement and progress look like? The point of education is to help students learn, but how do we know what they've learned, and how well?

To answer these questions, American educators typically look to Bloom's taxonomy of educational outcomes (next page). The taxonomy describes a hierarchy of educational outcomes, with basic learning — 'remember' — at the bottom, up through progressively higher levels of learning to 'create'. Mastery is not simply being able to remember facts, but being able to apply, explain, and ultimately create new knowledge.

The taxonomy emerged from a meeting of college examiners at the American Psychological Association, which led to a series of conferences from 1948 to 1953, and finally to a book, *Taxonomy of Educational Objectives*.[12]

---

11 'Sensitive' here means how well the instrument detects errors.

12 Bloom, Benjamin, et al. (1956). *Taxonomy of Educational Objectives.* (London: Longmans). p.5 at https://tinyurl.com/yvfjw4zp

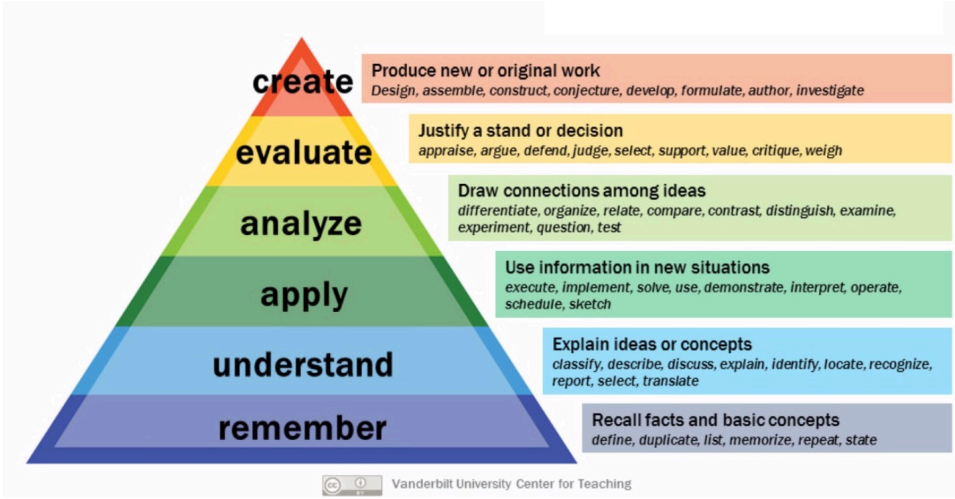


Figure 2.1 Blooms Taxonomy of Educational Objectives

A fair amount of research has gone into validating and extending Bloom's taxonomy in educational psychology and pedagogy. Seventy years later, it remains the most compelling description of what learning actually looks like. It is not simply a description, though, but also a values statement that reflects the kinds of learning outcomes we want from children in our schools. This is what we wish to measure.

The biggest problem in grading is that we assign **numbers** to these outcomes. We will see more of the math soon, but for now consider this: how much harder is 'analyze' than 'apply' in percentage terms? 10%? 50%? It does not make sense because these are different cognitive processes. And yet the basic premise of our grading system is that we can think of 'analyze' as a percentage of 'apply', 'understand' as a percentage of 'apply' and so on. Any numbers we assign these levels will ultimately be arbitrary, because they are different tasks.

Another problem that arises is that some of these outcomes are much easier to measure than others. It is easy to measure whether a student remembers facts or concepts, and much harder to measure whether they can evaluate that knowledge or

create with it. And so grading often favors low-tier outcomes, especially **recall**, and not mastery. This ends up being busy work for many students, and excess grading for teachers.

Once we have decided what to measure, we need to begin creating instruments to do so. In order to ensure that our measurements are accurate, we need some process to **validate** our measurement instruments — that is, our assessments. That process should look somewhat like this for assessments:

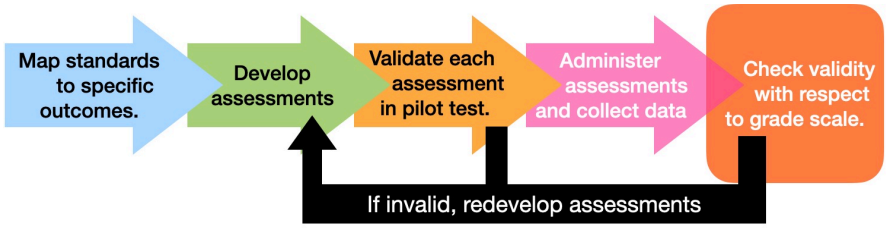


Figure 2.2 Assessment Validation

Something similar happens for SOL, ACT, and SAT exams, but it almost never happens for the assessments teachers use in the classroom. In fact, the SAT and the ACT go through extensive validation to ensure they measure achievement aptly, but they are still subject to stiff controversy.

The lack of validation means we are putting tremendous trust in teachers to design and give assessments that accurately measure students' progress. This is not necessarily a bad thing, but later we will see how we can leverage this trust to make grades more meaningful and less biased. In the meantime, the variety of assessments and their varying validity means that even before students sit down to an assignment, it is likely some degree of measurement error is built into the assessment.

The subjective nature of most grading compounds these errors. One of the first studies to look at the problem — more than a hundred years ago — found that teachers gave grades

ranging from 50 to 97 on the same paper.[13] A similar experiment for math teachers found even wider variation: a range from 28 to 95.[14] The original study was repeated recently with astonishingly similar results — despite the teachers spending 20 hours training in a writing workshop.[15] In my own teaching, I often ask students and other teachers what the difference is between an 89 essay and a 90 essay. Nobody can describe the difference, and yet that one point can have significant consequences — as we will see.

Multiple choice assessments were created in part to try to address the subjective nature of grading, as well as make it easier for teachers to grade papers. But multiple choice assessments are easy to focus on recall, understanding, or very basic application, and are much less suitable for higher-tier outcomes. They also tend to be too long, with too many questions. Consider this: for a 30-question multiple-choice test, what are the chances of a student guessing 14 correct answers if each question has four possible answers? The odds are effectively zero: .00543, or one-half a percent. And 14/30 is 47% — a failing grade.

Some teachers justify big tests because 'there is a lot of knowledge to cover' — which makes sense if every fact or skill is a discrete point in the universe of knowledge.[16] But if knowledge is more like a web or fabric, then everything in a given assessment should be related and connected somehow. In which case, students need not pick out every single thread in

---

13 Starch, D., & Elliott, E. (1912). "Reliability of the grading of high-school work in English." *The School Review*, 20(7), 442-457.

14 Starch, D., & Elliott, E. C. (1913). Reliability of grading work in mathematics. *The School Review*, 21(4), 254-259.

15 Brimi, H. M. (2011). "Reliability of grading high school work in English." *Practical Assessment, Research and Evaluation, 16*(17), 1–12.

16 I honestly cannot imagine teaching this way.

a scrap of knowledge to prove they understand it. That is **sensitivity**: how many threads are enough?

Our grading scale aligns with undetectable differences in sensitivity. In our thirty-question test, 26 correct answers is a B and 27 correct answers is an A. The difference in probability is about the same as between 10 and 11 on a 12-question test. The figure below shows the sensitivity levels for various lengths of exam, assuming items are the same value and similar difficulty.

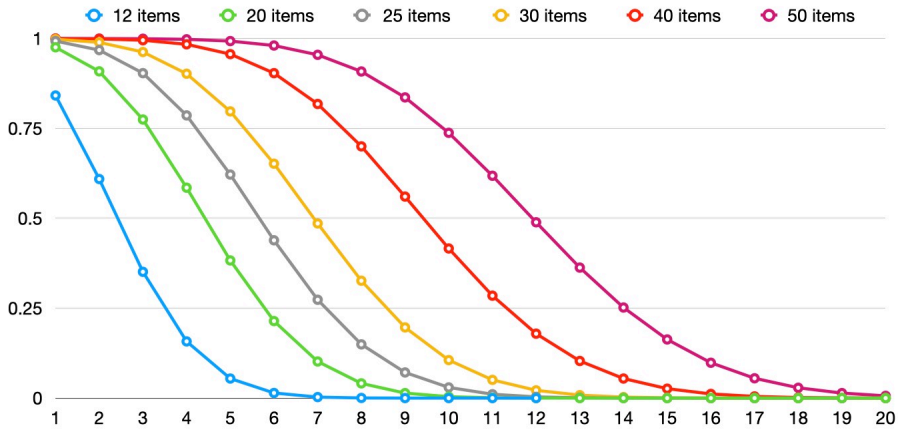


Figure 2.3 Sensitivity levels for 12, 20, 25, 30, 40, and 50 item assessments

As you can see, an exam with 12 questions of comparable difficulty offers plenty of sensitivity: there is effectively zero chance of guessing 8 correct answers, which is the lowest passing grade on our current scale. Note that a 50-question exam gets to near zero probability after 20 questions, so the remaining 30 questions are not doing any work in terms of sensitivity.

In practice, a teacher with a 30- or 50-question battery of multiple-choice questions can do better by giving students only 12-15 of those questions, and save the rest for potential retakes if students fail. Big, high-stakes exams and assessments are

often drastic overkill for the amount of information we actually need about students' progress. Fortunately, we have much better ways to design assessments than multiple choice.

Increasingly, teachers are turning to more qualitative assessments that are then graded by **rubrics.** In a rubric, a teacher sets out the criteria by which the assessment will be graded, and assigns those criteria specific point values. Rubrics are helpful in reducing bias in grading: in one experiment, a 4.7% bias in favor of students with white-sounding names was all but eliminated by standardized rubrics.[17]

While rubrics can make grades less biased, that does not mean the grades are any more accurate as measurement. In general, it is possible for measurements to have zero bias and still be completely inaccurate. And in practice, rubrics are often implemented in ways that contribute to the problems of grading, rather than resolving them. For example, the rubric on the following page was disseminated in an APS online course in social studies, to be used to grade all formative assessments.

There are three big problems with this rubric. First, it does not help us discern grades accurately. The qualitative criteria for 90 and 100 are exactly the same, and so are 80 and 85. But the difference between 70 and 75 is considerable, even more so the distance between 65 and 70. The rubric creates a staircase that is very steep for low-performing students but very shallow for high performing students. It should be the other way around.

Second, this rubric focuses on 'completeness' throughout, and does not distinguish educational outcomes per Bloom's taxonomy. In that respect, it does not comply with Policy Implementation Procedure I-7.2.3.34 PIP-2. A student could get an A in this class without ever 'creatively applying or extending' their subject knowledge.

---

17 Quinn, D. M. (2021) "How to Reduce Racial Bias in Grading." *Education Next*, Winter 2021. at https://www.educationnext.org/wp-content/uploads/2022/01/ednext_XXI_1_quinn.pdf (12/14/2025)

Formative Rubric - Social Studies

100 points	Full Mastery - Can Teach Others	-All parts are completed thoroughly -All sections show clear attention to detail and thoughtful answers
90 points	Full Mastery - Can Teach Others	-All parts are completed thoroughly -All sections show clear attention to detail and thoughtful answers
85 points	Mastery - Can Perform Without Supervision	-All parts are completed though some answers can be a bit more developed/ explained -Most sections show attention to detail and thoughtful answers
80 points	Mastery - Can Perform Without Supervision	-All parts are completed though some answers can be a bit more developed/ explained -Most sections show attention to detail and thoughtful answers
75 points	Partial Mastery - Can Perform With Limited Supervision	-Most sections completed though some answers can be a bit more developed/ explained -Most sections show thoughtful answers
70 points	Partial Mastery - Can Perform With Limited Supervision	-Most sections completed but answers need to be more fully developed -Most sections completed but at a basic level
65 points	Can Perform With Supervision	-Student does not grasp main ideas or answer questions properly -Most questions are answered but at a basic level
50 points	Cannot Perform	-Incomplete: Unable to fully evaluate mastery

Third, the rubric does not accurately reflect the grading scale. 100 and 90 are both an A. 85 and 80 are both a B. Similar with 75 and 70 — both a C. The rubric has no + grades — B+, C+, or D+ — on its scale, and the percentage points are set near the edges of the range for each letter grade. An 80 B is one point above a C+, a 70 C is one point above a D+.

Formative Rubric - Social Studies (Revised)

95 points	**Full Mastery - Can Teach Others**	-Student extends or connects main ideas to new domains -All sections show clear attention to detail in depth -All answers are fully responsive to questions -All answers are clear and comprehensible
85 Points	**Mastery - Can Perform Without Supervision**	-Student shows mastery of main ideas -All parts are completed -Most sections show attention to detail -All answers are responsive to questions -Most answers are clear and comprehensible
75 points	**Can Perform With Limited Supervision**	-Student shows partial grasp of main ideas -Most parts are completed -Some sections show attention to detail -Some answers are responsive to questions -Most answers should be clearer or better developed
65 points	**Requires Supervision**	-Student shows limited grasp of main ideas -Most parts are completed -Few or no sections show attention to detail -All answers need to be clearer or better developed
50 points	**Incomplete**	-Incomplete: Unable to fully evaluate mastery

A more accurate rubric (previous page) would reflect the actual grading scale (setting aside the problems with that scale for now) by centering each grade in its range. It would have consistent steps between grades and reflect Bloom's Taxonomy.

But even though this rubric is a significant improvement over the original, it still has important flaws. One flaw is the implication that every assignment captures the entire range of cognitive tasks in Bloom's taxonomy. Even using APS's language, the implication that every assignment allows a student to "Creatively apply and extend knowledge and skills in content area" is false. And if it were true, it would place a severe burden on students and teachers both.

We usually record our grading measurements — whether from multiple choice, a rubric, or some other instrument — as a point value. As we have seen, APS uses a 100-point scale in which 90 to 100 is an A, 87-89 is a B+, 80-86 is a B, and so on. Statisticians call this a 'ratio scale', because each data interval can be expressed in terms of a ratio of one to another. Note that this is a 100-point scale, but it has 101 values because 0 (no points) is also a possible grade.

The most important ratio in this scale is the ratio of passing grades to the scale itself. Passing grades on this scale are the 41 values from 60-100: we can call that the 'pass window' (next page). When we compare the pass window to the 100-point scale, the ratio is 41/101 — so about 40%. Not only does the 100-point scale imply that *every* assignment has the full range of Bloom's cognitive levels, but *all* of those outcomes occur in only a 40% pass window! This implies that almost 60% of the grading scale tells us nothing useful about students' progress.

As described above, the decision to set passing at 60 points is entirely arbitrary, as is the decision to use a 100-point scale. This system dates back to the era when mechanical grading devices were first available, and teachers had to assign grades to percentages because the machines only reported numbers (and note that this began before the Bloom committee formed

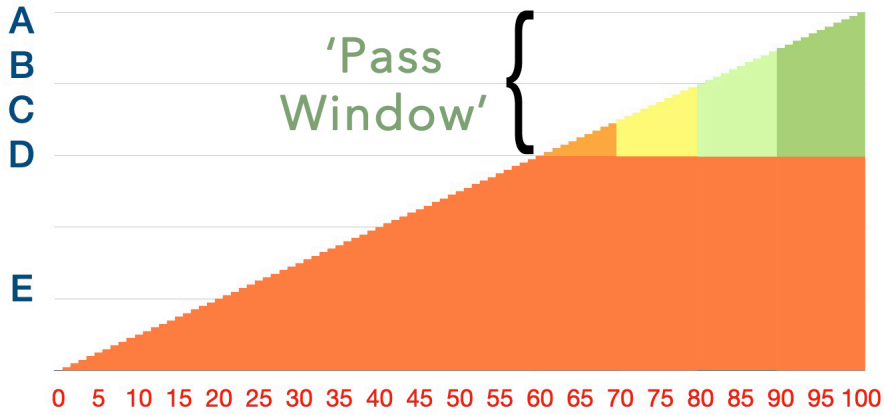


Figure 2.4 The Pass Window

its taxonomy). Even within the pass window, the different thresholds for grades are completely arbitrary.

To see this clearly, imagine a class that is easy to measure: weightlifting. If the teacher wants to assess students on the bench press, would it make sense for him to say 60 pounds to 66 pounds is a D, 67 to 69 is a D+, 70 to 76 is a C, and so on? So a student who starts at 75 and builds up to 90 is an 'A'. What about students who start at 10 pounds and work their way up to 59 pounds? If these grades make no sense in a subject that is easy to quantify, they make less sense in any subject that is not.

The problem with our weightlifting example is not just the arbitrariness of the numbers, but the implications with respect to effort: the student who increases their strength by 49 pounds has somehow accomplished nothing — an E, by our grading scale — while a student with a 15-pound gain gets an A.

Because the 100-point scale is a ratio scale, we can express every score as a ratio of the grade above it. This number tells us how much relative effort a student needs to move from one grade to another. In the table on the next page, I have simplified the grading scale to whole letter grades.

You want / You have	100%	90%	80	70%	60%
90%	do 11% more work				
80%	25%	12.5%			
70%	43%	28.6%	14.3%		
60%	66%	50%	33%	16.7%	
50%	100%	80%	60%	40%	20%

Figure 2.5 Relative effort from one letter grade to a higher grade

As you can see, the effort is not even across the grading scales. It takes more work, in relative terms, for a D student to get to a C (16.7% more) than for a B student to get to an A (12.5%). This means low-performing students have a steeper hill to climb. Higher grades require less marginal effort, per the chart below:

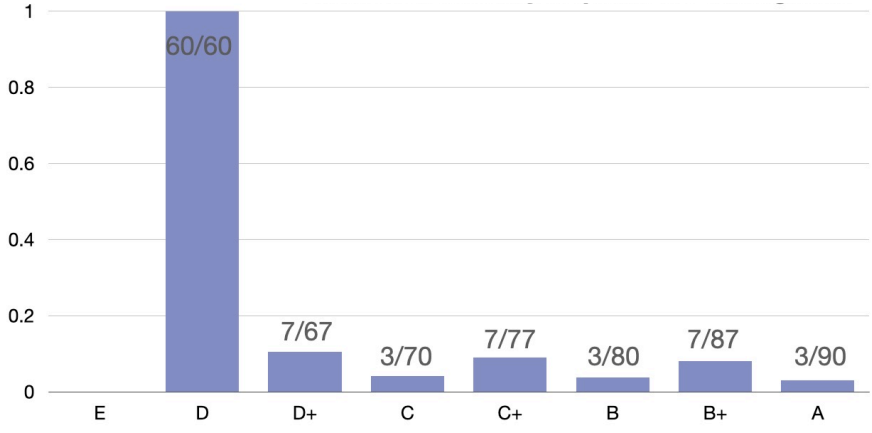


Figure 2.6 Marginal effort required for each grade step

A low C student requires *three times* the marginal effort to push their grade up to a C+ as a B+ student requires to push their grade up to an A. Our grading scale rewards *diminishing* marginal effort: it is harder for a low-performing student to improve their grade than it is for a high-performing student. This is almost certainly a discouraging climb for struggling students.

If we want to create an isometric scale — in which each step requires equal relative effort — we have a few ways we could do that, as below. We cannot start from 0, because the 0 to 59 interval is so massive. Our scale would something like 0-59 F, 61-108 D, etc. Instead we must choose the relative interval first, and build the scale from there. Each of the scales in Figure 2.7 has a *consistent* ratio from grade to grade, where our existing grading scale has diminishing effort up through higher grades.

A	**102 - 120**	**94 - 100**	**107 - 128**	**73 - 80**
B	**86 - 101**	**81 - 94**	**89 - 106**	**67 - 73**
C	**72 - 85**	**69 - 80**	**73 - 88**	**61 - 66**
D	**60 - 71**	**59 - 68**	**61 - 72**	**55 - 60**
E (=50)	**50 - 59**	**50 - 58**	**50 - 60**	**50 - 55**
rationale	**r=59/50**	**r=58/50**	**r=60/50**	**50+10%...**

Figure 2.7 Isometric (equal effort) grading scales

One solution meant to rescue low performing students from the sub-60 abyss is to record zeroes as 50. There is a solid rationale to this policy, but ultimately it is **masking tape** over structural flaws: it is low cost, it holds together okay, it is better than nothing. But it is ugly and does not solve the real problem.

0=50 mostly helps low-performing students, for whom one missed assignment can destroy their grade. Figure 2.8 (next page) shows the grade for a student with a 65 average across

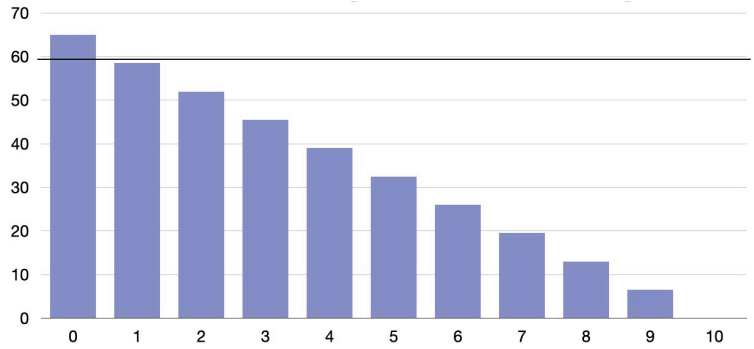


Figure 2.8 Effect of n missed assignments for student with 65 average

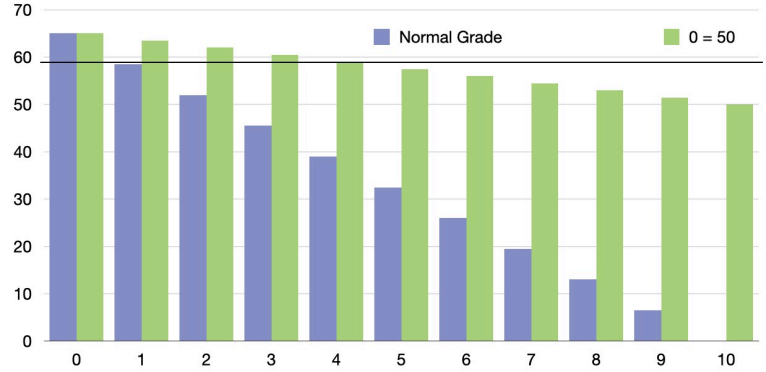


Figure 2.9 Effect of zero = 50 for the same student

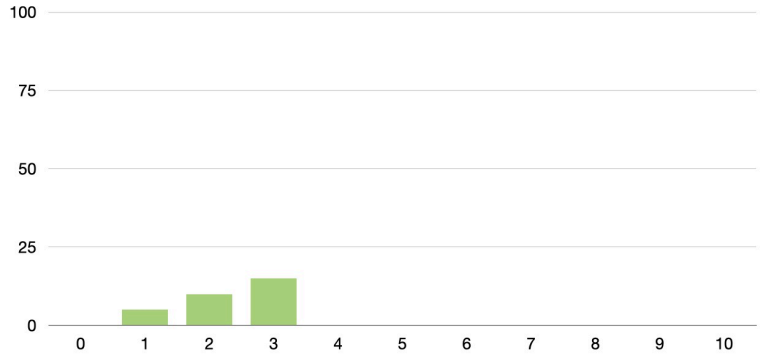


Figure 2.10 Net gain in points to a 65-average student for n missed assignments

ten assignments if n number of assignments are zero. Per figure 2.9, 0=50 gives the student only three more chances to attain a passing grade. After three zeroes, the student's grade is a whisker above passing. The fourth zero tanks their grade.

Many people — including teachers — feel this amounts to grade inflation. Figure 2.10 shows the net gain in points for the student in the previous charts. The 0=50 policy results in the student gaining perhaps 15 points that they did not earn. Suffice to say 0=50 is not the boon to students it is made out to be, and is only necessary because of the arbitrarily small pass window in our 100-point scale.

Once we recognize the arbitrariness of our grading scale, we can begin to appreciate that the glue holding it all together — what makes our grades even remotely plausible — is teachers. The less the math behind our grading system makes sense, the more implicit trust we invest in the professional and pedagogical expertise of our teachers to measure students' progress. As we will see, our grading scale makes so little sense that the trust we place in teachers ends up being massive. The upshot of this proposal is that our trust is not misplaced; it is misapplied. We can put it to better use.

III. REPORT

We use grades to report students' progress to the students, their parents, teachers, and other schools. Our grading scale provides false information:

- **Incommensurate grading scales distort and deflate.**
- **Curving (or lining) grades garbles the measurements.**

Grades are how we report students' progress to the students, their parents, teachers, and other schools. Their value as reporting depends on the choices we make in assigning meaning to the raw measurements. In the previous chapter we looked at only part of the grading scale. In this chapter we will look at the entire, three-part grading scale.[18]

100-point scale	Letter Grade	Quality Points
90-100	A	4.0 (3.75 to 4.0)
87-89	B+	3.5 (3.25 to <3.75)
80-86	B	3.0 (2.75 to <3.25)
77-79	C+	2.5 (2.25 to <2.75)
70-76	C	2.0 (1.75 to <2.25)
67-69	D+	1.5 (1.25 to < 1.75)
60-66	D	1.0 (0.75 to <1.25)
0-59	E	0.0 <0.75

Figure 3.1 APS grading scale

18 https://go.boarddocs.com/vsba/arlington/Board.nsf/files/DLYLJ256B13C/$file/I-7.2.3.34%20PIP-2%20%20Reporting%20Student%20Progress%20and%20Grades%20(Secondary).pdf p. 4

We saw in the last chapter that the 100-point scale is 'ratio' data. Statisticians and social scientists distinguish four different types of data, and have different tools to analyze them:

- **ratio** data have both intervals and true zero, so each data point can be expressed as a ratio — like weight or height.
- **interval** data are based on scales that have discrete steps but no true zero: temperature in Fahrenheit or Celsius.
- **ordinal** data show different ranks that may not have a quantitative scale: prime beef, choice, select, dog food.
- **nominal** data indicate categories that cannot be directly compared: apples, oranges, bananas.

We can see in our grading system that we have two ratio scales — the 100-point raw scale and the 4-point quality scale. We can compare the two ratio scales to each other, as below:

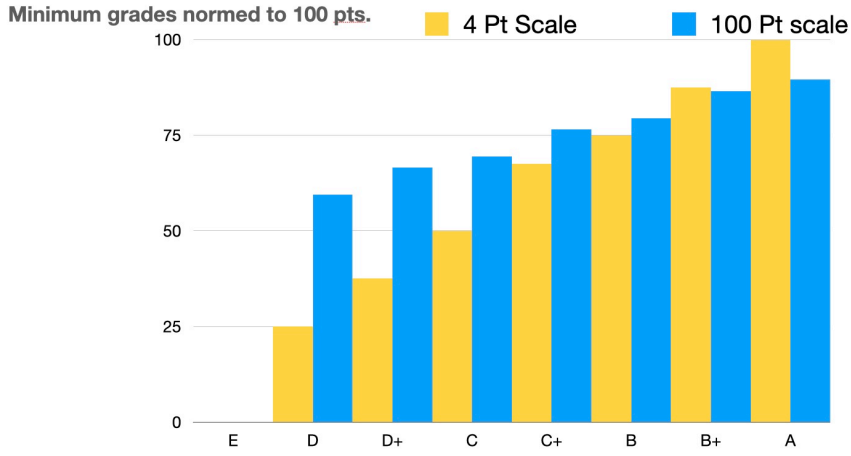


Figure 3.2 Comparison of APS ratio scales

We can see one problem clearly: the scales are completely different. They are so different as to be incomparable, that is: **incommensurate**. The previous chart shows the minimum score for each grade: we can see that the 4-point scale (yellow) is fairly steep, while the 100-point scale (blue) is fairly shallow. The 100-point scale makes differences between students look smaller, while the 4-point scale makes differences look bigger.

One scale is used to measure, the other scale is used to report, but the bigger problem is that these two ratio scales are connected through an ordinal scale. So when a teacher puts in a raw grade in the 100-point (ratio) scale, that grade is converted to the letter (ordinal) scale, and then reported as quality points on their transcript on the 4-point (ratio) scale. We can see that conversion process in the figure below (simplified to letter grades only, no + grades):

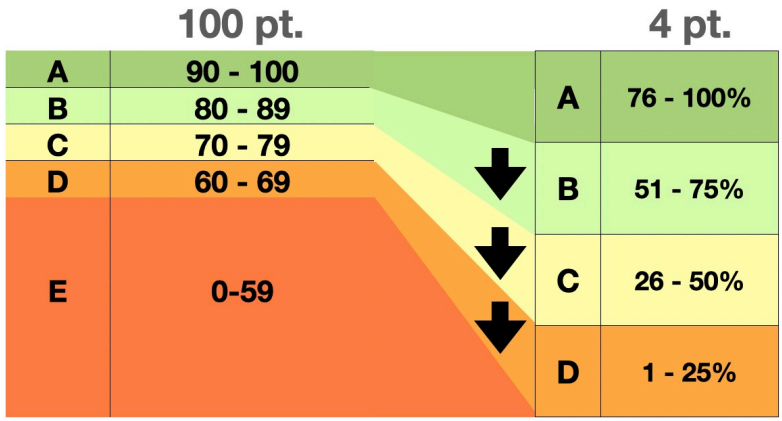


Figure 3.3 Transition loss across ratio scales (no + grades)

So, for example, a 75% ratio on the 100-point scale becomes a 50% ratio on the 4-point scale. In the plainest language we can use to describe this problem: this is bad math. As measurement, it is even worse. This is impossible to justify in

mathematical terms, and the actual consequences from combining these two scales range from bad to catastrophic.

Two key problems arise in the move from 100-point raw scale to the letter scale to the 4-point quality scale. The first problem is that we lose — throw away — a tremendous amount of data. We start with at least 100 possible grades, but really 1,000 or 10,000 depending on how many decimals we record in our grade book. We replace that data with no more than 9 possible grades (including + grades). The collapse from 10,000 possible data points to 9 is roughly akin to the move from CD-quality audio back to 8-bit audio (as in early Nintendo systems.)

Once the letter grade is reported, it is then converted from the letter (ordinal) scale into the 4-point (ratio) scale. It is an odd choice to throw away all that data and then try to reconstitute it — like recording a CD of 8-bit audio — but we could sort of make it work, if the two scales had the same ratios. They do not.

That brings us to our second major problem: the change in grades (in ratio) terms from one scale to another, which we can call **distortion**. For top students, it will increase their grades. For all other students, this distortion will result in a significant loss of points. For example, if a student scores a 65% on an assessment, that student's grade becomes a D for 1 quality point on the 4-point scale. 1/4 = 25%. That student loses 40 percentage points. A student who scores a 75% gets a C, which earns 2 quality points: 2/4 = 50%, a loss of 25 percentage points.

We can see the consequences for students on the chart on the next page. The red bars shows how many percentage points a student loses at each grade; blue shows how many they gain. From the chart, we see that most students lose points, but students with low A grades benefit by inflation. For everyone except A students, the grading system has built-in **deflation**.

Note that the pass window in the 4-point scale is far more generous than the 100-point scale — it covers the entirety of the 4-point range. But that is only because we are discarding *all* the measurements we collected on failing students — 59% of the

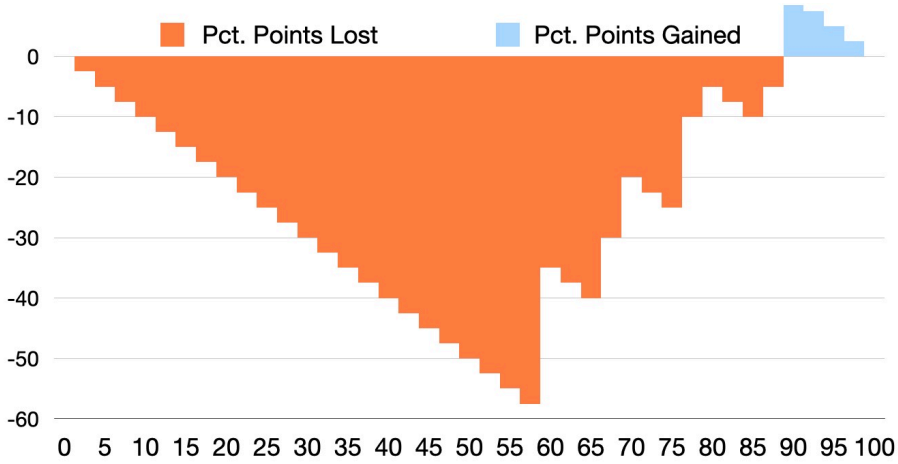


Figure 3.4 Grade distortion from 100-pt to 4-pt scale (in 2.5 pt. steps)

range of our 100-point scale — thus hiding how much they have accomplished. A student can complete more than half the work and yet we still report that as a zero. In terms of the 'report' function of grades, this is grossly inaccurate. Or consider a student whose grade is in the 30% range. In order to pass, they need to expend twice as much effort — for a 60% — and that grade is then reported as 1 on the 4-point scale for 25%. Even if the student is not proficient at math, at some level they might sense the reward is not worth the effort.

It is important to recognize that the ratio-ordinal-ratio grading scales are not the product of deliberate thought. Rather, this system has been cobbled together based on traditions and practices going back more than a hundred years, with each scale meant to serve a different audience.

The chart on the next page shows 100-point and 4-point scales that are commensurate, in that they have equal intervals and the same zero point in ratio terms. Adjusting the ratio scales to bring them into parity is one solution to the problems of loss and distortion. If this does not make sense, that confusion is entirely due to the indefensible math in our current grading.

	100 pt.	4 pt.
A	100	4
B	88.9	3.556
C	77.8	3.112
D	66.7	2.668
E	55.6	2.224

Figure 3.5 Ratio scales made commensurate

I have been told by teachers that the two scales are not meant to be compared, but as long as they connect via the letter scale, it is vital that they be comparable. If we cannot make the math work, we cannot claim any integrity for the grades we give students.

Even if our raw scale made sense, we would still have problems with our grades. Those problems are worse if teachers are curving (or lining) grades. **Curving** a grade means mapping a class's scores to the normal distribution. The normal distribution (below) is a statistical model that describes a lot of data — but not most data. Most of the time when we curve

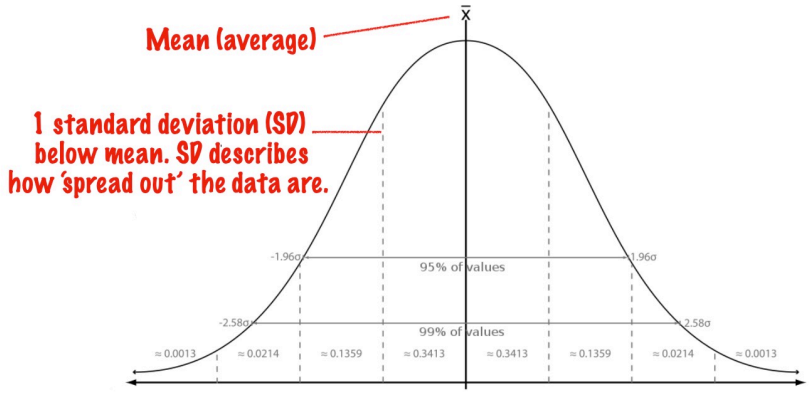


Figure 3.6 The normal distribution

grades, we are only assuming the scores are normally distributed, not actually checking the distribution.

To be clear, curving is a reporting decision, not a measurement decision, because it affects how the grades are interpreted after data are collected. Curving is not part of the data collection process itself.

Curving grades requires four steps:

1) find the mean or average of grades;
2) find the standard deviation;
3) choose a grade to anchor the mean; and
4) translate the scores in terms of the standard deviation.

This is easy to implement in a spreadsheet, but obviously requires a bit more work than recording raw grades. The figure below shows a proper grade curve for a class of 25 students.

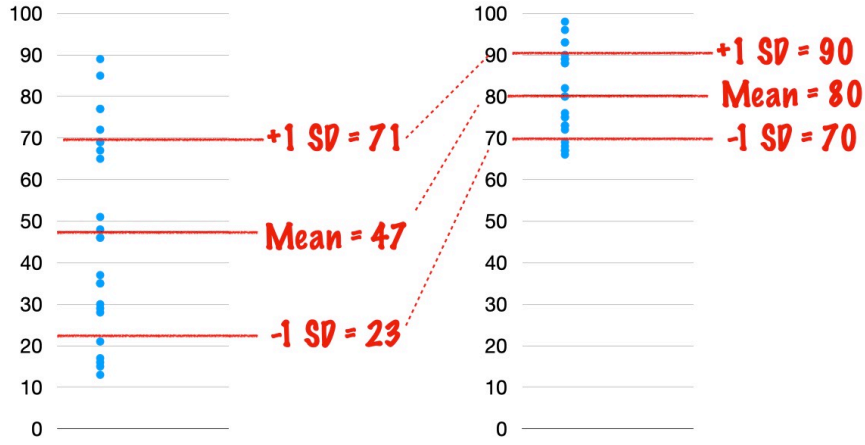


Figure 3.7 Grading curve set to the normal distribution

You can see that no student gets a 100, but no student fails. This is not always the case, but often curving results in a narrower distribution of scores. In effect, the curve says that

worse students were not so much worse than average, and the best students were not so much better.

As mentioned, this involves some amount of calculation, so many teachers instead use 'curve' methods that are in fact a 'line'. **Lining** the grade can include:

- Setting the class mean average to an arbitrary score — like 80. While this is straightforward, it can sometimes mean that top students experience grade deflation, if the highest possible score is 100. A student who scored a 94 in a class that averaged 70 might see everyone else's grade increase by 10 points, but their own grade only increase by 6 points.

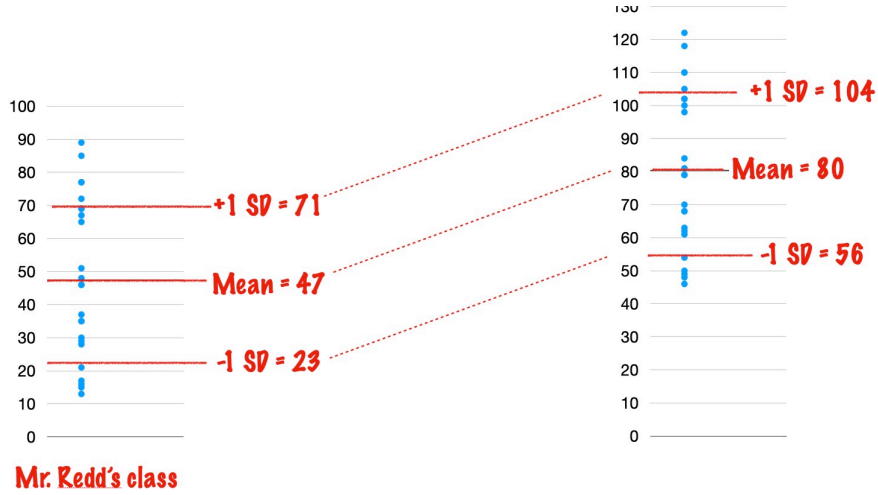


Figure 3.8 Grades 'lined' by mean arbitrarily (from 47 to 80)

- A teacher might also 'line' the class median average to an arbitrary score. Median and mean can produce very different results, depending on how lopsided a class's scores are. The

chart below shows a class lined to each. Though the lines look similar, the score distributions (blue dots) are very different.

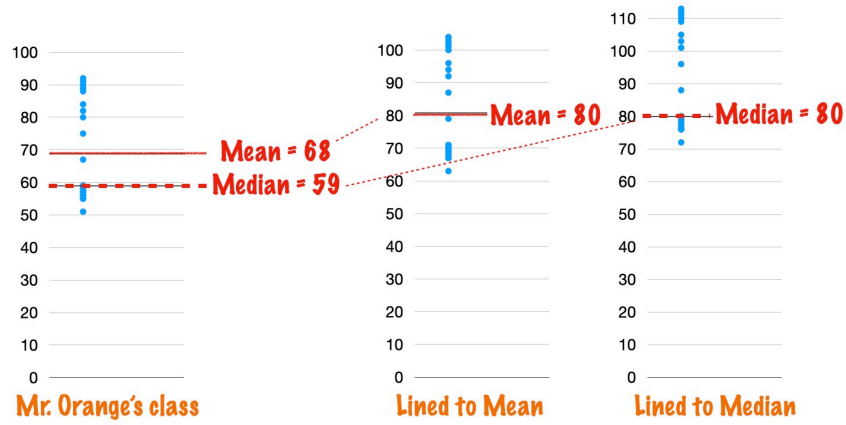


Figure 3.9 Grades lined by mean and median arbitrarily (to 80)

- A teacher might also line grades by setting the highest score in the class to 100, as below. This makes the entire class's score dependent on that student; the social pressure to perform worse may become intense.

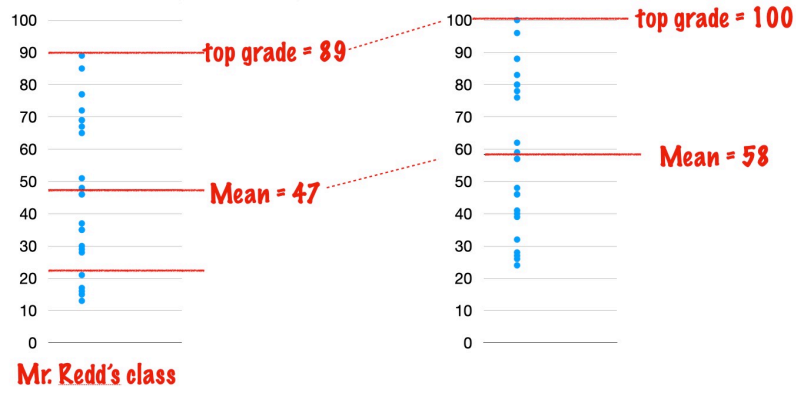


Figure 3.10 Grades 'lined' to top score (89 to 100)

Each of these adjustments can have very different consequences. The chart below shows a comparison of one hypothetical teacher's use of different systems. The median and mean give the same result because the distribution of grades is symmetric, which is not always the case.

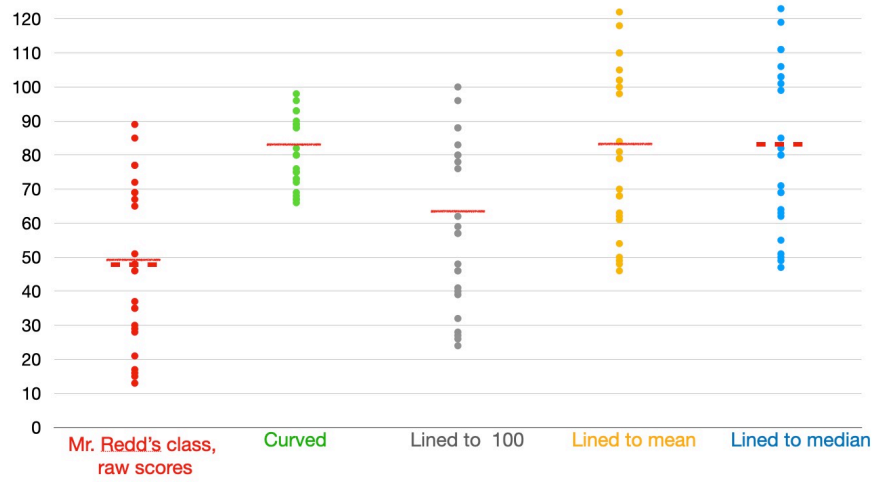


Figure 3.11 Comparison of different curve/line grading algorithms

Even teachers who do not explicitly curve grades often face the problem of class scores being lower than expected or accepted. The teacher has several options:

- Cover less material
- Cover less demanding material
- Make assessments easier
- Offer lots of make-up opportunities
- Offer extra credit

All these adjustments to the curriculum can result in an implicit curve that makes even less sense than an explicit curve. This chart illustrates the problem:

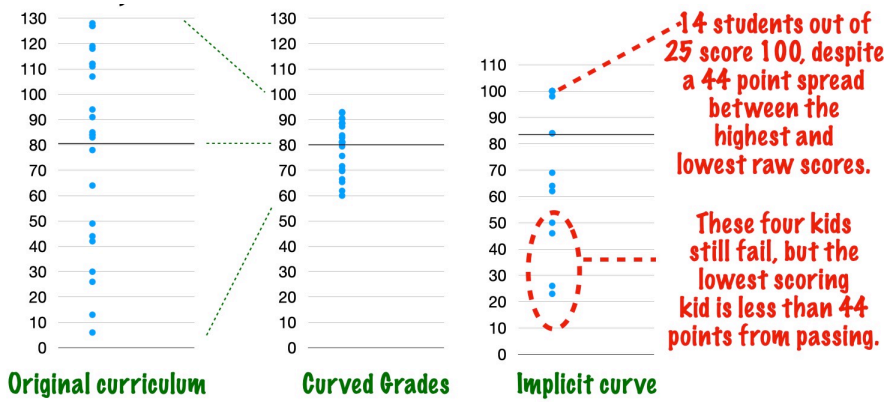


Figure 3.12 Explicit vs. implicit grading curve

A class that has a high number of perfect scores — say more than 5% of students — very likely has an implicit curve intended to make the class easier for failing students. This is not always a bad thing, but over time it is likely to lead to top students showing less effort and not performing to their full potential.

So far in our discussion of lines and curves, we have not seen any zeroes in the raw scores. Zeroes add an additional complexity to the math. The chart on the next page shows the same class with and without a single zero score. The no-zero scores are on the left, the zero scores are on the right. If scores are lined to the mean (the second column), then a single zero means scores are higher for everyone (the fifth column; again, we are looking at the distribution of dots, not the lines). If scores are curved (the third column), then a single zero means scores are higher for students below the mean, and lower for students above the mean. That is, the top-performing students get less of

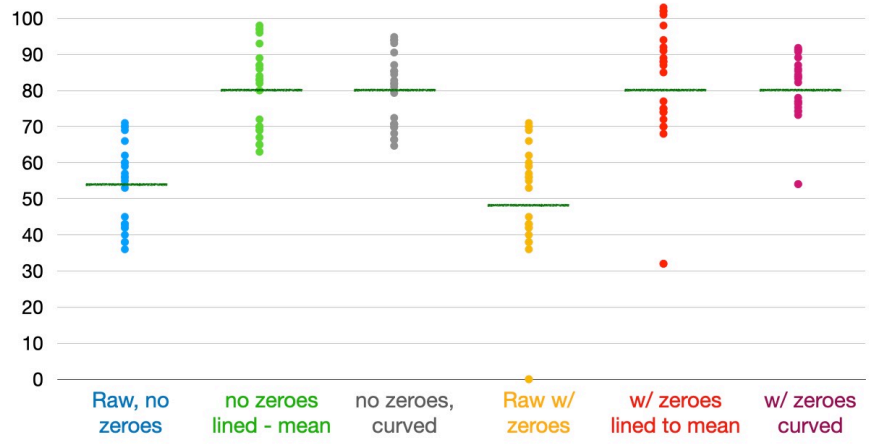


Figure 3.13 Various grading algorithms with and without zero scores

a boost, and the lower performing students get more boost. 0=50 helps solve some of these problems, of course.

Yet the bigger problem driving all of this math is that our grading system is based on fixed ratios and a very narrow pass window, as previously discussed. That means making the class easier for lower students also makes it easier for top students. Instead of making the class easier for everybody, responsible pedagogy would better support low-performing students while keeping the material adequately challenging for top students.

Our schools' main way to give top students more challenging classes are Advanced Placement, International Baccalaureate, and Dual Enrollment classes. Unfortunately, our grading system means these classes make things worse for everybody, a problem we will explore in the next chapter.

IV. COMPARE

We use grades to compare students' progress in a class from start to finish, to compare students to one another, to compare classes and teachers, and to compare schools. Our grading system gives us flawed comparisons by:

- **Distorting grades in favor of top-performing students.**
- **Devaluing low-performing students' efforts.**
- **Rewarding 'grade grubbing'.**
- **AP/IB/DE premiums make everything worse.**

Comparison is the crux of the equity problem with our current grading system. When we use APS's grading scales to compare our students, we systematically overstate higher performing students' achievement and deprecate struggling students' progress.

We saw in the last chapter how the loss and distortion in our scales harms struggling students, making it look like they performed worse than they did. If this still does not make sense, we can look at how our transcripts report these outcomes to see the difference. In transcripts, a student's GPA is the mean average of student's quality points across all their classes.

Imagine two students: William is a bright but unmotivated student who does the minimum to satisfy his parents, managing to earn a string of 89.5 grades on the 100-point scale. Because our practice is to round up, he is a straight A student.

Our other student, Val, is bright and really does apply herself, resulting in 100s in all her classes. In fact, she could probably score higher if her classes were not too easy because of the narrow pass window and the implicit curve.

Let's compare Val's transcript to William's on the next page:

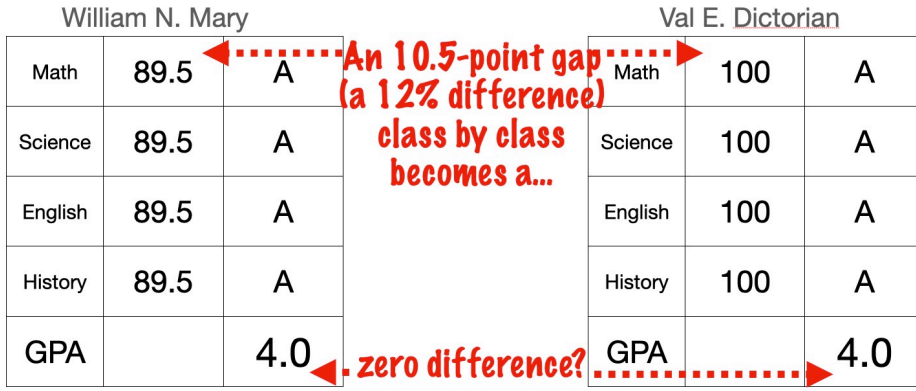


Figure 4.1 Transcript comparison of William and Val

Although our measurements recorded a 10.5 point gap (a 12% difference, relative to William's grades) between these students, we are reporting zero difference — the exact same grade — in their transcript. If that seems inconsequential, let's meet a third student: Bobby is diligent and studious, but his parents never graduated from high school. He also works weekday evenings to help support his family. Despite his solid effort in school, he barely misses earning a C+ in each of his classes: he scores 76.4 all the way, resulting in a 2.0 GPA. The tables below compare William and Bobby's transcripts.

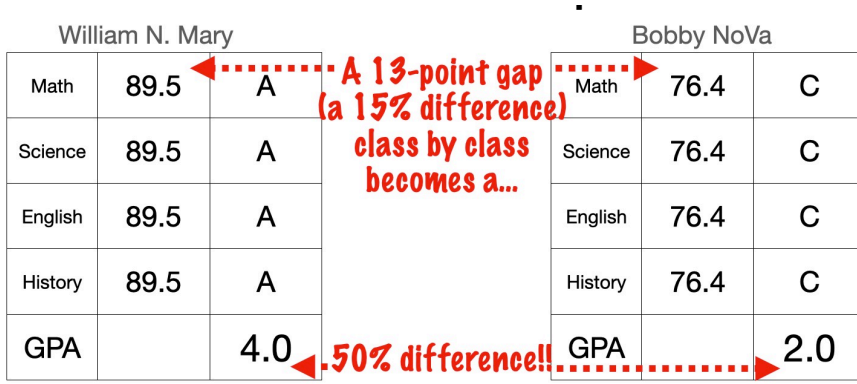


Figure 4.2 Transcript Comparison of William and Bobby

As you can see, a 13-point difference in raw scores (which is a 15% relative difference) becomes a 50% difference by the time it is output on their transcript. For William and Val, a 12% difference was nothing. For William and Bobby, a 15% difference is *everything*. William is accepted to one of the best schools in the state, while Bobby stays home and perhaps attends community college. And yet that 13-point gap is not enough to justify putting these students in different classrooms, much less different campuses.

The **devaluation** of Bobby's achievement is not only difficult to defend as methodology, but it will likely have serious consequences for his future education and career options. If this does yet not seem like an equity issue, replace 'Bobby' with 'Roberto': you can see exactly this dynamic in dozens of classrooms across Arlington.

This devaluation is a real consequence of the grade distortion problem we looked at in Chapter 3. Recall that Figure 3.4 showed the grade distortion students experience across the 100-point and 4-point scales. It is reproduced as Figure 4.3.

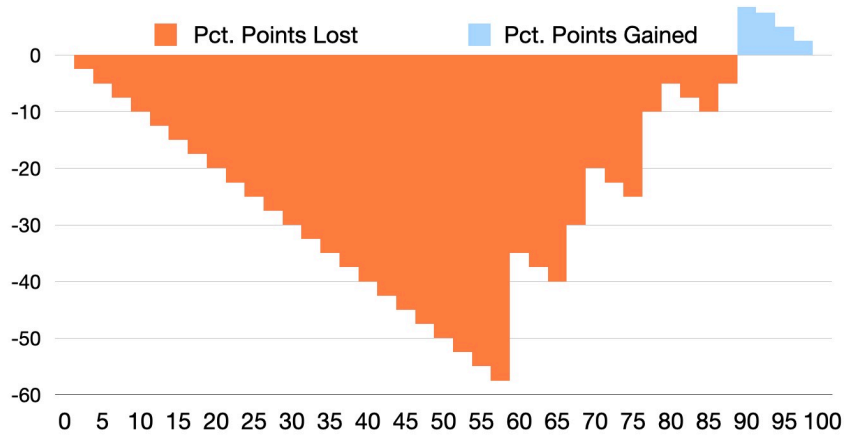


Figure 4.3 Grade distortion from 100-pt to 4-pt scale (in 2 pt. steps)

On this chart, William's 89.5 scores in each class win him 10.5 points from distortion. Bobby's 76.4 scores lose him 26.4 points. It should be obvious that students like Bobby will tend to lack motivation in a system like this.

In fact, research shows that students judge their progress not in absolute terms, but in relative terms. When higher standards are put in place, lower scoring students often do worse on standardized tests: "students near the bottom may perceive themselves as losing ground and give up on graduating as a result."[19] By exaggerating the difference between low- and high-performing students, our current grading system encourages exactly this invidious comparison — without the benefit of higher standards.

Granted, both William and Bobby are imagined to be edge cases, where two students land very close to a grade cutoff. The problem gets worse the more assessments a student has to complete. Assume a teacher grades on nothing but multiple choice tests (which is not strong pedagogy). If the teacher gives 10 assessments over a quarter with 20 items each, the total 'grade space' is 200 items, as in this table:

# Assessments (20 pts. each)	10	15	20
'Grade space'	200	300	400
Lowest A	180	269	358
Highest B+	179	268	357
Percent diff.	0.55%	0.37%	0.28%

Figure 4.4 'Grade spaces' and edge cases for A/B+ students

19 Betts, J.R. and Grogger, J. (2000). "The Impact of Grading Standards on Student Achievement, Educational Attainment, and Entry-Level Earnings." NBER Working Paper No. 7875, September 2000, p. 21; (12/1/2023) at https://www.nber.org/system/files/working_papers/w7875/w7875.pdf

A student with an 89.45 will receive a B+ — or 3.75 quality points — while a student with an 89.50 will receive an A — a 4.0. If the grade space is 400 items, the two students might have a .0028 — 0.28% — difference in their raw scores, but that translates into a .0625 — 6.25% — difference in their transcript grades. Our grading system amplifies the difference more than 22 times!

One effect of this is to encourage **'grade grubbing'**. Teachers complain about students begging for higher grades, but those students might get a 6% boost in quality points for a class from grubbing just one question in 400. Our system makes it profitable for students with edge-case grades to 'grub'. Our transcript comparisons ought to take into account whether William merely badgered his teachers into those A grades.

We also saw in the last chapter that because of the narrow pass window in our grading system, grading practices that make low-performing students more likely to pass tend to make school less challenging for high-performing students. One way we try to address that problem is by offering Advanced Placement, International Baccalaureate, and Dual Enrollment courses (AP/IB/DE). Unfortunately, the grades we give these classes make our system a lot worse for nearly everyone.

AP/IB/DE students earn an additional quality point, implying that higher quality work deserves a higher reward. Unfortunately, the rationale for our grading scale does not support the idea that this premium point reflects quality. Per APS Policy Implementation Procedure I-7.2.3.34 PIP-2, an A in *any class* indicates a student "Creatively applies and extends knowledge and skills in content area". That means an A in a premium class cannot represent a higher cognitive outcome, and at best shows more work at that level.

Here we run into the problem with ratios again: we saw earlier how the two ratio scales — the 100-point raw scale and

the 4-point quality scale — are incommensurate, and that causes problems for most students' grades.

We are now introducing a third ratio scale, the 5-point premium quality scale. However, there is not a new raw scale for premium classes, which means we have to again translate raw scores into premium points to determine the loss involved:

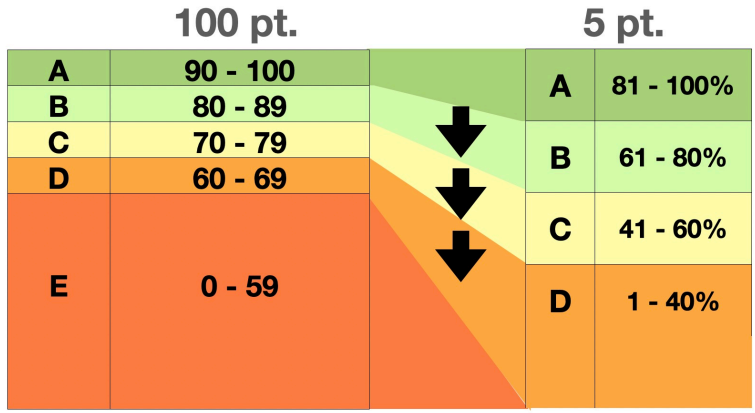


Figure 4.5 Transition loss for premium grades

The 4-point scale implies 76% is an A, 51% is a B, and so on. We can see the 5-point scale implies *less* A-level work for premium classes, contrary to the intent of the extra quality point: 20% rather than 25%. In fact, the scale implies AP/IB/DE students complete *a lot more* D-level work — 40% of their reported grade, instead of 25%. In actual practice, premium points often reflect students' outcomes over a higher *quantity* of work, not work of higher quality.

Recall that our calculations for Figure 3.4 were based on the assumption that all students were enrolled in regular classes, and so the quality points possible were limited to 4. If we redo the numbers for students not in AP/IB/DE classes, we get much worse results (next page). Every student in a regular class loses:

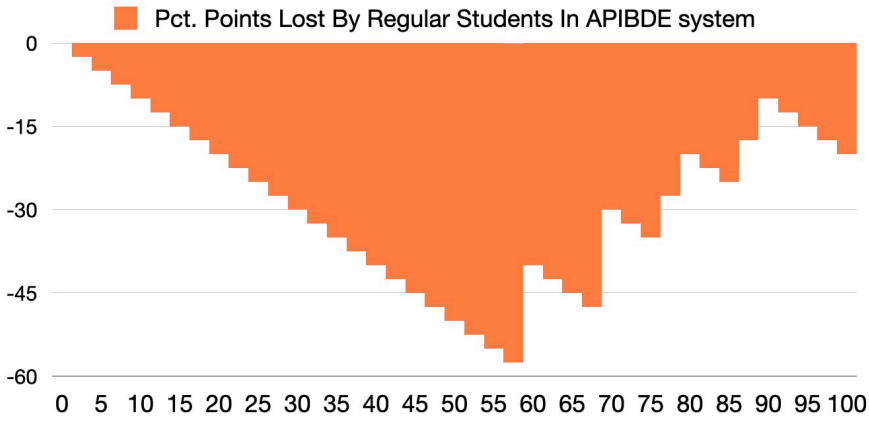


Figure 4.6 Distortion for non-AP/IB/DE students on a 5-point scale

In fact, the premium for AP/IB/DE grades does not offset distortion for most students enrolled in premium classes. A truncated view of the distortion for AP/IB/DE classes (below) shows that premium points offset losses only for students who

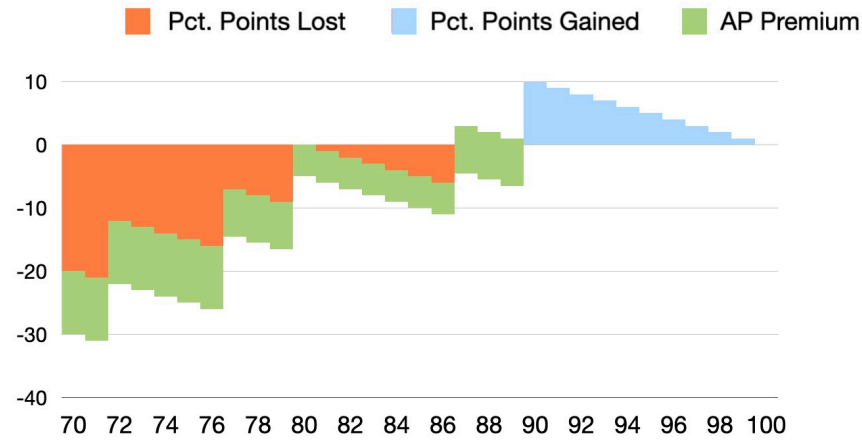


Figure 4.7 Premium offset of distortion loss for high-performing students

score 80, 87, 88, and 89. Top students still receive inflated grades, and everybody else suffers deflated grades.

Not only does the relative benefit from the AP/IB/DE point premium not offset losses for most high performing students, but the lowest-graded students benefit the most (see below). An A student receives one point, or 20% of their total grade.

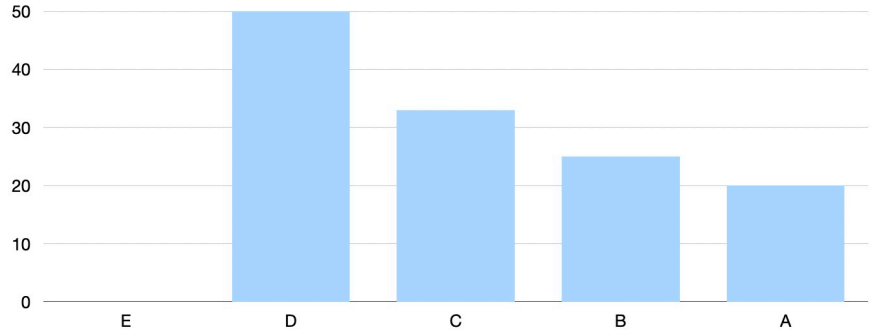


Figure 4.8 AP/IB/DE premium as a percentage of total grade

Meanwhile, a D student receives the same point, for 50% of their grade. The premium most rewards the least effective students in these classes.

Back to William: if he takes a premium class, he is not necessarily doing higher quality work. He is likely doing a larger quantity of work at a similar level, work that he should be comfortable completing. He should have little problem earning an 89.5 in his AP class, in terms of his cognitive abilities. Whether he is diligent enough to do more work is the only question at issue. If Bobby is in the same classes and manages to find the time to complete the extra work, his relative position improves somewhat due to the premium. Yet he is still pretty far from William in their transcripts (next page).

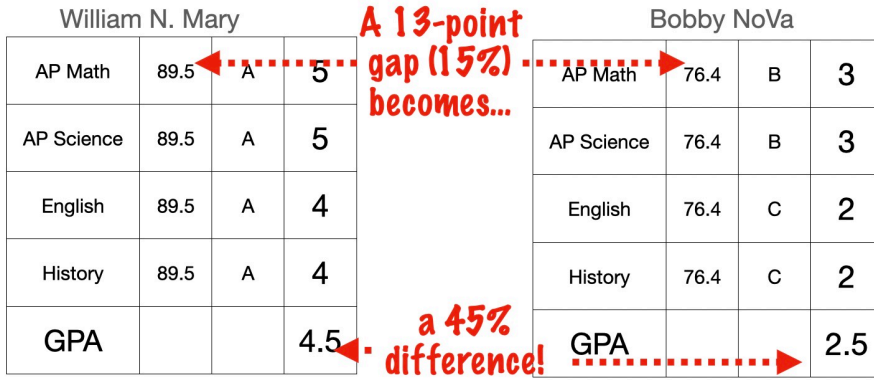


Figure 4.9 Transcript comparisons for 2 students in 2 AP classes

The quality premium narrows the gap between the two students' GPAs to a 45% difference, where it was 50% for regular classes. But if Bobby and William are in different classes, then the outcomes are much worse (below). A 50% gap becomes a 66% gap! In raw score terms, Bobby would have to average no more than a 40 in all of his classes to be this far below William.

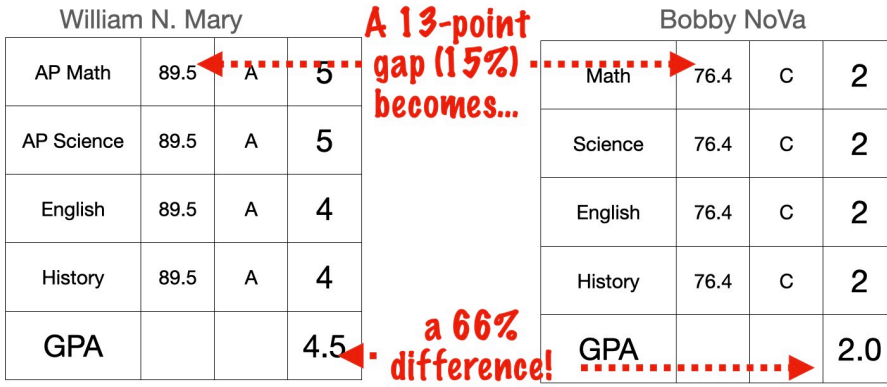


Figure 4.10 Transcript comparison for students in 2 and 0 AP classes

In the Introduction to this document, we saw that the graduation rate for white students is 96% and 89% for Black students. While that seems almost comparable, consider that around 75% of white students take courses with the AP/IB/DE premium, while only 26% of Black students do.[20] That suggests that Arlington's Black students are much less likely to attend college and less prepared when they do.

We can use data from VDOE School profiles for Arlington Public Schools to see the disparate outcomes in college enrollment.[21] The profile reports on graduation rates, types of diploma, and subsequent enrollment in Institutions of Higher Education (IHE). Because IHE enrollment is based only on students who graduate, we have to adjust by the percentage of students who graduate to determine how likely a disadvantaged student is to attend college. Those calculations are below:

	Graduate	Advanced	Drop-out	IHE	Adjusted IHE
All students	92.9	61.5	5.9	80	74.3
White	98.5	82.1	0.8	88	86.7
Black	91.9	45.3	5.7	80	73.5
Disabled	95.1	30.1	3.6	69	65.6
Econ. dis.	91.4	41.3	6.9	69	63.1
Hispanic	84.7	38	14	65	55.1
English learner	77.7	17.3	21.1	71	55.2

Figure 4.11 Outcome percentages by demographic, with adjusted IHE

20 Chan, M. "Black Parents of Arlington raises new concerns over APS performance gaps." ArlNow 6/23/2022, at https://www.arlnow.com/2022/06/23/black-parents-of-arlington-raises-new-concerns-over-aps-performance-gaps/ (12/1/2023).

21 https://schoolquality.virginia.gov/divisions/arlington-county-public-schools#desktopTabs-4 (6/6/2026)

We see a sharp difference especially in Advanced Diplomas. AP classes are not required for Advanced Diplomas (which are based on credit hours), but students who take AP classes typically complete Advanced Diplomas. We see that white students who graduate are about twice as likely to earn an advanced diploma as Black students. We also see that Black students are less likely to graduate and less likely to attend higher ed — 74% compared to 87%, per our adjusted rates. And these data do not reflect whether students enroll in 4-year or 2-year college programs, nor whether students are full-time or part-time in higher education.

For Disabled, Economically Disadvantaged, Hispanic, and English Learner students, the disparities are even more dire. The chart below helps us see the differences in IHE enrollment.

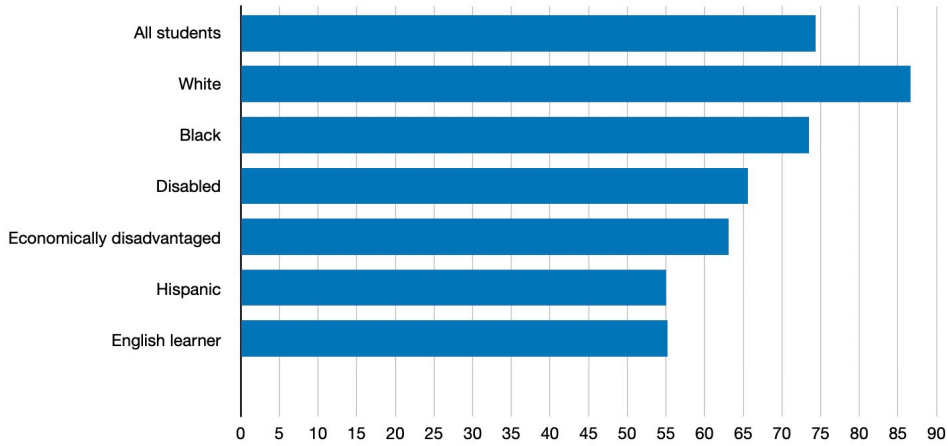


Figure 4.12 Difference by demographic in adjusted IHE enrollment

The AP/IB/DE system helps drive this outcome in three ways: first, our grading system likely amplifies and distorts the achievement differences between wealthy, white, and enabled

students versus marginalized students in their secondary years, leading to lower enrollment in AP/IB/DE classes. Second, lower AP/IB/DE enrollments create a structural bias against marginalized students for college admissions compared to students with those classes on their transcripts.

Finally, the AP/IB/DE premium penalizes marginalized students even further, reporting their progress as worse relative to white and wealthy peers. This is especially true for disabled students who take "Study Skills" or other special education classes, and English Learners who take ESL classes. Those classes are only reported on the 4-point scale, creating a structural limit on their total GPA. These classes are meant to help those students, but they hurt their GPAs. We are penalizing and punishing students who most need our help, so we can make our best students look a tiny bit better.

The AP/IB/DE premium system helps drive false and harmful comparisons of our students. Fortunately, there is an easy fix: instead of weighting points, weight AP/IB/DE classes for credit, as in the table below. Counting the premium as additional GPA credit will reward students for the extra workload, which is the real difference between regular and premium classes.

	Letter Grade	Equal Points	Extra Points	More Credit	
AP History	D	1	2	1 x 1.25	1.25
Physiology	B	3	3	3	3
Ceramics	A	4	4	4	4
DE Trig	B	3	4	3 x 1.25	3.75
French I	B	3	3	1	3
AP Lit	A	4	5	4 x 1.25	5
GPA		3	3.5		3.33

Figure 4.13 GPA comparison based on point vs credit premiums

Because all grades would be reported on a 4-point scale, and not a 5-point scale, the credit premium would not penalize students for taking regular classes and would not penalize classes required of students for their contingent status (like students with disabilities or English learners), while still rewarding students who did well in the premium classes. While it would require permission from the Virginia Board of Education to allow AP/IB/DE credit premiums to count towards graduation (and this is something we should explore), there is no obstacle to APS using credit premiums to calculate GPAs for students.

However, students in all classes would still be subject to the distortion, bias, and demotivation built into our grading scale. We will see how we can solve these problems soon, but first we need to look at one more problem our grading scale creates: excessive work for students and teachers.

V. WORK

In order for our grading system to function, students must actually do the work and teachers must grade it. Our grading system demands students and teachers do way too much work for little or no useful information.

- **Up to 60% of students' work gives us no useful information about their achievement.**
- **Teachers must grade up to 200% of assessments in order to justify failing a student.**
- **Our assessment and grading system discourages student achievement.**

In the last chapter, we considered a student's grade as a 'grade space': that is, a landscape of all the points a student could earn across all the assessments in a given term. Let's assume a teacher uses ten assessments with 20 questions each to arrive at a final grade for the quarter. We can represent that grade space as a grid of 200 squares, as in the figure below.

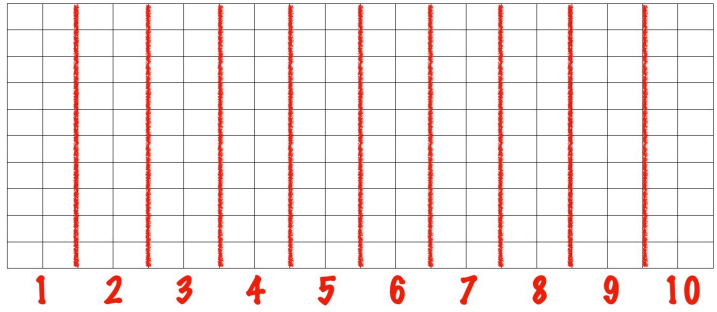


Figure 5.1 A 200-point grade space

In order for a student to pass, 60% of these squares must be 'correct' responses. In the easiest grading scenario, the result

looks like the table below (where green is 'correct' items, and white is wrong or incomplete items):

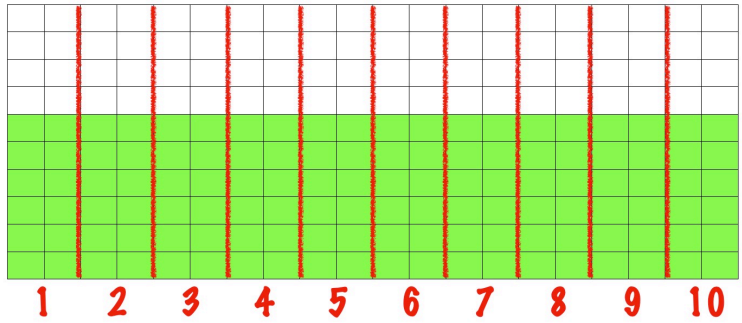


Figure 5.2 A 200-point grade space, consistent performance

But a student who does six assignments perfectly and then bails on the last four is also a passing grade, as below:

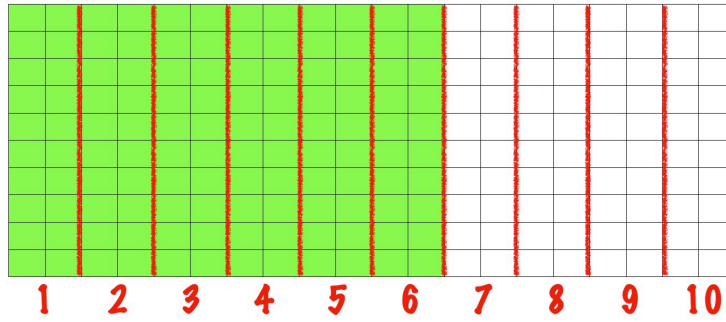


Figure 5.3 A 200-point grade space, 4 missed assessments

Both are the same grade, but usually these two are very different students. In practice, both students are rare. It is more likely that a student's performance from assignment to assignment is uneven, as on the next page:

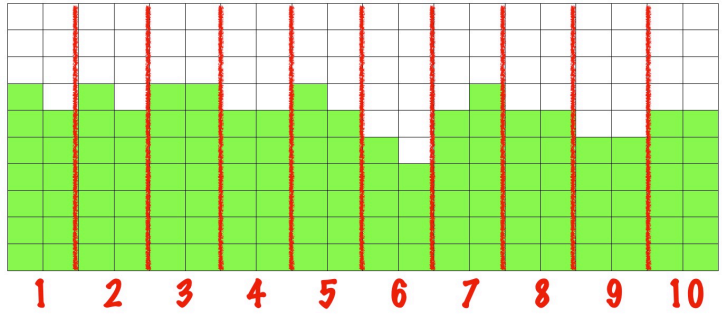


Figure 5.4 A 200-pt. grade space, varied performance

What grade is this? It takes a little bit of work to calculate the total grade, which is of course much easier with computers and grading software. And in fact, that student's performance will not be clustered neatly, but will look more like the figure below. That is, the teacher will have to cover the entire grade space.

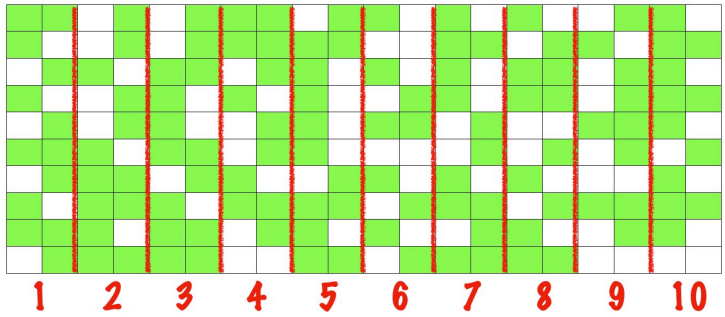


Figure 5.5 A 200-pt. grade space, highly varied performance

If a student turns in 100% of the work, the teacher has to do 100% of grading. Which seems fair — so far. Yet the student who turns in 6 perfect assignments and bails on four is doing the teacher a big favor: they slashed the teacher's grading by 40%!

What about our next student (next page)?

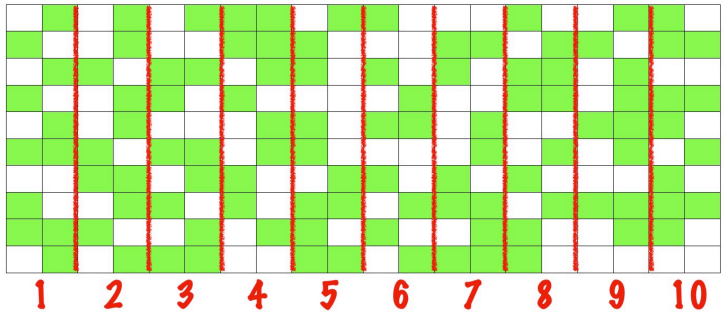


Figure 5.6 A 200-point grade space, low performance

They failed — which is bad for the student, but also means more work for the teacher. Fortunately, this teacher believes students deserve more than one chance, so they offer retakes. But retakes require the student and the teacher to do the same work twice. Below, the student has retaken every assessment, and the teacher has graded every assessment twice, and the student still fails.

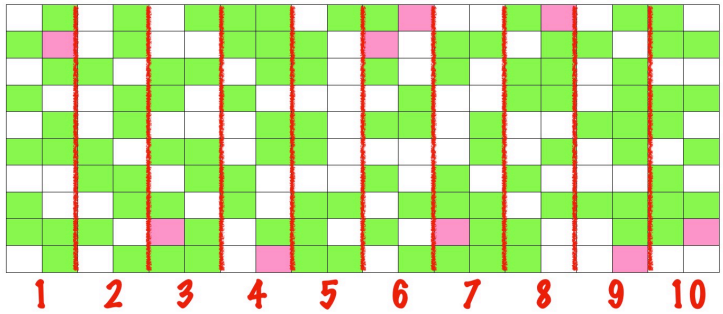


Figure 5.7 A 200-point grade space, with retakes

This dynamic represents a real obstacle to teachers' willingness to provide retake opportunities for students who are struggling. As a matter of equity and equitable pedagogy, we must provide those opportunities. As a matter of pragmatism, it

is a tremendous amount of work for both student and teacher. Low-scoring students create a particular burden on teachers to document students' deficits, when teachers could better put that time towards helping the student succeed.

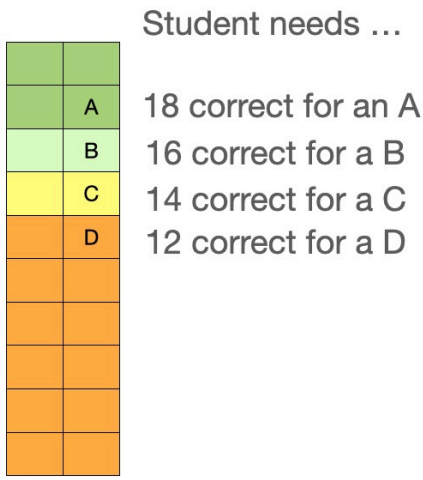


Figure 5.8 Gradation in one assessment

So far we have assumed that all items on all assessments in the 'grade space' are the same difficulty. However, that would be weak pedagogy and inconsistent with APS's own grading policies.

Instead, assessments should be designed to increase in difficulty — which we can call "**gradation**" — across the assignment. That gradation looks more or less like the chart to the left for a 20-item assessment.

Note that this chart shows the relative difficulty of the each item, not the student's grade. All 12 items necessary for a D should be D-level difficulty. The 13th item is C-level difficulty, but the student must get that and the 14th correct for a C. In theory, A items should be harder than B items, B items harder than C, and so on.

Across all assessments, the grade space would then look something like the table on the following page. Each assignment is gradated, so the grade space as a whole reflects the different cognitive tasks explicit in APS grading policy.

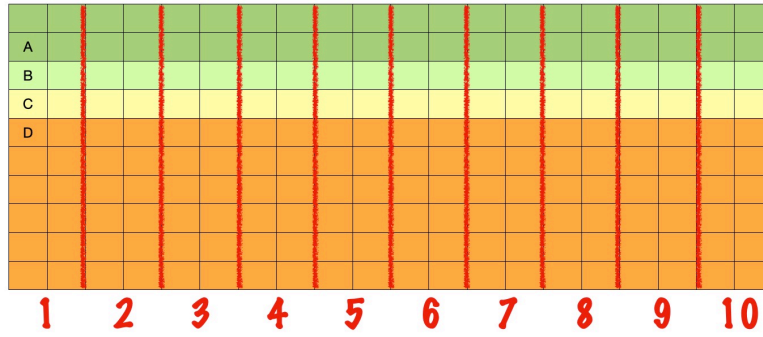


Figure 5.9 Gradation across the grade space

While this seems like a good idea in theory, in practice this approach tends to **discourage achievement**. For starters, a student who gets a perfect score on the first nine assignments knows that they do not need to turn in the last assignment to receive an A, as we saw before. Sometimes, teachers even make this explicit by excusing students from taking final exams.

A more consequential problem is that if knowledge is cumulative across a quarter, it is much easier for a student to go down in their performance, per this chart:

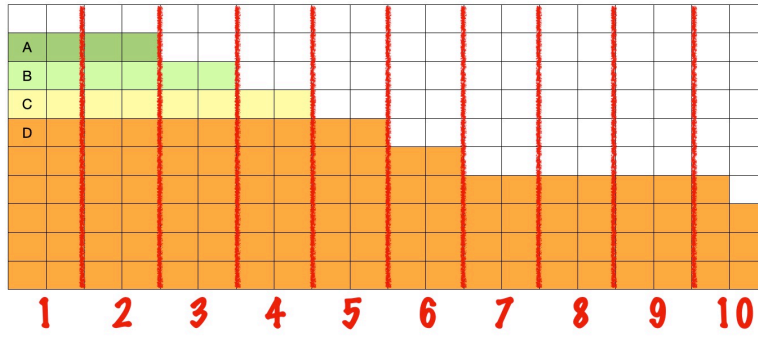


Figure 5.10 Descending a grade space

... than up, per this chart. Progress is harder than regress.

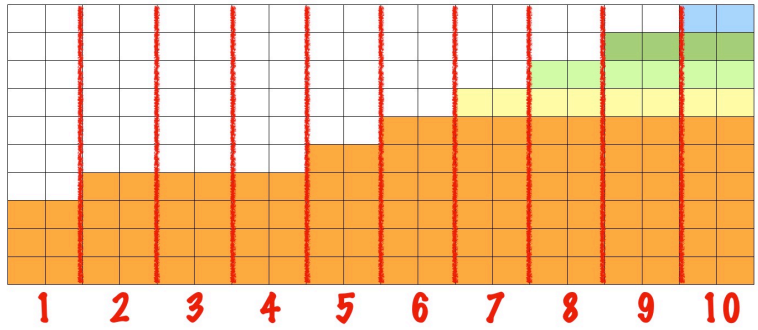


Figure 5.11 Ascending a grade space

And even though the student in the second chart has shown extraordinary progress, they still fail the quarter.

The difference between these two trajectories comes down to whether missed items in early assessments lead to less chance of success in later assessments — which is very typical — versus assessments in which correct items point to increased likelihood of success later. It may feel like there is no difference, but in pedagogical practice and assessment design there often is. If a student does poorly early in the term, they are often deeply discouraged from any effort to improve their mastery.

Both previous charts amount to nothing — 0 — for the students' transcript grade, which means a tremendous amount of no-value work for both teacher and student. That is also the case even if the student passes. Recall that earlier we saw a 30-question test had excessive sensitivity, especially for students with low grades. We can apply the same principle across the grade space: our imagined grade space has far more sensitivity than we need to accurately record a student's grade. Our students are doing too many graded assessments, and teachers

are doing too much grading. We could cut the assessment (and grading) load in half and still get reliable information that accurately reflected student progress.

More than that, most of the information in the grade space is essentially zero value. Consider that a student only needs a 90 to get an A (=4.0). That means that any score higher than a 90 does not give us usable information about that student. So we can remove 20 items (181-200) from our grade space model.

The same is true for most of the fail range: because there is no final difference between a 59 E and a 0 E (ignoring 0=50 for now), all of the grade space that result in E grades amounts to a single data point in the end.

This means, in theory and practice, all the usable information in our grade space covers just 61 items out of 200. The chart below shows the usable grade space:

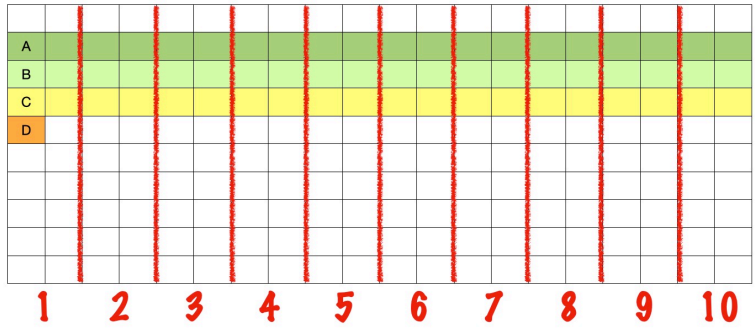


Figure 5.12 Portion of grade space with useful information

This means that a tremendous portion of student work and teacher work is literally point-less, in terms of the information it contains on students' learning. Nearly 70% of students' work and teachers' grading gives us no useful information! And that is without factoring in retakes. Students and teachers are wasting massive amounts of time on no-value work.

The good news is that if teachers can justify the implicit gradation in their assignments as shown above, they can justify ***any*** system of gradation that makes sense according to the curriculum and the standards. As we have seen, *any* grade cut-offs are arbitrary, so a teacher might well adjust those cut-offs from assignment to assignment, varying assessments in difficulty across the grade space. The chart below shows what that could look like. It looks messy, but it is much more defensible than curving, and probably reflects a more deliberate and proactive approach to designing assessments:

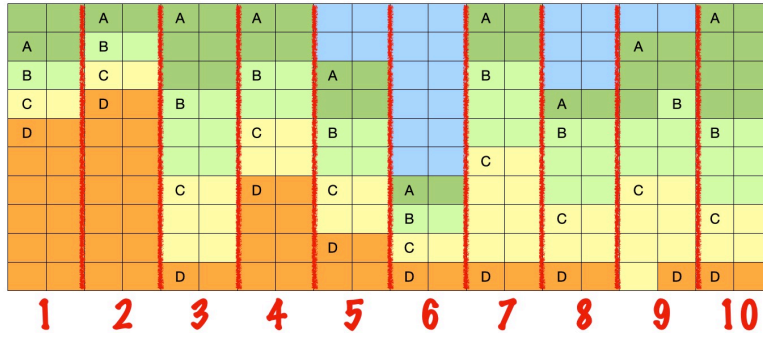


Figure 5.13 A grade space with varying difficulty assessments

Gradation is a good thing; it is an important skill that all teachers should develop. Most teachers already use it implicitly. As we saw in the chapter on measurement, we already trust teachers to develop their assessments correctly. But the tools we give teachers — the 100-point scale and its inconsistent intervals — distort those measurements. Rather than making teachers' lives easier, it drives demands for more assessments and more grades, as if more grades are a substitute for accurate grades. In the next section, we will see how explicit gradation can help us measure, report, and compare grades more accurately and much more fairly.

VI. GRADATING

Using explicit gradation based on Bloom's outcomes offers us a grading system that measures accurately, reports reliable information about student achievement, and allows for meaningful comparisons. Among other benefits:

- **Students can focus on work at their level.**
- **Curving becomes unnecessary and irrelevant.**
- **We solve the problem that 0=50 only masks.**
- **Teachers do a lot less grading.**

So far we have looked at the problems with our grading system; in this section, we turn to a solution that uses assessment design to ensure accurate and unbiased grades.

Here we need to revisit Bloom's taxonomy:

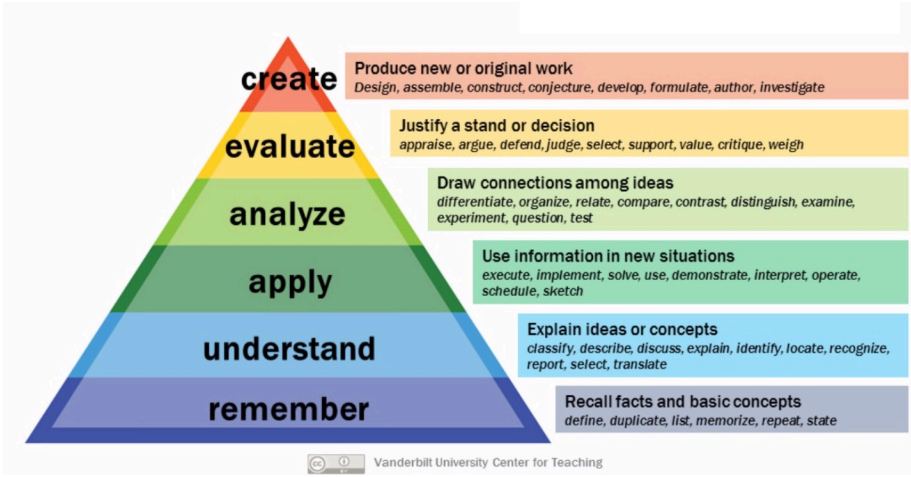


Figure 6.1 Bloom's taxonomy of educational outcomes

The last chapter introduced the concept of 'gradation' — that our grades reflect varying difficulty in the cognitive tasks our assessments measure. Again, Bloom's taxonomy provides

the best rationale for gradation of assessments. In fact, APS Policy Implementation Procedure I-7.2.3.34 PIP-2[22] uses Bloom-like language such as 'creatively' and 'applies' in its descriptors.

For our purposes, we can simplify Bloom's taxonomy to describe gradations in our assessments as follows:

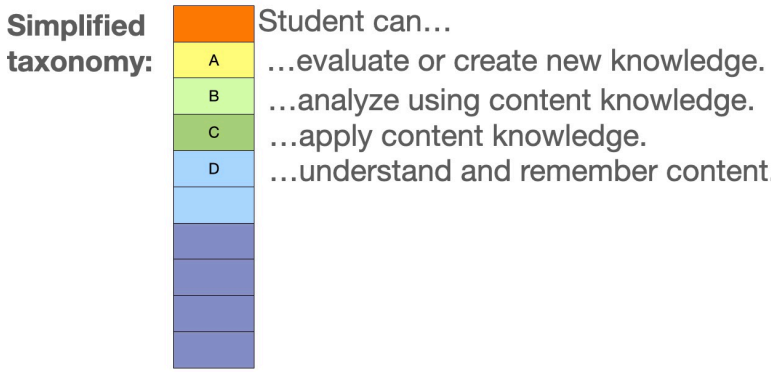


Figure 6.2 Bloom's taxonomy in one assessment

For a test on Congress, the gradation might look like this:

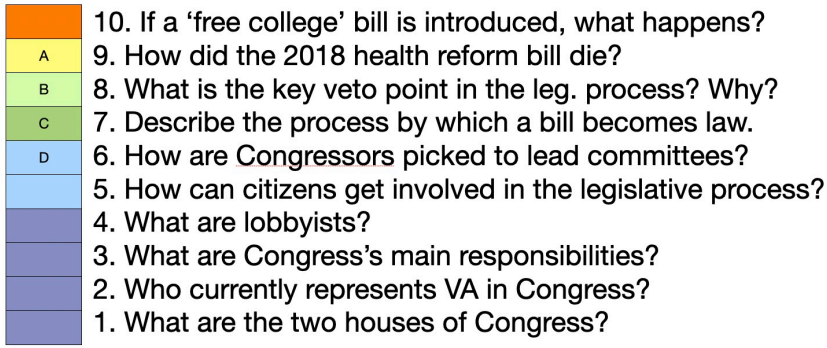


Figure 6.3 A civics test gradated according to Bloom's taxonomy

---

22 At https://shorturl.at/kAUY1 (12/1/2023).

Another way to implement the test in Figure 6.3 might be to have multiple choice questions for the D-level work, and short answer and essay questions for the higher-tier work. But each tier should be clearly identified, so that students know what is expected of them and can calibrate their effort. Teachers can then grade the test bottom to top, and stop when the student no longer demonstrates competency at the higher-tier tasks.

As we saw, our grading system implies that every assessment has gradation built into it, though this is often not the case in practice. The kinds of problem sets assigned for math homework or the reading notes for a social studies class will rarely rise to the 'apply' level of educational outcome. Many teachers already grade these low-tier assessments on a rough pass/fail scale, basically checking off that the student successfully completed the work.

In any case, if every assessment is gradated the same, our grade space looks like this:

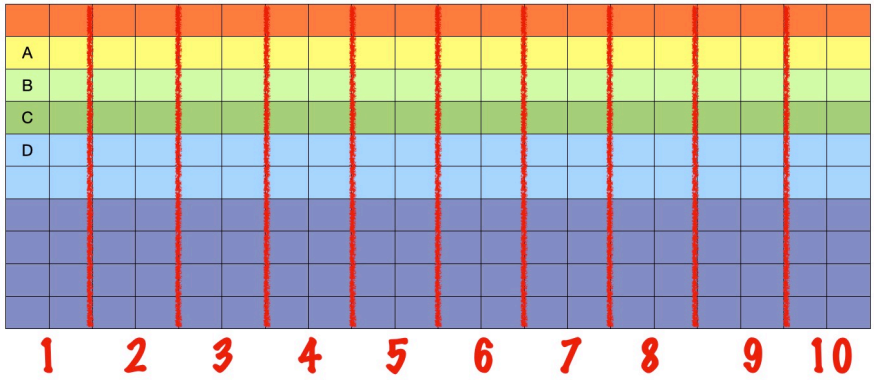


Figure 6.4 A grade space gradated according to Bloom's taxonomy

In every one of these assessments, students have the opportunity to demonstrate that they remember the content and understand it (D), that they can apply it (C), that they can analyze it (B), and finally to evaluate and create with it (A).

Gradating each assessment is a lot of work, but the logic of the grade space means we can also gradate by assessment. That is, we can design assessments that measure a single level in Bloom's taxonomy. In the grade space below, students will complete three assessments that focus on recall of content, three assessments on whether they understand it, one assessment on application, one assessment on analysis, one assessment on evaluation, and one assessment that requires students to create new knowledge. Among other benefits, this approach allows for *much* simpler rubrics in grading.

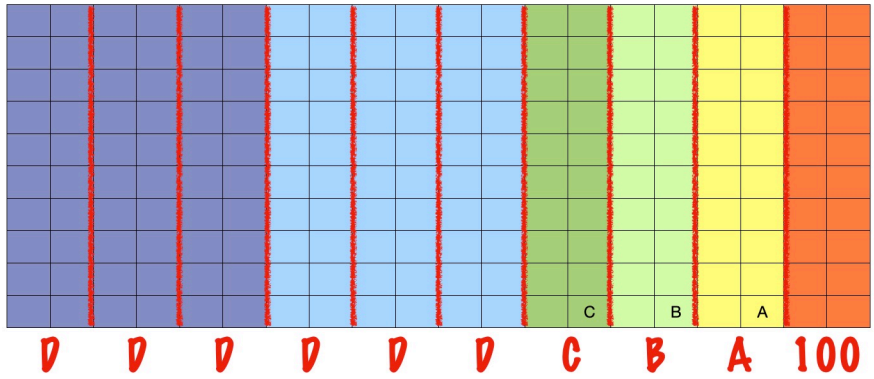


Figure 6.5 Gradation per assessment

It is important to understand that figure 6.4 and figure 6.5 have *exactly* the same validity in measurement terms, but the per-assessment gradation in figure 6.5 is much easier to grade. It is also better pedagogy.

A significant advantage of gradating by assessment is that we **focus** struggling students on only the work they need to pass. If a student fails lower-tier assessments, they should not attempt higher-tier work. Consider a student who performs as follows across the grade space, and ends up with a 49% (below). If they turn in perfect work on the next higher assessment, their grade will still be 59% — not passing. So this student

should focus their energy on D-level work. This means the student has to do less repeat work and the teacher does far less repeat grading.

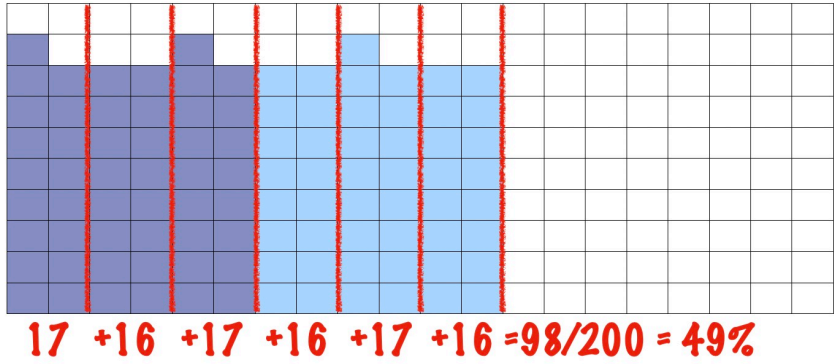


Figure 6.6 A student who should revisit lower-tier outcomes

Students should attempt higher-level work only when they have completed the lower assessments. A student (below) who has not turned in two assessments should complete those

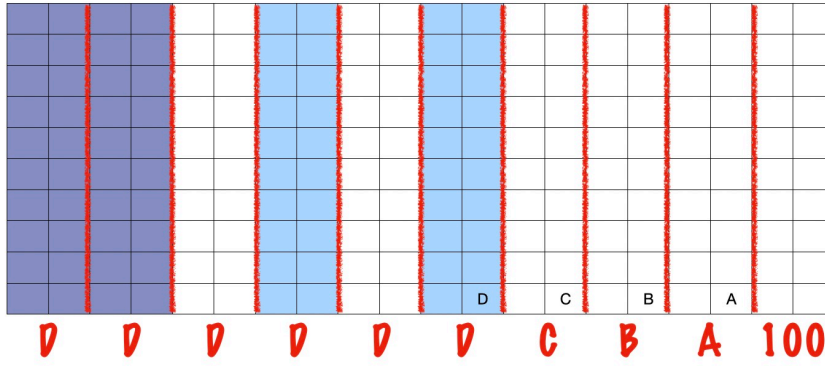


Figure Six-Seven. A student who should complete lower-order assessments.

before they attempt the higher level. Note that this student could then miss four assessments and still pass the class, which is a much more reasonable solution to the **zero problem** than 0=50. It also provides students a clearer path from a failing grade to passing, and a clearer sense of the expectations for higher grades.

It is important to see that **the same amount of teaching and learning is happening** in the classroom as before — if not more, because of the teacher's reduced grading load and the decrease in redundant retakes. There is only less *assessment* taking place, only because the teacher already has clear evidence the student is unlikely to succeed at higher levels.

For students who are reasonably complete in their work but still need an extra 'bump' to get a higher grade, the next-level assessment is in effect extra credit, as below, even if they do not attain the next level outcome. This encourages students — even low-performing students — to reach beyond their comfort zone and attempt more difficult work. That is, it encourages achievement, which our current system discourages.

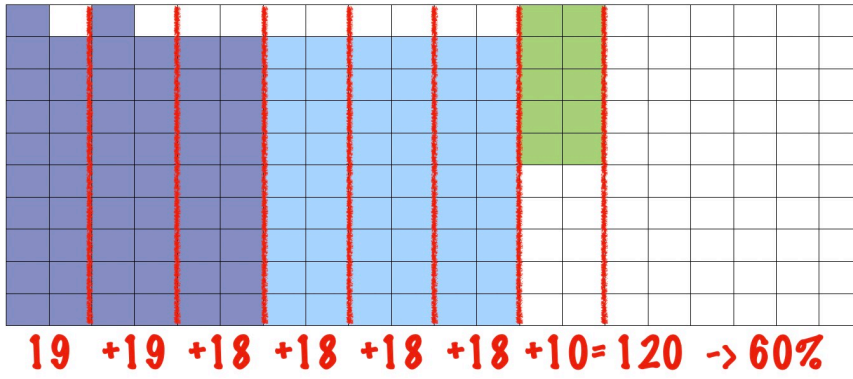


Figure 6.8 A students who attempts the next-order assessment

One of the advantages of this approach to assessments is that it **does not allow curving grades**. It makes no sense to report that a student 'analyzed' where they only 'explained', simply because much of the class had difficulty 'remembering'. In a properly gradated system, outcomes are not fungible; different assessments give discrete information about students' mastery of content.

In order for this system to work, the gradation that is usually implicit in assessments and our grading scale has to be explicit per assessment. The rationale for each gradation should be based on the teacher's command of the standards and curriculum. The ordinal nature of the scale mitigates the need for formal validation that quantitative, ratio measurements usually need. This affords the teacher far more flexibility in design of assessments for a given standard, unit, or term. In fact, a teacher proficient in their content area can justify gradation mapped to *any* point scale. The chart below shows assessments gradated to the 4-point scale.

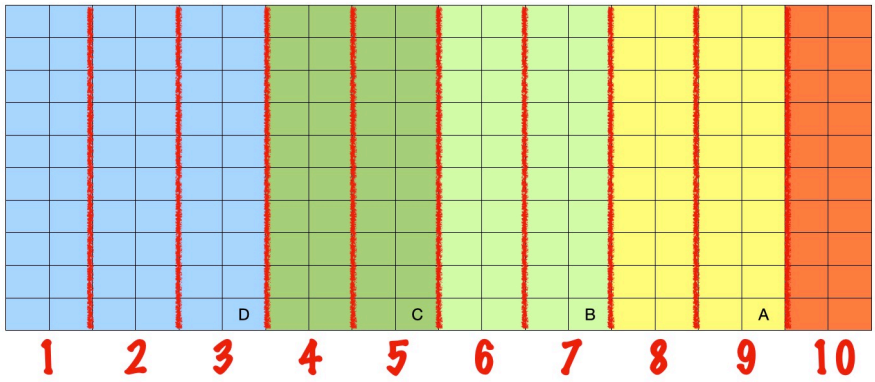


Figure 6.9 A grade space gradated to the 4-pt. scale

In this new gradation, because each assessment has an explicit rationale for the outcome difficulty, they cannot be mapped to the 100-pt. scale as a matter of math and

measurement. As a result, this teacher can justify passing a student even if they miss seven assignments, as below:

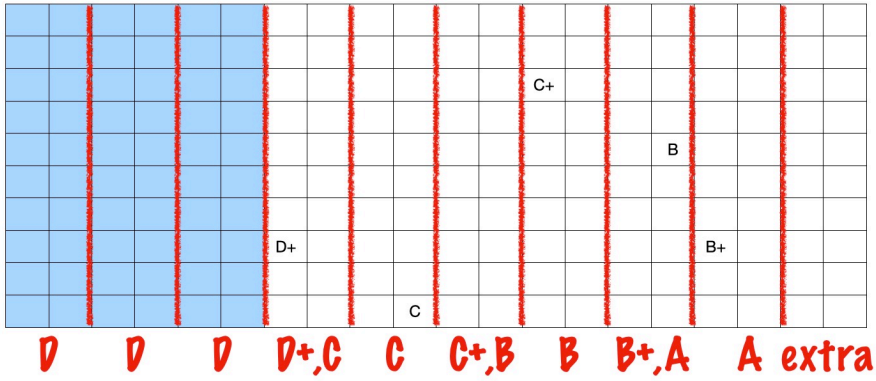


Figure 6.10 This student passes!

We saw that on the 100-pt scale, most of the work — about 70% — that teachers and students do is literally point-less. The grade space in figure 6.10 means we have cut the point-less workload by more than half, and made the rest of the grade space far more valuable as measurement.

Again, the *same amount of teaching and the same amount of learning* — if not more — occurs in this class as with traditional grades. Only the assessment load is different. (If a teacher is unable to articulate a rationale for their gradation, whether implicit or explicit, they may require training in order to be effective in designing and grading assessments. This would also be the case for teachers creating traditional assessments.)

One way to think of this grade space is that the teacher has deleted the easiest 40 items — the 'recall' items — from their grade space, and added 40 items across higher levels of outcome. This larger 'pass window' gives struggling students a reasonable chance at passing, while ensuring that the class remains challenging and engaging for top students.

And though it seems like this approach skips a lot of foundational work at the lower levels, there is a significant body of research to support the idea that higher-order assessments better measure student achievement across all levels of Bloom's taxonomy.[23] The massive amount of low-tier work that most assessments demand of students is probably unnecessary, if not counter-productive.

For our unit on the legislative branch, the resulting assessments might look something like the list below:

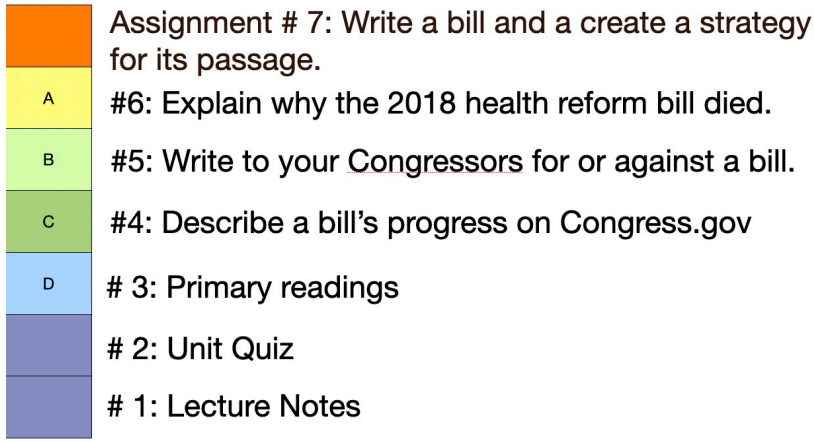


Figure 6.11 Gradated assessments for a unit on the legislative branch

23 For a recent article that discusses and extends this research, see Agarwal, Pooja (2019). "Retrieval Practice & Bloom's Taxonomy: Do Students Need Fact Knowledge Before Higher Order Learning?" *Journal of Educational Psychology (*111:2), p. 189-209 at https://www.researchgate.net/publication/325639446_Retrieval_Practice_Bloom's_Taxonomy_Do_Students_Need_Fact_Knowledge_Before_Higher_Order_Learning/link/5c9ab6bf299bf1116949990d/download

Of these assessments, #7 is the most difficult — and in fact, it is way too difficult for middle school Civics & Economics students to complete. But they do not need a perfect grade on the seventh assessment if they have already done well on all the previous assessments. The highest gradated assessment should be the most challenging, and it should go beyond even the best students' comfort zone. That cannot happen with our current system, because of the narrow pass window.

The lower (D-level) assessments in this unit are typical for social studies classes. Some of them, like Lecture Notes, offer little cognitive challenge. In fact, lecture notes do not even ask students to remember or recall, the very bottom of Bloom's taxonomy. We need to recognize that it is meaningless to give a student an 'A' on a task like this. Even 'Describe a bill's progress on Congress.gov' does not offer students an opportunity to 'creatively apply and extend knowledge and skills', per APS's grade descriptors. It makes little sense to award these sorts of assessments an A, when they do not involve the desired level of outcome. We should not pretend that every task has the potential to be A-level work. Even as teachers, we do not use our higher-order cognitive abilities in every task we complete (certainly not most grading!). Outcome gradation better reflects how people actually apply their cognitive skills in real life.

And because educational outcomes are qualitatively different tasks in cognitive terms, the different gradations are better suited to the ordinal letter-grade scale. We can in fact cast these as distinct and discrete assessments, rather than occurring along a specific interval scale. We can justify whatever difference makes sense from assessment to assessment in terms of the content and the appropriate level of gradation — rather than anchoring our scale to a specific ratio. This affords teachers far more flexibility in assessment design than measuring grades according to fixed ratios. Figure 6.12 on the next page shows the assessments as discrete tasks.

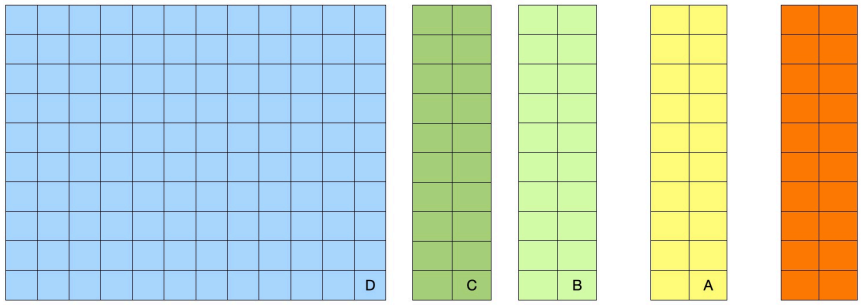


Figure 6.12 Gradation as discrete tasks

For a real-world example, I used this approach for a unit on the Judicial branch for middle school Civics. The main assignment was the unit quiz, which was C-level work — that is, a lot of knowledge application. The highest grade any student could get on the quiz was a C. The quiz included remembering and understanding, because a person cannot apply knowledge they do not remember and understand. The questions were based on formative assessments, so the students had already practiced each task; students who did poorly were allowed to retake specific sections of the quiz as needed.

I then gave students a choice. First, they could research and write a short, structured analysis of an important Supreme Court decision. This was B-level work, and I made clear to the students that a B was the highest grade they could get.

Or, the students could participate in a mock appellate court, and potentially earn an A. Throughout the school year, students had been immersed in a government simulation that included proposing and passing laws for their country. These governments were based on the U.S. Constitution, so I created a series of cases in which laws the students had passed were struck down by Federal courts. Students then paired up to research similar cases, draft briefs, and argue opposing sides in a mock appellate court hearing. I even borrowed a black robe and made a gavel from scrap wood. This assignment incorporated B-level work in appraisal and analysis of case law supporting their arguments, but it also reached farther to extend and create new knowledge. The qualitative nature of the assignment meant I

could set the bar for a 'perfect' score very high: there was in effect no 100% possible.

Giving students this choice worked extremely well. First, I was surprised by how many students picked the A-level assignment. Several students I had thought were relatively indifferent put serious effort into the A-work (though not all earned an A, in the end). Not only that, but most of the students who did the B-work, including students who were genuinely indifferent to the class, ended up doing a very credible job on the assignment. I worried we might leave some students behind, but the results were opposite: students ran to catch up.

The least surprising outcome was how deeply invested top students were in the A-work. They threw themselves into the research process, with some producing near-college level work (again, these were middle school students). As their teacher, I am quite confident these students understand how courts work at a level I could not determine from a multiple choice test or even an essay assignment. And note that there was no final exam for the unit, only the quiz.

We should note that while gradating is equitable because it benefits marginalized and disadvantaged students, it also has significant benefits for overachievers. When A-level work is limited to discrete assignments that are only a fraction of the work load, top students no longer have to stress out trying to get perfect marks on every assignment. They can hold back some of that energy and invest it where it matters, in the A-level work — the way some of my middle schoolers skated through the quiz but absolutely crushed the appellate case assignment.

And one implication of this approach is that there is no intrinsic reason why students need to accomplish all levels of mastery within a single year. Virginia's Standards of Learning certainly expect students will complete each course in one school year (or less), but in theory students should not need to repeat the whole year to demonstrate appropriate mastery. Instead, they could simply return to the assessments they

missed. That is true not only for students who fail, but for students who fall short of an A. When mastery is gradated into discrete cognitive outcomes, it is entirely reasonable, from a developmental perspective, to allow some students to develop cognitive skills later than others.

Again, this system discourages curving grades (or lining them) within a given assessment, and it all but bars the practice across the unit or term. Students' grades in this unit reflect discrete differences in educational outcomes, rather than two-decimal-point differences in recall.

This system can also discourage cheating. For example, it is rarely the case that the problem sets typical of math class homework rise to the level of "creatively applies". A teacher using outcome gradation could peg all homework at the C and D levels — "apply" and "understand", respectively. Tests and in-class projects could then serve as B and A level work — "evaluate" and "create" respectively. Students would face much more difficulty in cheating their way to top grades. In the legislative branch unit plan above, the higher level assessments especially can be easily structured to promote original work.

For teachers, the main benefit of gradation is that it requires **a lot less grading**. Gradation does require more forethought in assessment design, but on the back end it could cut teacher's grading workload by more than half. While that may seem optimistic, recall that in our current system most of the work students and teachers do ultimately provides no useful information about student achievement.

Our system also encourages low-motivation students to do shoddy work across all assessments, leaving teachers to grade vast amounts of low-quality work. The gradated approach requires those students to do quality work for at least a few assessments. Meanwhile, outcome gradation gives students more agency and ownership of their education in terms of choosing or refusing to do work based on their interest, energy, time, and skill level.

The process for recording grades for outcome gradation is also more straightforward. The table below shows how a portion of the grade book might look:

	D work	D work 2	C work	B work	A work
Adam	✓	✓	✓	p	
Beth	✓				
Carl	✓	✓	✓		
Dinah	✓	✓	✓	✓	✓

Figure 6.13 Grade book portion for outcome gradation

In this unit, Adam has successfully completed the D and C-level assessments, but has only partially completed (that's the "p") the B-level assessment. His final grade for the unit would be a C+. For a higher grade, he should reattempt the B-level work, instead of trying for the A-level or asking for extra credit. Meanwhile, Beth is in danger of failing: she should complete the second D-level assessment. Carl has a solid C, and Dinah has completed all the work and so earns an A.

While outcome gradation produces ordinal data, it is inevitable that our high schools will report grades and calculate GPAs on a ratio scale — the 4-point scale — simply because that is what colleges and universities expect. There is no sound methodology to make the math work: it is meaningless to report that 'creatively extends' and 'applies' are one point apart. But because we have abandoned the 100-point scale, we will no longer see the gross distortion in translating ratio to interval to ratio. Not only does outcome gradating help focus students who are struggling, it also does not penalize them simply for being struggling students.

As described above, this is not a purely theoretical proposal. I have used outcome gradation in the classes I teach, and it works. My grading load is lessened, I can focus struggling students on the work they need to pass, yet still provide top students with challenging work. Students appreciate the ability to calibrate their effort to specific outcomes. I can explain to students and parents very clearly what a given student's grade in my class means and how they can improve.

We saw that gradation disallows curving: the idea that a student ascends to higher-order cognition just because the rest of the class disappointed the teacher's expectations is nonsense. In the same way, gradation disallows the AP/IB/DE premium: it makes no sense to report a student performed at a higher cognitive level than they did, simply because the class is more work. However, it is entirely reasonable that AP/IB/DE classes require more A-level and B-level work from students, and offer less D-level work. Credit weighting is a much more reasonable reward for students' effort in these classes, and it does not penalize students in regular classes.

By using a system of outcome gradation, we make our goals and expectations for students much clearer. We provide students, teachers, parents, and other schools with a much more solid rationale for the grades we report. We create valid comparisons across students, among teachers, and between schools. We can allow students to focus on work at their level, while encouraging them to aim higher.

More than that, outcome gradation can serve as the keystone to a far more equitable approach to grading. Our grading system denies students equity in their choices during their career with APS, and even more so once they graduate. Outcome gradation with credit weighting for AP/IB/DE classes will go a long way towards solving these problems.Outcome gradation will not solve all our problems with respect to equity, but it will make it easier to detect, measure, and address other equity problems across the district.

VII. BETTER GRADES

Arlington can drastically improve our grading system and take a big step towards equity through outcome gradation.

As we have seen, Arlington Public Schools' grading system is an algorithm that drives inequity — not just for marginalized students, but for the vast majority of students. Our system fails each of the three purposes grades serve in our schools: it gives us inaccurate measurements, reports biased outcomes, and creates invidious comparisons. Meanwhile, this system demands a sizable amount of meaningless work from both students and teachers.

Fortunately, there is a straightforward solution, which I call "outcome gradation": instead of grading assessments as if they cover *every possible* cognitive outcome (as implied by our grading system now), we can design assessments pegged to a *specific* cognitive outcome per Bloom's taxonomy. We can assign those outcomes to specific letter grades, as suggested by APS's current grading scale. The resulting system focuses struggling students' attention on the work they need to pass, while encouraging them to push up to the next level of outcome — where our current system is unfocused and deeply discouraging.

Outcome gradation means rejecting the 100-point scale and the distortion and bias it creates. We have seen how nearly every student loses percentage points in the transition from the 100-point scale to the 4-point scale. Outcome gradation also means ending the AP/IB/DE premium point, which gives the most advantage to the worst students in those classes. Instead, premium classes should be weighted for credit, which helps the best students more than it helps the worst, without penalizing students in regular classes, Special Education classes, or English-learner classes.

The resulting grades will measure more accurately, report with far less bias, and provide fair comparisons — in stark contrast to our current system. This approach builds on the professional skills and expertise of our teachers to create more meaningful work and less point-less work for everyone.

With that in mind, I propose three specific change to give Arlington Public Schools better grades:

1. **Use outcome gradation pegged to Bloom's taxonomy.**
When assessments and grades are based on explicit gradation, students have a clearer understanding of the steps to mastery. Teachers have less work grading and can better focus struggling students on what they need to pass. Grades are then more meaningful to students, parents, and schools.

2. **Reject the 100-point scale.**
The transition from 100-point to letter to 4-point is disastrous for low-performing students, leading to invidious comparisons from student to student, with serious harm to their life prospects after school. We should stop using the 100-point scale to record raw grades. The 4-point scale is only necessary to report GPAs.

3. **Weight AP/IB/DE classes by credit, not points.**
The point premium for AP/IB/DE classes penalizes any student enrolled in a regular class, and makes no sense with respect to educational outcomes. Weighting for credit rewards students for taking high-level classes in proportion to their effort, without penalizing peers in regular, special education, or English-learner classes.