# Part Three: How to Motivate a Teenager

In the first two parts of this paper, we saw an example of a grading structure that results in high and equitable student achievement. Now it is time to explain the theoretical basis for why this set of grading policies works. This is a longer section of the paper, because I will go in depth into both educational as well as cognitive psychological research studies and will critically compare peer-reviewed scientific findings to the actual lived experience of students within my classroom. The short version is that the policies in place in my classroom have two beneficial effects: 1) they are practices which help to develop students’ self-efficacy, and 2) they are practices which measure major steps of the known processes by which our brains progressively learn new material. Both of these effects result in increased student achievement, while the second one specifically helps the grades’ accuracy as measured by such things as alignment to AP exam scores.

Basically, if studies say that Steps X, Y, and Z are necessary to help a student learn the subject, I assign some graded measure to each of those steps. Students then earn a good grade by taking steps that directly increase their achievement in the subject. The steps start small/simple, and grow to comprehensive, so that the students gain confidence in their abilities as they learn. At its core, this really is that simple. Now, here are the details.

## Bonesrønning and Achievement

I am generally skeptical of much of what counts at present for education research, for reasons that will become apparent later. Occasionally a good study leaks through a flawed system, though, and we get actionable knowledge out of it. Such is the nature of a Norwegian research study from 2004 by Hans Bonesrønning. The paper is titled, “Do the Teachers’ Grading Practices Affect Student Achievement?” At the outset, I will say that although I believe the paper was well studied and reported, I disagree slightly with the conclusion the author reaches.

The study subjects high achieving and low achieving students to two types of grading schemes: one that is harsh and one that is lenient. For example, the harsh system might assign a 3-point grade penalty instead of a 1-point penalty for the same mistake on an assignment. Later tests show both types of students benefit more from harsh grading than from lenient grading, but the high achievers benefit disproportionately. The author therefore concludes that teachers should be harsh graders, because students will be motivated to work and learn only if the grading is harsh enough that the student risks losing their apparent status as a high achiever (or similarly, risking not getting into a good college) if they do not work and study. The exact words from his conclusion are as follows:

“The second part of the analysis shows that all students, and high-achieving students in particular, perform better when exposed to hard grading. The strong message from this analysis is that effective teachers are characterized by being able to manipulate student effort.”

For the record, I strongly dislike the use of words like, “manipulate,” when discussing students. On the rest of the substance of his argument, in my opinion, Dr. Bonesrønning is half correct. I will agree that if you tell a student who sees themselves as a B student that if they do not do an assignment, they will still get a B, the student will tend not to do it, even if doing the assignment could increase their grade to an A. I also agree that a student who sees themselves as an A student will be much more likely to do the assignment that raises their grade from a B to an A. This could explain why being a harsh grader seems to help the higher-achieving students more.

What I think the researcher misses is the potentially flexible nature of the student’s self-identity. His study groups students into low- or high-achieving based on previous academic performance. That is fair enough, but it treats each group as if they are static. Is there any way to encourage a low-achieving student to become a high achiever? Might a high-achieving student become a low achiever if the wrong conditions are present? And to what extent can a grading policy alone be used (or misused) to accomplish either outcome?

## How a Student’s Self Perception Influences Their Motivation

Students enter our high school classrooms with several types of preconceived notions and expectations for their own performance. Sometimes these are grade-centric, as in the case of someone who self-identifies as a C student. This is far from complete, though. We have students who see themselves as Math/Science students, who generally will do only what is necessary for their History classes but are intrinsically motivated in, for example, a Physics class. There are students who see themselves as artistic, musical, or theater students, who see Physics class as a thing they have to complete rather than a course they really care about. There are students who see themselves as Honors students, who will take high-level courses no matter the subject because they see themselves as belonging at a higher level. There are students who think school is “not for me,” and thus view themselves as non-academic. And of course, there are more identities and intersections between identities to explore. A student can see themselves as a nerd, extrovert, leader, and much more. Also, none of this even touches a student’s perceptions of their parents’ expectations or of external entities like colleges, whether or not those perceptions are based in reality.

In the absence of action by teachers and other people in the student’s life, these students’ views and perceptions of expectations are liable to become self-fulfilling prophecies. In some cases, as the aforementioned paper points out, this might result in a student trying harder than they normally would in order to protect their identity, so to speak. In some cases, that can be a good thing. Do note the “in some cases” though. I hold that this is far from a universal truth.

As to manipulation, any high school teacher worth their salt can design assignments and assessments that are so easy and lacking in rigor that almost every student will automatically get an A. Curves and adjustments can also be made to accomplish this feat. Likewise, every high school teacher who knows their subject could, in theory, make assignments so unreasonably hard that most students are guaranteed to fail—either due to the difficulty of the questions and tasks themselves, or due to time constraints, say. Curves and adjustments could be made to suit this purpose, too. I take it as self-evident that neither of these approaches constitutes good educational practice.

Instead of thinking of grades as a tool for manipulating student performance, I believe the best practice is to think of ways to help encourage and develop students’ perceptions of their own abilities and expectations of their own academic performance. How can we design assignments so that the feedback (including the grade) result in a student who thinks of herself as a B student start to believe she can and should try for the A? Can the same system help change the mind of a student who starts the school year genuinely not caring whether they pass or fail?

Research from various sources suggests that an answer does exist, and graded feedback is a crucial piece of the puzzle. We will delve into that research, but before we go about answering how best to communicate feedback with a grade, let us take stock of what messages we wish to convey with a grade in the first place.

## What Messages Should Grading Policy Communicate?

A good grading policy, and the assignments that are graded within it, should primarily aim to maximize the amount that students learn in a class. It is crucial to gain each student’s partnership in this endeavor. To that end, the grading system should be a communication tool that specifically communicates the following:

* Success is possible, and the teacher wants the student to succeed.
* If the student’s idea of success is an A, a 5 on an AP Exam, or any other measure of highest possible achievement, then the message needs to be, **“It will not be easy, but you can do it.”**
* If the student’s idea of success is a B, a C, or another middle-performance standard, then the goal should be to **develop the student’s identity so that they start to identify as someone who can get (and should try for) something higher.** This means that the A or, at least, the next rung on the ladder has to be in reach for them. They have to be encouraged to expect that higher achievement level of themself.
* If the student does not see themself as belonging in the class or in school to begin with**, they must be made to feel welcome, cared for, and capable.** This means there must be at least some form of assignment on which even the lowest student in that class can easily earn an A. Once that is achieved, one may attempt to build up the student’s sense of self and achievement from there. This is what I will refer to later on as a foothold or gaining buy-in.
* **All necessary work should be directly relevant to the subject,** so that the student’s increased perception of themself as academically successful in the subject directly aligns with the their actual skills for the next academic level and/or potential employment in a related field.
* All success on assignments, no matter whether it is at the level of lowest foothold or the highest possible achievement on a test, must be acknowledged by the teacher. This means it must be worth something in the gradebook. Recognition without the grade risks garnering cynicism and sarcasm from students; it is at best a mixed message for the teacher to say, “Great job,” all the while not giving the student actual credit for it. Recognition via the gradebook is necessary to help develop the student as a learner.[[1]](#footnote-1)

This building up of student identities does not happen overnight, and it does not always happen within a single school year. I have seen all sorts of variations—a student who is suddenly motivated and believes they can succeed because they see their *friend* get an A, for example. The mechanisms for success vary as much as the human brain varies, which is to say no two students are identical, even identical twins. If I may hazard an analogy though, the results of this approach are much like a radioactive sample of *N* atoms with a known half-life of one year. One cannot predict which particular atoms will decay within a year, but one can say with relative certainty a set amount will decay within a given timeframe. All in all, this is a statistical-minded strategy toward setting up classroom practices, assignments, grading and other rules, so that students are more likely to develop their perceptions of themselves as learners.

## Using Grades to Develop a Student’s Self-Efficacy

I am hardly the first person to suggest a theory similar to this. The “self-efficacy theory” was published as early as 1977 by a psychologist named Albert Bandura. Bandura showed that there was a correlation between how well students thought they could perform specific tasks, and then how well they did on those specific tasks. Since then, a number of papers have noted a correlation between students’ perceptions of their own abilities and their achievement at various levels of education. Causality, however, is difficult to prove. My hypothesis regarding the self-efficacy theory is that this is because the problem is naturally chicken-and-egg in nature; a student getting a good grade can help boost confidence, which can (but does not necessarily) lead to future good grades. Likewise, bombing a test can (but does not necessarily) lead to less confidence, which can lead to worse grades.

The goal of a good grading system would ideally be to help boost the confidence of students who need it so as to increase buy-in to the educational curriculum, while at the same time making high achievers work to maintain their high achievement, but only to the point that their confidence builds rather than erodes. In so doing, one develops motivation in students to do that which is necessary to learn the course material, because they have developed the belief that they can *and should* work to learn it.

Eventually, one hopes to build toward what might be considered an optimal academic identity: a student who is confident but not cocky, has developed a good work ethic by habit, is resilient when there are setbacks, is not only unafraid of but enthusiastic toward learning new material, and is, in short, *always learning*. In the Aristotelian spirit of people becoming that which they repeatedly do, the goal of my physics classroom is to help create people who are good at learning physics. In that spirit, ***grades should represent whatever they need to represent in order to help students develop as learners.***

Note how that statement differs in substance from other opinions out there about what grades should measure. For example, consider the difference between that statement and the following: “[We should make] grades as pure a measure as possible of student achievement; that is, make them reflect *only* student performance in mastering the public, published learning goals of the state/province/district/school.” Some will agree with that statement on principle. I will only agree with it if limiting grades to this purpose helps the students learn more. A similar statement, this time applied to a specific policy, is, “Grades are broken when they include penalties for student ‘work’ submitted late.” Some will agree with that statement as a matter of fundamental belief. A student’s behavior is not the same as their knowledge, after all, and a grade should not reflect a student’s behavior—or so goes the argument of Ken O’Connor, an independent educational consultant and one of the leading proponents of a program called Standards Based Grading (SBG).

I have a different fundamental priority when evaluating such policies: does the policy, e.g. giving a penalty for late work, increase the amount of knowledge students gain from my class? Ken O’Connor would not care about this; he only cares that the grade represents what he wants it to represent. In his own words, his sole priority is the *purity* of the grading method, and that’s that. (If he wishes to protest that statement, he can kindly rescind the passages I just quoted.) I care about the students learning the most they can learn, and I will only favor a policy that improves that *result*. Speaking of late penalties, it turns out that there is research on that. Let’s take a look at that now.

## Late Penalties and Core Principles

If putting a late penalty into the grade of an assignment has no effect on or hurts a student’s ability to learn in my classroom, then we absolutely should get rid of late penalties. However, studies have shown that the opposite is generally found to be true. The research on this subject is somewhat interesting and perhaps counterintuitive.

A 2019 doctoral dissertation by Melanie Walsh of Montana State University came to the conclusion that although giving no penalty at all for late work resulted in significant reductions in student work completed and learning associated with that work, it was also true that very strict late penalties (e.g., giving a zero if the assignment was not precisely on time) had no benefit over a smaller late penalty (e.g., deducting 10% or 20% of the assignment’s value). So FCPS’s current policy of allowing a 10% penalty for work handed in late is already well in line with best practices.

Other quantitative studies generally agree with this. A research study at Purdue University looked at homework late penalties over the course of three semesters of a college engineering class. The researcher started with a lenient policy the first semester, but got stricter in the second and third semesters. Not only did increasing the strictness of the policy result in more homework being turned in; there was also an effect on the overall scores of the class: the researcher watched the course’s failure rate of 39.1% in the first semester decrease to 23.3% in the second semester and 8.7% in the third semester. This means that, at least in this particular case, *introducing homework with a late penalty cut the course’s failure rate by over 87%*.

“The results of the study described in this paper strongly indicate that collecting homework for grading, if it is accompanied with penalties for late submissions, enhances learning among those students who do and submit their homework for grading on time. It appears that the stricter the penalties, the better the enhancement of learning among those who comply.”

However, in the same study, they noted that students *perceived* the professor as “less helpful” when the penalty was stricter, and thus the professor received poor reviews from the students when the policy was at its strictest.

This leads us to an important point about perceptions and beliefs versus reality. Late work penalties represent a prime example of where philosophy meets policy and where our priorities need to be clear. One can easily imagine someone being in favor of late penalties because “the student deserves to lose credit,” or the opposite, that work being late “does not reflect accurately what the student knows, and therefore should not be included.” These are not straw men; both views have been posited many times over the history of education. I strongly disagree with both of these statements.

Instead, my view in this regard is simply, “Which policy results in the student learning more?” I am here to help students learn, not to impose some idea of a moral high ground. This is a utilitarian, nuts-and-bolts issue, nothing more, nothing less. If I apply late penalties and watch grades and work turned in and exam scores increase, which is *exactly* what I observed after reinstating late penalties once we came back to school after the pandemic, I then have zero patience for anyone who complains about how applying those penalties is bad on principle. *No principle supersedes the fact that I am here to do what is best for the students’ education*.

As Walsh points out, accepting late work with no penalty results in fewer students turning in work, mostly due to procrastination.

“The results were clear. Giving the students more time to turn in late work led to a sizable decrease in the amount of work turned in. When students were given one extra day, 80.9% of the work was turned in. When they were given one week, only 55.3% was. When I asked the students what they thought had caused the precipitous drop, most of them offered a variation of what one student put quite succinctly, ‘a lot of kids procrastinated and just forgot about it.’”

A well-intended policy thus ends up effectively training procrastinators, whereas a small late penalty results in students developing better work habits. I am only in favor of keeping the penalty to no more than one letter grade, because, the Purdue study aside, all evidence we have at the high school level suggests there is no significant educational benefit to applying a harsher penalty, and, in fact, the students mainly learn to resent the teacher. If a high school data study showed that not accepting late work at all resulted in higher student achievement, though, I would advocate for not accepting late work. And if data clearly showed that getting rid of late penalties increased student achievement, I would advocate getting rid of them. In this, and in all other policy decisions, the right decision to make is that which helps develop the best academic identity and supports each learner’s development of self-efficacy, thus maximizing student achievement both in my classroom and beyond.

If we believe all learners can succeed, we should adopt policies that best *result in* all learners succeeding.

## Developing Self-Efficacy Through Different Formative Assignments

So, how does one develop good self-efficacy practices? The answer is multi-part, but I have laid out the groundwork already. With the numbered list at the beginning of this section in mind, let us look carefully through every section of the gradebook laid out in Part Two, as well as the rules by which each one is scored.

### The Role of Homework

Few topics in education have more opinion pieces written about them than homework. Data abounds, although often the studies are done seemingly by someone with an ax to grind. The actual picture is complicated. Some types of homework help, and some do not. The optimal volume of homework varies greatly by age, and there is a good case to be made for not giving too much of it. People hotly disagree on how much is too much. Rather than quote a hundred studies and opinion articles and compare/contrast the finer details of each, I will sum up by saying that FCPS’s current limits on homework—30 minutes per class period—are generally fine. Limits of up to an hour for advanced classes might be considered, but I am not going to fight hard over this one. Generally, the quality and consistency of the homework seem to be more important than the volume.

An example of a study that examines what constitutes good quality homework is by Pedro Rosário et. al. and published at the NIH’s website. This study addressed elementary and middle schoolers in particular, but its findings are reasonably consistent with high school and college experience, too. In short, the two most beneficial types of homework are the simple-and-fast, and complex-and-open-ended. For example, as math or physics homework increases in complexity of the topic being assigned, the homework will be most beneficial when there is more than one way to answer the problem correctly. On the other hand, for rote and introductory processes, practice should be relatively straight and to the point.

My homework category of the gradebook is split into two grades: the Stamps and the Unit Portfolio. The logic for doing things this way is based upon this dichotomy or spectrum of useful homework practices. For example, the Stamps are earned for the student’s completion of a rote process: Picture & Information, Equation, Solution (PIES).

It makes a nice acronym,[[2]](#footnote-2) but the more important point to be made is that the student receives repeated training in how to demonstrate that they have fully attempted to solve a problem. As I walk by their desk with my rubber stamp, I expect to see a useful and well-labeled picture summarizing the problem. I expect to see equations with variables only. Then, I expect to see numbers plugged in and solved. This is a standard process, and various versions of it are expounded upon in physics textbooks.

This is why I can stamp the work so quickly. The piece of knowledge being assessed is not whether they got the right answer, but whether they executed the proper steps. I have no problem giving a perfect score even if several negative signs are missing or the student misunderstood what the problem was asking for. They can correct that when we go over it or when they see the answer appear on Schoology. As for each stamp being a one-chance deal each day that cannot be recovered for late credit, that policy forces the student to constantly keep practicing the material. The student can still get a perfect Unit Portfolio grade because it does not require them to have completed it the first time for a stamp, only to have completed it correctly by the unit test.

The Unit Portfolio demands that students go through their work with a fine-toothed comb and correct it. Often, there is more than one right way to get the answer, and I will describe some of those in the video answers I upload to Schoology. Sometimes the problem is single-step and was really asking for a quick answer. That is fine, too. Complexity and creativity of student answers go hand-in-hand in a good homework assignment, because the student is simultaneously learning the standard conventions (the must-knows of the subject), as well as developing creative problem solving skills. One cannot develop creative problem solving skills with rote single-part answers alone, just as one cannot easily develop the one-step must-knows with a problem that is too complex. The Stamps grade emphasizes the rote PIES process and daily practice, while the Unit Portfolio emphasizes the more in-depth analysis and multi-step problem solving by requiring that the answers be corrected after review.

Both the Stamps and Unit Portfolio are worth a grade, and both are footholds—easy by design to attain a perfect score on them. They are collectively only worth 10% of the gradebook, so although they are a foothold, their completion does not guarantee an A. Nonetheless, this grade is significant enough that a student who sees themself as an A student cannot ignore the assignment.

There is one other major feature of the homework strategy I employ, and it is backed by a wealth of data on what types of practice best help a student learn. The principle at work is that of what psychologists call *retrieval practice*. Cognitive scientists Henry Roediger III and Mark McDaniel explain it thusly:

“Retrieval practice—recalling facts or concepts or events from memory—is a more effective learning strategy than review by rereading […] When you *space out practice* at a task and get a little rusty between sessions, or you interleave the practice of two or more subjects, retrieval is harder and feels less productive, but the effort produces longer lasting learning and enables more versatile application of it in later settings.”

Although the quotation above comes from a book summarizing current research for a popular audience, there are many technical psychology experiments that back the statement if you wish to dig further. For an example of a typical technical study, here is a quotation from the abstract of a 2008 study on the subject:

“The final recall of non-relearned items exhibited a cumulative effect of retrieval-induced forgetting such that the size of the effect increased with each block of retrieval practice. Of most interest, and very surprising from a common-sense standpoint, items that were relearned benefited more from that relearning if they had previously been forgotten.”

I will explain this by example. Suppose I give two students, Adele and Beyonce, a list of 20 elements from the periodic table and say they will be quizzed on those 20 at a later date. I quiz Adele one week later, but I wait a month to quiz Beyonce. Chances are, Adele will do better on the quiz, and this is no surprise; Beyonce had a whole month to forget everything. Say that Beyonce’s score is 11 out of 20 correct, while Adele correctly remembered 15. Something odd happens after that point, though. If I go over the quiz with each student and give them a redo one week later, Beyonce is statistically likely to perform *better* than Adele. Not only that, but Beyonce is more likely to get the 9 elements she *missed* the first time than she is to remember the 11 she *got right* the first time. We learn better when we have had a chance to forget a little. This is a well established phenomenon in cognitive psychology, albeit a counterintuitive one: if you want students to retain information longer, wait to quiz them until they have had some time to forget.

The scientists quoted above were not talking about homework specifically; any quiz or test or practice assignment could count for the purposes of retrieval practice. What I am doing is applying their general principle of spaced retrieval practice to my homework assignments’ structure. It is one thing to have students do a few physics problems for homework. It is another thing entirely for them to revisit and correct the problems later, just before the test. The time delay means that the student may have forgotten some of the material, which makes reviewing the problem a bit harder than it would be had we collected it on the stamp date and stopped there. This in turn actually helps embed the content of the homework more firmly in the student’s long-term memory.

Grading the work in two phases has two consequences. First, the two-part grade structurally places a good learning practice into the course: attempt the work now, and retrieve the knowledge and revise the work later. Second, the grade is easy to get a perfect score on, so the negative emotional effect (“…is harder and feels less productive…”) is thus mitigated. More bluntly put, if the Stamp part of the grade were absent, I suspect a lot of students would wait until the night before the test, copy the homework answers from Schoology, and thus only store the knowledge of that content in their short-term memory. If I left out the Unit Portfolio grade, I suspect a lot of students who missed a day’s homework assignment or whose first attempt was wrong would not revisit the material. We need incentives to do that which is uncomfortable, and the process of long-term information retrieval unfortunately falls in that category.

Low-achieving students benefit from this homework structure because they earn easy recognition for something that really does help them develop as learners—that is, it is an easy grade, but it is still relevant to learning the course. It is not as if I am giving five points for sitting with good posture; I am giving credit for the encoding and retrieval of course content. High-achieving students benefit by the same long-term retrieval mechanism, albeit for the slightly different motivation of securing the easy grade that validates their perception of their own academic strengths and abilities. Thus, my homework policy is designed to cater to the lowest-achieving student in my classroom as well as the highest. The natural implication of this is that it therefore also caters to every student in between, and I believe my students’ high grades and AP scores are evidence to this effect.

One caveat should be noted. It is not always possible or optimal to assign homework. In my regular physics class, students benefit from similar work, but we do it in class instead of assigning it as homework. This means that we do not get to some more advanced things during our time in class, but that is okay, because the level of the class itself is not as advanced. In other words, the foothold effect does not need to be established at home. But building a small foothold (10% of the gradebook or so) of basic practice exercises is still beneficial, as is spacing out similar exercises over the course of a unit so as to get the highest benefit out of retrieval practice. Other subjects and levels should use that knowledge to adjust their policies as they see fit.

### The Role of Curves on Quizzes and Tests

When I first assigned quizzes, they were graded out of 10 points and not given a curve. Since adjusting this practice to have the curve so that a student could get two questions wrong and still earn a perfect score, I have observed a substantial improvement in how well students perform on average as well as how anxious students get in advance of a quiz. To be clear, I mean that I observed the students’ *original, un-curved scores* improve once a curve was applied. My contention as to why this was the case has to do with the specific type of curve being used.

The bulk of research studies out there on curving students’ grades have to do with competitive curves, where students’ scores are fit to something resembling a normal distribution (i.e., bell curve), and students’ grades are assigned based on where they fit relative to their fellow students. I reject this form of curving on the basic principle that any decent grading system should maximize student performance, and that has to mean that it is theoretically possible for every student to earn a perfect score. The designers of the AP Exams, apparently, agree; the highest score of 5 is given for scores within a certain range, rather than for perfection. Also, there is no rule that says too many people can earn a 5. For the 2023 Exams, 43% of Calculus BC students earned a 5, whereas only 6% earned a 1.

There are several things I should point out about the quizzes. First, although the questions are simple in the sense that they are single-step and no partial credit is allowed, that does not mean that they are statistically easy. Often, the questions go straight at the most common misconceptions students have, which is no small number in physics. So, there is a good chance that even a great student will now and then be tricked by a particular problem. Students at the top end of the gradebook benefit from seeing that they can afford to get a few wrong without losing their grade. Students at the bottom end can see themselves go from D minus to B minus by going from 3 to 5 correct. This is actually huge for gaining buy-in to the quiz and testing process in this course. On the other hand, an A student cannot get too complacent, because going from a perfect score to a B minus is only 2 points (8 to 6).

This concept is crucial to what I believe is the ideal correlation between assignment grade and student achievement. Most of us in education would generally agree that higher grades should correlate to successful learning of the subject material, but this system begs the question: who says that the ideal correlation has to be *linear*? An un-curved quiz grading system out of 10 points would be a perfect line with a slope of 1 (i.e., 7 correct gets a 7/10), at least until we hit the 50% FCPS minimum score. My quizzes’ curved grades reveals a graph of a shape that is concave-down (albeit with one hitch at 6/10). See the figure below for details.

Let’s look at it a different way. Try to imagine someone climbing up the scatterplot above as if it were a mountain. Without the curve, a person near the bottom needs to improve significantly before they are able to climb even a little bit, whereas a student at the top is holding on for dear life lest they fall down the slope. The gentle slope assures that 1) there is still a correlation between how a student does and what their quiz score is, 2) no students are punished by the success of another as happens in a competitive curve, and 3) people near the bottom can see the fruits of their labor more quickly if they partially improve, and people near the top are relatively safe even if they get some small thing wrong.

I have searched many papers on curving, but none of the ones I find show anything like the curve I have presented here. Rather, my quiz policy is the result of lived experience in the classroom and tweaking the parameters until I found what worked best. That said, it is not entirely haphazard and without justification; this policy is perfectly consistent with the principles of helping students develop, or keep improving, their academic identity and self-efficacy. They must put in the work, it is doable, they get credit/acknowledgement for it, and so on.

The test curve works similarly. Being nice for the first several points, but then being nice no longer, results in a similar shape of un-curved vs. curved results. Remember this later when we look at what went wrong with some FCPS schools’ attempts to implement certain reforms, most notably the 4.0 scale or M-P-D (Mastered-Proficient-Developing). We will get there. But bookmark this for when we do, because just as informative to my practice as seeing the quiz curve work was seeing other policies fail.

Quizzes and Tests made up 20% and 30% of the gradebook, respectively, for AP Physics 1 and 2. On the other hand, AP Physics C had no quizzes; their tests made up 50% of the gradebook. This year, I am introducing quizzes to the AP Physics C curriculum. So far the students seem to be doing well on them overall, but it will take time to assess the full effect of introducing quizzes to the more advanced course.

While most of these notes and my own research have been based on AP classes, this year I am teaching a non-AP (regular) Physics 1 course, and I implemented this grading course there as well. I did this for a few reasons. Implementing a grading curve to the AP, college level, honors, or other advanced courses, but not the other courses is inconsistent and unfair. Additionally, the reasons that the curve works are universal. It is designed to work for both low- and high-achieving students, and those students may present in any given class. This is my first year doing it, but so far the results look good.

### The Role of the Quizzes and Tests Themselves

Quizzes and tests are not merely measurements of how well students know course material; assessments play an active, formative role in education, and the act of testing students’ knowledge actually causes retrieval (see the homework section above) and affects how well the students retain the information. This is a well-studied area of educational psychology, and the technical term for it is the *testing effect*.

Two things happen as a result of tests or quizzes being given and then graded. First, the student must retrieve information from their long-term memory, which helps to cement that information in the brain. Application of the retrieval effect to testing was first observed experimentally over a century ago and is presently the subject of research aiming to maximize its effect by altering such things as the assessments’ design and presentation.

Second, the student receives feedback that targets the knowledge on which they are weaker. Grading is a form of feedback, and losing two points for getting an answer wrong can cause the student to revise their mental model of the information. Cognitive scientists H.L. Roediger III and A.C. Butler performed a study titled, “Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing,” which Roediger would later summarize as follows:

“The experiments show that feedback strengthens the effects of testing alone, and that feedback may be more beneficial when it’s slightly delayed.”

One might point out that the word, “feedback,” may or may not imply a grade, but Roediger and his co-authors address that elsewhere:

“In the classes where [practice] exercises count toward the course grade, even at very low stakes, students achieve higher success over the course of the term compared to students in classes where the exercises are the same but carry no consequences for the grade.” (Brown, Roediger III, & McDaniel, 2014, p. 233)

A study of introductory biology students at the University of Washington verified not only that students’ failure rates dropped as a result of including formative quiz grades, but that final exam scores rose significantly. So, to anyone who suspects that such quiz grades are a fluff score that artificially boost a student’s grade, the evidence clearly says otherwise. Also of interest is the finding that lower-achieving students disproportionately benefited from the practice exercises counting toward their grade compared to high-achieving students, which means that grading this formative work was fundamentally an equitable practice.

“To summarize, our data show that prescribed [graded] active learning benefits students in introductory biology more than voluntary or ‘unenforced’ active learning. This study also shows that if introductory science courses are reformed in a way that prescribes constant student participation and practice, it is likely that more students, especially those who are at high risk of failing, will gain the discipline and intellectual tools required to be successful in the life sciences.”

One other result stands out from this study. The researchers looked not only at whether practice quizzes were more effective teaching tools when they were graded versus ungraded; they also looked at the question of whether grading the students for correctness or grading them for participation made a difference. They found that although students were more likely to answer quiz questions correctly if they were graded for correctness, final exam scores were unaffected—meaning the students actually learned and retained as much information in the end if they were graded for completion/participation on this practice work as when they were graded for actually being correct. Both grading for correctness and grading for completion resulted in significantly better achievement than ungraded practice.

So now we know that 1) formative assessment is a useful tool for learning, and 2) formative assessment is more effective as a learning tool when it is graded—even if it is only graded for completion or participation. The result is unambiguous, and as this paper points out on page 137, this study is just one of many that reach similar conclusions.

What is less clear is *why* grading formative assignments is better for students’ learning than just giving ungraded feedback. The researchers in the Freeman, et. al. study hypothesize that it may be a combination of two factors. First, the act of grading assignments provides a clear structure for students who may benefit from increased structure in education—specifically, at-risk students. Second, attendance was observed to increase when the course had daily graded quizzes, and attendance is highly correlated to college courses’ pass rates. These may both be true, but I will add one more hypothesis based on my own experience. By grading an assignment, certain qualitative types of feedback are more likely to be present than if the assignment is ungraded. A grade communicates not just whether something was correct or incorrect, but whether it was a minor or major mistake (depending on how many points were deducted); whether it was a one-time or repeated mistake several times on the assignment; whether a topic was important enough to show up on a graded assignment in the first place and is thus worth remembering; and whether any mistakes have similarities to each other and might thus be remedied by a common correction or addition to the student’s knowledge.

Take this all together, and what you get is the inevitable conclusion that testing and quizzing, if done in a way that is consistent with practices that best enhance a student’s learning, are *formative exercises rather than summative measurements.* I am hardly the first person to say this. Psychologist Herbert Spitzer of the State University of Iowa came to this realization after reviewing the literature and conducting his own experiments with the help of over 3600 middle schoolers in the 1930s:

“Achievement tests or examinations are learning devices and should not be considered only as tools for measuring achievement of pupils.”

It would take a long time for Spitzer’s work to be taken seriously, but this is in fact now the consensus view of educational psychologists. We will expand upon this concept at the end of this part of the paper.

The timing of quizzing is also important. Here are the principle findings of a 2011 study on the subject, as summarized in its abstract:

“In Experiment 1, three quizzes on the content were spaced across the coverage of a unit. Quizzing produced significant learning benefits, with between 13% and 25% gains in performance on summative unit examinations. In Experiments 2a and 2b, we manipulated the placement of the quizzing, with students being quizzed on some content prior to the lecture, quizzed on some immediately after the lecture, and quizzed on some as a review prior to the unit exam. Review quizzing produced the greatest increases in exam performance, and these increases were only slightly augmented when the items had appeared on previous quizzes.”

The researchers are careful to note that not all such quizzing has to be graded in order to have a positive effect. Throughout a unit, I will give practice problems, warm-ups, and other basic checks for understanding to the class. Only the most important of these is graded: the quiz that is used on the unit’s review day, the day before the test. During the rest of the unit, the homework and lab experiments are the only things completed for a grade. This ties the act of assigning a grade to the specific assignments and exercises whose successful completion best actively causes the student to learn and retain their knowledge. Also, by the time the student is being graded for correctness on something, it is not the first chance they have had to answer a similar challenge.

The clear message from all of the research on this subject is that graded formative exercises help all learners, but they disproportionately help the students who most need assistance. This includes learners with special needs and/or accommodations. Any school system that wishes to decrease achievement gaps, or for whom equity is a goal, should take note of this.

### The Role of Labs and In-Class Activities

Now we reach what was the largest portion of the AP Physics 1 and 2 gradebooks by percentage: labs. Lab experiments in a physics classroom are an odd mix of training rote processes similar to the homework and creative analysis at the end. It would be completely inappropriate to assign these as homework, because they break the aforementioned homework rules. Recall that homework should ideally either be quick and closed-ended or longer and open-ended; however, our labs are by their nature longer and closed-ended assignments because of the strict rubric they follow. It is thus worth differentiating our lab experiments from, say, a Science Fair project or other presentation homework assignment. Our labs involve considerably less creativity and presentation skill than a project like that.

I grade labs using a repetitive rubric based on the student reporting their lab’s mission, method, data, a graph of their data, analysis, and uncertainty of both accuracy and precision. The goal of assigning the labs with the repetitive rubric is to instill solid science experiment practices that can be used in any college lab environment or extended to science jobs anywhere. Most of the rubric is in line with things I either had to do as part of earning a physics degree at the University of Maryland or are else consistent with practices my mother used as a researcher in infectious disease at the University of Maryland Medical Center in Baltimore. The labs also follow the AP Physics Science Practices, but that is so much more boring.

Homework can be seen as general preparation and retrieval practice, quizzes can be seen as targeted assessment, and tests can be seen as comprehensive assessment, but the labs are something else altogether: work experience. They are also relevant assessment practice. Every year, one of the most valuable free response questions on AP Physics exams has students design a lab investigation from scratch. The lab assignment is more important than the grading. However, insofar as grading policy goes, the rules for labs are quite simple: follow the rubric exactly.

Labs may sound like they are specific to a science classroom, but the basic idea of a content-specific major type of assignment that mimics work experience in some way is well studied as an educational technique. In some classes, a research paper might fill the same role. The same may be said of projects, presentations, and portfolios of some kind. Whether or not they follow as strict a rubric as my own or are much more open-ended, it has been shown that students best learn and improve at professional tasks and procedures (such as doing an experiment) when the practice is *interleaved*. This is another term from cognitive psychology, and they mean applying the knowledge alongside other pieces of knowledge so that connections are made between the desired skill and other things to be learned in the class (we will apply our lab techniques to Newton’s laws, to conservation of momentum, to rotation, to fluids, and so on).

Benefits of interleaving education, both within a lesson as well as over the course of a semester or year, are well established at this point, although applying the general principle to specific levels of education or areas of study remains an ongoing topic of research. In my opinion, the main takeaway from papers on interleaving is that there is more than one good way to incorporate it into a classroom, and any method of interleaving an essential skill of the class into multiple other class subject areas is likely to see some measure of success as an educational technique.

A related educational concept used in labs (or their equivalent in a class outside the science department) is that of *generative learning*. I design my labs so that they are not “cookbook” labs with prescribed methods and data tables. Instead, I give the students a mission (e.g., “Get a linear graph of data whose slope is the mass of this object”), and the students then have to generate their own solution to the problem. It has been shown that students better learn and retain information when they are at some level responsible for generating the answer as opposed to reading or otherwise being provided with the answer. [[3]](#footnote-3)

The goal of the lab rubric is therefore to train students in a set routine that can in turn be used as a format by which a wide range of experimental problems may be solved. Students invariably mess up here and there, but most mistakes are small and only cost one or two points out of 50. So most students are able to earn a higher grade, even though there is no curve. This is a different way of achieving the same effect as the curve for quizzes and tests. By making the labs a practice in processes, it is easy for them to earn enough points for a good grade, but the rubric is sensitive enough that a perfect score, while completely doable, really does represent perfection.

As with homework, quizzes, and tests, a good lab grade fundamentally represents the successful completion of a formative process that is shown to lead to greater knowledge and retention of the course material. Not all formative processes are alike, though. I mentioned back in Part One that there was a good reason why labs and homework were better predictors of who would get a *low* AP exam score, while quizzes and tests were better predictors of who would get a *high* AP exam score. Well, let me put it this way:

* Homework = easy (in terms of what it takes to get a high score) and simple (in that it tends to address one single learning objective at a time).
* Labs = easy but comprehensive.
* Quizzes = hard but simple.
* Tests = hard and comprehensive.

There is benefit to scaling assignments in terms of difficulty in a course. There is also benefit to scaling assignments in terms of complexity. The former is how you build confidence. The latter is how you build experience. Build a student’s confidence and experience, and you will build their self-efficacy. You will help the highest and lowest achievers progress, and you will help everyone in between, too.

### The Role of Extra Credit

Next, let’s next talk about extra credit. There is data from various sources out there on how well extra credit works, but there are also a lot of opinion articles. I will steer clear of the opinion articles here, and I will also stay away from data regarding superficial extra credit, such as bringing in a ream of paper or tissue box for extra credit points.

Substantive extra credit is made of challenge problems, tasks beyond the regular scope of the course, or extra practice. For this type of extra credit, studies indicate that advanced/gifted learners benefit most, while struggling learners would benefit more from reviewing the core course materials. Therefore, I take the view that in an Honors or AP class, extra credit is a valid form of differentiated instruction in advanced coursework.

Indeed, research shows that gifted learners tend to do better and learn more from challenging problems given as extra credit than when given on the test. Furthermore, performance on extra credit in advanced courses correlates strongly with advanced exam performance. In fact, it correlates so strongly that it can be a better predictor of exam performance than a student’s GPA. Based on this data and what I have witnessed over the years in my classes, I will be applying extra credit in my AP Physics classes going forward, but I have no such plans for the on-level Physics class. With them, we will spend our Learning Seminars making sure that anyone who has fallen behind gets a chance to catch up on course content, assignments, or both.

## All Learning is Formative, So All Grading Should be Formative

Now is a good time to pause and reflect on how the grades I give go about measuring learning. To see this, it would first be best to give a better definition to what learning means, exactly. In cognitive science, learning is a three-stage *process* of “Encoding (or acquisition of information); storage (persistence of information over time); and retrieval (later use of the information).” If all of my graded assignments represent tasks at which these three processes must happen, then what I am measuring with a grade is fundamentally this learning process.

Put another way: *learning is a verb, not a noun*. To use an analogy from physics, it is much more similar to measuring velocity (speed in a direction) than measuring position. *Never in my entire process do I single out or grade students’ mastery of course material explicitly.* The assignments are tied/referenced to the course standards, but that is not the same thing. I never tell a student that they have mastered the application of Newton’s 3rd Law to a collision between two objects; rather, I grade a test or quiz question on which the student had to retrieve and apply that concept. The grade fundamentally represents this application and retrieval—the learning practice—rather than the individual concept’s mastery.

This means all of my assignments are formative, and all of the grades I give are formative grades. The homework is largely there for the purpose of short-term encoding (stamps) and retrieval (portfolio). Quizzes, labs, and extra credit serve more advanced ongoing storage and retrieval roles. Their function is to help move encoded information to the long-term memory. Tests and the final exams/reassessments do this even more thoroughly, because they require retrieval of mixed information rather than massed/related information, come with a time delay from the original presentation of the material, and require more advanced generation of student responses (i.e., students have to apply their knowledge to situations they have not necessarily seen before).

The result is that the student’s course grade represents their learning process over time. When I grade a test, it is true that I am marking whether a student’s answers are correct, whether they used correct units, whether their reasoning is solid, and so on. But at a deeper level, what I am really grading is whether they were able to retrieve appropriate knowledge when given cues/prompts meant to trigger such retrieval. I never sort the information by learning objective. I don’t have to. If the student is able to retrieve a variety of course knowledge on the test, and they have done the relevant labs and quizzes and homework assignments, I can state with relatively high confidence that this student not only knows the information, but will retain it. The grade ends up being a very good indicator of present as well as future content mastery. Is it any wonder that students’ grades in this system end up predicting the students’ AP exam scores so well?

Along with the cognitive science behind the assignment and assessment design in my courses, there is also a pattern in how easy/hard it is to earn a good score, and the scoring system I have applied is designed to help a student’s self-efficacy at every turn. The theme in every grade category is that it is relatively easy to get a B, hard but doable to get an A (except for the homework grade, where it is a foothold), and almost impossible to fail under this system as long as the student is consistently doing their work. If that work is aligned with what they need to know on the end-of-course assessment (in our case, the AP exams), then we have a successful grading system in place.

## Reassessment as a Formative Tool

Last but not least, we should talk about the reassessment system and those final practice exams in particular.

In the past, I used a reteaching assignment that the student had to complete first. Then, they could come in after school or during a Learning Seminar and take an alternate version of the test. This had many problems. For one thing, not every student had the extra time to commit to coming after school or doing a reteaching assignment. For another thing, students would often cram and memorize all they could for the retake, take it as soon as possible after the test itself, and immediately forget the information afterward. My most frequent retakers usually earned a 1 on the AP Exam itself, even if they managed an A or A− in the class. Additionally, I limited retake scores to 80% just so I would not get 30 perfectionists after every single test trying to improve their A by 1 point. I had to do this, because the work load from the teacher’s perspective was unbearable otherwise.

Since moving to the strategy of delaying retakes until April, one or two students each year will ask early on whether they can please reassess sooner. The plea is almost of the nature, “Let me do it while it’s fresh on my mind.” That phrase, “fresh on my mind,” might as well be an admission that they do not believe they will retain the information. Thus I insist that the student must wait until April to reassess.

To be absolutely clear, that is the conclusion I reached after years of witnessing students forget information after retaking a test, and at the time, I did not know of actual research studies that agree with this observation. In researching this paper, I have found this is actually a well known phenomenon in psychology, and it is a much better practice to delay a reassessment opportunity than to allow or require a reassessment be done immediately following the original assessment. We already met this concept while going over retrieval practice in the discussion of homework policy above, but the exact term from educational psychology is *creating desirable difficulties*, if you care to research further.

I have also found one paper that specifically looked at a system of tiered reassessment similar to what I do, but in a college math course: “When there are quizzes followed by a test, the test can replace the quizzes,” for example. This 2009 study concludes that the practice raised students’ grades by a full letter on average, reduced the number of students who dropped the course, and increased information retention. Again, I developed the strategy organically without knowing about studies like this—they were never a part of my Masters in Education training. Nonetheless, it is gratifying to find there is a real basis behind my observations and the practices I have developed as a result. Now if only we could get educators’ professional development courses to train teachers and administrators in such things.

In addition to effectiveness, consider how equitable the new system is in contrast to the old one. Now, the unit tests are automatically counted as a retake for quizzes and there are 3 AP-level tests at the end of the year that everyone completes. The AP test questions correspond to subjects on the unit tests and count as an automatic retest. So without ever having to do extra course work or coming in after school, *every student* is *automatically* reassessed on every single quiz and every test for the entire year. As with the test and quiz curves, the students do not need to be perfect, but must demonstrate proficiency.

There are also additional benefits that come out of this system. The timing of the test encourages students to retain information for the entire year, and the lack of retakes on tests throughout the year represents a benefit to teacher workload. If students have not perfectly retained their content knowledge from earlier in the year (which is probable), then they must use extra effort to retrieve or reconstruct that knowledge, which we know from cognitive science helps to solidify their knowledge of the material. Because the end-of-year retake represents long-term learning and there is no longer a need to discourage obsessive retakes, I allow all tests to be reassessed up to 100% credit, so the students can see a benefit to this system that the 80% limit could never provide.

Anecdotally, of all the strategies I have employed in my grading system, the reassessment system has gained the widest student and parent support and seems to have created the greatest benefits in increasing student achievement, lowering achievement gaps, and increasing the alignment between course grades and AP exam grades. It worth stepping back for a moment and looking at the big picture as well as the research on this topic. We have already seen that the testing effect means students learn better when they are tested on a subject. We have also seen that the best practices in assignment grading and design involve the student having to retrieve knowledge from a considerable time ago, and that interleaving multiple topics results in better retention of information.

Put this all together, and here is the conclusion I have reached: properly designed, the most effective *formative* assignment we have the ability to give might just be something like a final exam. It is worth repeating: all learning is formative. If grades are supposed to measure and maximize learning, it follows that every grade should be a formative grade—a measure of a process, not just a result.

## Summary: Structurally Encouraging Motivation through Tiered Formative Assignments

When a student develops their identity as a learner, they develop their education generally. Feedback is crucial to this development, and grades can be structured so that every action toward this is encouraged and recognized. Because grades are given increasing importance as the assignments gain complexity, and graded knowledge is spaced out and interleaved, the student’s grade ends up being an accurate measurement of their learning process. The system guards against the dumping of information while still giving students multiple chances to demonstrate they know the course material. Simultaneously, the system builds the student’s confidence and expectations for their own achievement.

This is what motivation looks like: a combination of belief that one can succeed (self-efficacy) and recognition/acknowledgement of that success, for learners at all ends of the spectrum. It helps struggling learners buy into the system and start climbing, while keeping advanced learners engaged and applying effort. Note that I am not claiming that my policy is the *only* policy that can achieve this feat; any policy that helps students at both ends of a class’s achievement spectrum is likely to help all the students in between.

We are about to head into the part of this paper that talks about SBG in depth and detail. SBG proponents have their own ideas about student motivation, and we will meet these ideas in due time. First, we should examine some of the features and policies advocated by SBG proponents and see how their results and theories stack up versus the results and theories I have shown thus far.

1. There is a time and place for ungraded practice, too. We will look shortly at the effects both of ungraded practice and of graded formative assignments on students’ learning. It turns out there are studies on this, with actual data to consider. [↑](#footnote-ref-1)
2. I have rejected a fellow teacher’s suggestion of changing the word Picture to Diagram on the grounds that it makes the acronym less fortunate. [↑](#footnote-ref-2)
3. There is an abundance of research on the benefits of generative learning if you wish to look it up, but be warned that modern research databases are liable to confuse the term for “generative AI,” a subject on which there are hundreds of studies available that have nothing to do with our present subject at hand. One has to dig to find the research relevant to human education. [↑](#footnote-ref-3)