Grading for Equity: A Rationale and Guide for Instructional Change

Andrew Berns & J. Philip East
Computer Science Department, University of Northern Iowa

This paper argues that traditional instruction and its grading practices are inherently inaccurate, inconsistent, and, too often, address students' non-academic behavior which introduces bias. We begin by briefly describing our background. Next we describe "traditional grading practice" to ensure our readers understand the practices to which we refer. We then elaborate on some of the issues we believe exist based on our reading of Feldman (2019) and also discuss alternative practices we believe overcome these issues, and then describe a planning process for implementing these better practices. Finally, we discuss the benefits we have experienced or expect as a result of using equitable grading practices.

1. Relevant Author Experience

Philip's educational background is a mixture of computer science and education. He was hired at the University of Northern Iowa in 1985 to oversee a master's program and an undergraduate minor for high school computer science teachers. Very early in that effort he realized that he didn't have a rationale (other than that's the way it's always been done) for any teaching practice. Since then, he has been working to identify educational practices for which he can provide a rationale that is based on research and theory/principle. One aspect of that work has been focused on grading in computer science, generally with the theme "I hate grading" and how to make it better, e.g., East (1999), East (2000), East & Schafer (2005).

Andy's background is in software engineering and computer science. After completing his PhD at the University of Iowa in 2012, he spent three years as an assistant professor at the University of Wisconsin-La Crosse before moving to the University of Northern Iowa in 2016. Grading has always been the least pleasant task for him and as such he has spent significant amounts of time investigating how it can be improved for him and his students, e.g., Berns (2020).

In general, our experience is much like that of other college instructors. As students, we spent 18 or more years in school receiving mostly traditional instruction. As instructors, we then spent the bulk of our careers using traditional instructional practices and at some point began searching for ways to improve what we were doing. We made some progress toward better practice but seemed mostly to be doing trial and error debugging of our teaching (and grading).

A central issue with grading is that too often we spend an inordinate amount of time providing feedback to students as we grade, but the students seem to look at the points awarded and ignore the feedback given. And, too often they come to office hours purportedly to better understand the grading but seeming to mostly argue about points. We really wanted to do something different. In that frame of mind, we encountered Feldman (2019) and finally found a framework we could use for revising our grading practice. Even better, we have found that it is revolutionizing our thinking about teaching.

2. Review of Traditional Grading Practice

When we talk about traditional grading practice, we are referring to a number of characteristics which differ slightly depending on the class being taught. Typically, there will be several categories of grading elements. For example, a programming class might have a number of lab activities, some homework sets or problems, perhaps some quizzes, probably attendance and participation, some larger exams, and perhaps some extra credit activity. There may be some pair-programming involved. Each category will typically total or average the scores on individual elements with those values combined to produce a value indicative of the final grade. Alternatively, a project or capstone class might focus on group projects having some milestone reports, individual reports of group process, an overall project submission, and a final exam. Again, the scores on components would be totaled or averaged and category scores combined

to produce a final grade value. Additionally, missing work typically receives a zero score. Students whose work seems nearly identical will receive some sort of penalty, perhaps a zero score or a failing grade.

Grades are typically determined by scaling the overall possible points to a percentage or 100-point scale and identifying ranges of scores for each grade. A very common set of ranges is one where 90-100% indicates an A; 80-90%, a B; 70-80%, a C; 60-70%, a D; and 0-60%, an F. Some instructors use the grade ranges as a maximum and reduce some of the cutoffs in light of natural "breaks" in the set of scores or by examining individuals near the cutoff points or some other rationale.

Feedback is typically provided when assigned work is graded, often requiring substantial time and effort on the instructor's part. Additional feedback is provided by discussing homework or exam results in class and, of course, via individual discussions with students during office hours. Sometimes, or in some courses, feedback is also provided while supervising student work in lab or during class time.

3. Issues with Traditional Grading Practice and Some Alternatives

Feldman (2019) indicates that grading practice should be 1) mathematically accurate, 2) free of activities subject to bias, and 3) motivational to students. When we evaluate our grading, however, we find that our traditional grading practices fail to satisfy any of the three criteria.

Grading Homework

While reflecting on our practice (prior to encountering "Equitable Grading") we had mostly concluded that including homework in grading was not good—whether the grading was for correctness or just for getting it done. Our reasoning was that homework was a learning activity. The homework was assigned so students could learn. If they had already learned the desired content or skill, they did not need to do the homework. For those not already capable, doing the homework would be useful but some students would "get it" right away while others would struggle. It seemed wrong that those who struggled to learn —but did learn—would get lesser grades than those who learned more readily or already possessed the desired capability.

When we read Feldman's book (2019), we encountered a well-formed rationale for not including homework in course grades. The rationale contained several points.

> "We can't deduct points for incorrect answers on homework—which penalizes students for mistakes —**and at the same time** (emphasis added) tell students that homework is just practice and that mistakes are a necessary part of learning." (p. 219)
>
> [When homework is graded,] "The weaker students—those who most need our help—won't attempt the practice, will learn less, and therefore will be less prepared for the summative assessment." (p. 130)
>
> Grading homework is meant to provide an incentive, beyond desired learning, in the form of points or grades. "For student in these situations we could make each homework assignment worth 10, 1000, or a million points and they still wouldn't complete it." (p.136)
>
> Finally, the grading of homework allows (or encourages) students to think in terms of points rather than learning the content. It leads to students arguing for more points rather than trying to understand their misunderstandings.

"Without us realizing it, including homework in a grade acts often to perpetuate inequalities and the achievement gap." (p.137) It is **biased** in favor of students with better background knowledge and student skills.

An alternative one might try, then, would be grading for completion or attempts at the homework, as this seems at first to be less biased and more reasonable. However, this too can cause problems. A struggling student might attempt all the homework but get many parts incorrect and as a result do poorly on the assessment. When final course grades are calculated the homework completion grade shows that the

student did well on homework but poorly on the assessments, thus raising the overall grade and **inaccurately** reflecting the student's capability.

Instead of grading homework, even based only on completion, one could choose simply to exclude homework from grade calculation completely. Choosing to not include homework in the grade has multiple benefits.

- It removes any incentive students have to copy.
- Instructors do not have to spend massive amounts of time grading homework.
- Students no longer argue about points and become more amenable to receiving feedback since it is not tied to evaluation.
- Intrinsic motivation (being successful at learning) replaces extrinsic motivation (grades) thus enhancing personal responsibility
- Students are no longer punished twice—once for not doing the homework and again when performing poorly on quizzes and exams.
- Students are not penalized for environmental factors beyond their control—illness, transportation issues, home/work responsibilities, family issues, etc.

Of course, one of the primary purposes for grading homework is to provide feedback, an alternative to grading must be developed. Some possibilities include providing answer keys when appropriate, peer-review of work, and student demonstrations or code reviews during class time (with students volunteering or being selected by the instructor).

Some of the discussion below may seem less pertinent if homework is not graded, however, it relates to other elements of grading and is still relevant.

The Grading Scale

The primary issue with the 100-point grading scale—90/80/70/60—is the inherent lack of logic. Are there really 10 gradations of an A (or B, C, & D) and 60 of an F? Additionally, is there really a difference between a 92 and a 93 (or 94). In our experience, we typically count off for a few things for each grade level, but not 10 and certainly not 60 things. Once we recognized this, it became clear that our grading practice was not mathematically **accurate**.

Several alternatives exist and likely depend on the assessment used and the individual instructor's preference. When a general sense of the work is being assessed, one could use a four-point scale roughly equivalent to letter grades. Having rubrics (known to students) that align with each mark could enhance this alternative. A 15-point scale could also be used. It would offer three different values within each grade range—likely more accurate and defensible than the 10 variations of a grade range available in the percentage scale.

Unequal Grade Ranges

With the 90/80/70 grading scale a student receiving two B scores (e.g., 85) and a low F score (e.g., 0) would fail (with a score of 57). Two B scores (85) and a low C (70) would yield a B- (80). Two B's (85) and a low D (60) would be near a C+ (77). Even a middling F (e.g., 30) would still bring the two B's down to a D (67).

If the F range were the same size as the others, the two B's and a low F would result in an average score of 73 or a C-. That seems more reasonable and more **accurate**. And, as discussed above, using a smaller range of scores might be produce a more accurate assessment of student capability.

The Zero Score

As seen above a zero score dramatically affects a student's grade. Typically, a zero is used to indicate that a student did not complete the work or was absent. Thus, the zero score does not actually represent an assessment of the student's capability and, therefore, is **inaccurate**. An accurate grading system would take this into account.

A score of zero suggests absolutely no capability with the indicated content/skill. That might occasionally be the case, though rarely. In our experience a zero is often used as a place holder allowing for a calculation of an approximate grade at the present time. Avoiding this requires developing an alternative way of determining an overall grade. The discussion below also addresses the use of the zero.

Penalizing Late Work

Points deducted for submitting work after a due date/time is automatically **inaccurate** since the penalty has nothing to do with a student's academic capability. Additionally, doing so may well disadvantage (be **biased** against) certain kinds of students, e.g., those with:

- Weaker prior knowledge of the content
- Poorer academic/student skills or slower learning or work rate
- More life entanglements such as overwhelming schedules, family illness or death, psychological problems, etc.
- Less engagement with the content

Furthermore, late penalties may cause students to quit working (if they don't believe they can finish in time, for instance) or strengthen the temptation to cheat (to complete the work by the prescribed time).

An alternative to late penalties is simply to not have them. Feldman (2019) suggests that not penalizing late work (and, perhaps, explicating the rationale) will lead to an increase in the quality of work submitted. Of course, there is an ultimate deadline for working on past-due assignments. The individual teacher will need to adopt one that makes sense, e.g., when the unit assessment is done or a week before the term ends.

One might ask, of course, about developing a student's "soft skills", including timely completion of work. For those interested, we suggest examining chapter 13 of Feldman (2019).

Penalizing "Cheating"

Penalizing "cheating" is thought to be a tool for teaching appropriate academic behavior. As with late work, however, reducing a student's grade for cheating is **inaccurate** since it reflects a student's behavior rather than academic capability. And, the students inclined to attempt "cheating" are likely those with lesser background knowledge or poorer study stills. Additionally those who get caught in "cheating" behavior are likely those who are less able to disguise it.

Certainly, teachers should not condone or reward academic dishonesty and we are not suggesting that. But other responses might be more appropriate. One possibility is that the student be required to complete the work without cheating.

Leaving issues of scoring aside, there is an additional point with respect to cheating on homework: if the student is able to develop the desired capability and demonstrate it in an assessment, what does it matter if they copied their submission for the homework? If they did not learn, the assessment will so indicate and nothing was gained from the copying.

Including Attendance & Participation in Grading

Attendance and participation scores are often used to foster active learning, to encourage those with lesser academic skills, and to develop soft skills such as focusing attention. From an equitable grading perspective, however, their inclusion in grade determination makes the grading **inaccurate** as they measure student behavior and not academic capability. Attendance behavior may be beyond student control (transportation, alarm clock, illness, etc. issues) or may result from student "choice" (an alternative educational experience or oversleeping/sleeping in). An instructor cannot know with certainty what caused the absence and may well end up making **biased** decisions when excusing absences. It is better to not include attendance in grading.

Like attendance, class participation seems particularly important for active learning. However, the rationale for not including it when grading is even stronger than for attendance—it is highly susceptible to cultural **bias**. Some black cultures, for example, consider class participation as "acting white". Some

Native and Asian cultures venerate authority and would not consider offering their opinion/thought in deference to the instructor. While we may well wish to encourage class participation, we should not include it as a part of student grades.

Including Extra Credit in Grading

As with attendance and participation, frequently our rationale for including extra credit when grading was to provide those in need with a grade boost, and then allowing all students to complete it in the sense of fairness. The result is that those who are most able—either academically, due to supports they have, or because of life characteristics—are the ones taking advantage of the opportunity. Thus, those who most need the "extra" credit are rarely able to receive it. For this reason, extra credit introduces **bias** into grade calculation.

Furthermore, extra credit also results in the final score being an **inaccurate** representation of what the student was to learn. If the skill in question is important, then it should be required for everyone. Allowing "extra" points means two students with the same final grade may not have demonstrated the same level of competence. While we often wish to provide some reward for students going beyond the coursework or participating in outside activities, we need another means to do so to ensure our grading is equitable and accurate.

Including Group Work Scores in Grading

Pair programming and project work often result in a product produced by multiple students being graded and the group score being given to all members of the group. Often, one of the goals of the group work is to enhance the development of soft skills involved in group work. Giving individuals the group's score on the product produced, however, is by its very nature **inaccurate** as it does not measure individual capability. For some students the score is too low and for some too high.

Additionally, when the goal is the development of soft skills, soft skill development is frequently not actually assessed, thus we gain no knowledge of instructional success with respect to that goal.

In this case analyzing our work with an equitable grading lens shows a flaw in our instruction. We are not directly assessing group-work soft skills and, in our experience, are not actually providing instruction in those skills. It now becomes incumbent on us to rethink what we have been doing and to provide explicit instruction on soft skills and develop assessments to gauge student learning and the success of our instruction.

Single Opportunity Grading (Early Failure)

Traditional grading practice seldom allows student do-overs. Presumably, this is done to avoid additional grading and to keep everyone moving forward. The problem created, however, is that a student who learns more slowly or had difficulty with (or was absent during instruction for) an early, basic capability will not only have more difficulty later but will be put at a large disadvantage grade-wise even if they do ultimately "get it". For example,

> Philip once had a student who consistently made the same errors during the first weeks of a programming course. Eventually, it became clear that the student was not learning from the feedback and was told, "You have to pay attention to correcting the errors you are making and avoid them in the future." The student's work began improving and by the end of the semester he was doing A and B work. However, the early failure (and the grading scheme being used) meant the best grade the student could receive was a C-.

It seems inordinately unfair that a student who does A work at the end of the semester on activities that require all desired capabilities can, at best, barely pass the course.

An alternative to once-and-done assessment is to offer multiple opportunities for students to demonstrate the desired capability. This can be done by offering re-assessments of individual outcomes or using later

assessments that are cumulative in nature. It really should not matter that students do not "get it" the first time so long as they demonstrate the capability at the end of the course.

A number of issues might arise when deciding to provide retakes of assessments. Whether the assessments will be binary (credit/no credit) or points-based, e.g., a 4-point grade scale or 0-1-2 scale (not okay/ok/good) needs to be decided. When retakes are no longer allowed also must be determined. They will stop at the end of the course unless some mechanism exists for work beyond that time. A time-dependent cutoff might be established, perhaps the beginning of final exams. Alternatively, particularly if points-based assessments are used, there might be a limit on the number of trials allowed for students wishing to improve their score. There may also be some dependence among the assessments, for example, success with the an early assessment might be required for later retakes (but only retakes, not the initial assessment). Other issues might also arise.

We have introduced this capability into several of our courses and are quite pleased with the results. Doing so was made easier as we quit grading homework. That allowed us to develop multiple versions of each assessment. For a programming course the binary-scored assessments are typically comprised of several multiple choice or short answer items and a couple longer-answer items that can usually be graded one in a minute or two. It is relative easy to determine whether students "get it". Sometimes we are uncertain and ask students to come see us and explain or expand on their answers. We have discovered that students no longer question us as to whether they have demonstrated the desired capability. They accept our judgment and there are no more points discussions/arguments.

Using Grades As Reward/Punishment

It would seem that students would work to get good grades and/or avoid bad grades. According to Feldman (2020), however, research indicates otherwise. Apparently, extrinsic motivation (e.g., grades) has several effects with respect to rewards or the expectation of good grades:

- "The promise and contingency of a reward in the future … reduced students' intrinsic interest" (p. 154), i.e., they only do that kind of activity if there is the promise of a reward—not the result we wish to develop in our students.
- "Those [students] who were motivated entirely intrinsically were more creative and successful when completing complex tasks. (p. 155)
- "[E]xtrinsic rewards may reap short-term results but … can cause undesirable side effects", e.g., "students work only to the level that triggers the reward—and no further" and "the healthy behavior disappears once the reward is removed." (p. 155)

Additionally, the hope that desiring to avoid (more) bad grades does not motivate students, rather it causes students to withdraw from the activity, lose confidence, blame themselves, and feel helpless.

The idea is that teachers need to change their perception of grades. Grades need to be thought of as an indication of student success (and teacher success) rather than as a reward or punishment based on student performance or behavior.

Averaging/Totaling Weighted Categories

Our experience with grading (as both students and teachers) has pretty much always involved one of two approaches. One approach is to average of scores in various categories that are then weighted and added to provide an overall score. The alternative in our experience was to plan in advance the weighting and award individual assignments a given number of points that built in the weighting. The scores were then just averaged. Ultimately, in both cases, a percentage of possible points was calculated and a grade determined.

An equitable grading analysis of this practice leads us to believe it contributes to **inaccuracy** because, at least in our case, the scores did not reliably relate to particular outcomes. At the end of the course we were unable to divine which particular capabilities our students had or had not developed. Not only were

we unable to provide accurate information about individual students, we were also unable to use our results in an effort to improve our instruction.

A better approach seems to be to have scores or grades that are readily relatable to particular outcomes/ capabilities. If this were the case, we would be able to inform students of their particular shortcomings and we could relatively easily identify any outcomes where numerous students had difficulty. In both cases it would be useful to have scores that relate to single outcomes or sets of outcomes.

Once all scores relate to outcomes we can analyze them for both student grading and for instructional improvement. If we use binary grading, as in Berns (2020), we merely decide whether any particular outcomes are more important than others and combine them in an additive manner. For example, if we have six individual outcomes and a cumulative outcome, we might treat the putting-it-all-together outcome as double the others and have eight total "points". We could then decide what constitutes the various grades. Perhaps we feel that more than half the outcomes must be demonstrated to get a C, i.e., five of the eight points. Then, six points might demonstrate B work and seven or eight would be considered an A. We might add into the mix that the putting-it-all-together outcome must be passed to get an A or B regardless of the independent outcomes.

Alternatively, we might use a four point scale for each outcome thinking of them as unknown[1]/failing (F), marginal (D), satisfactory (C), good (B), and excellent (A). Combining them might be something like:

- A—all A's and B's with mostly A's
- B—all A's, B's, and C;'s with mostly A's and B's
- C—90% of outcomes were D and above with as many A's and B's as D's
- D—at least 60% of outcomes were assessed and most of them were C and above

Note that this example is an example only, not a recommendation. Each teacher or department/school will need to decide how grades would be determined after careful consideration.

Conclusion

Given the **accuracy** and **bias** issues associated with traditional grading system, and the added issue that traditional grading offers us little data for evaluating and improving our instruction, it behooves us to consider an alternative. Feldman (2019) offers an alternative that is well-presented in his book and supported by research (which the book identifies). The discussion below suggests a process for revising instructional and grading practice.

4. Revising Our Instruction Towards Equitable Grading

For the reasons documented above, we had come to the conclusion that our prior grading practices were inequitable and needed to be changed. Knowing change is required and actually implementing change, however, are two very different things, and we are still working on the latter. In this section, we further discuss the steps we have taken towards equitable grading. We reiterate that this is a work in progress for us, but we feel our efforts are at least worthy of discussion.

Focus on Student Outcomes and Capabilities

In order for our grades to be accurate, they must reflect how competent a student is in the outcomes of the course. Therefore, the first step we took was to look closely at our classes to identify the student outcomes and produce assessments that would measure if a student has the desired capabilities or not. While this has been encouraged practice for some time, it was not until we thought about outcomes in terms of equity that we really focused on this step.

---

[1] At the end of the course, all assessments must be accounted for. This is the situation where a zero might be appropriate.

We suggest Wiggens & McTighe's (2005) notion of "backward design" as a good process for designing instruction. It is very similar to top-down design in programming or ends-means analysis from AI. The suggested process is composed of three steps to which we have added a fourth.

1. Identify desired results.
2. Determine acceptable evidence.
3. Plan learning experiences and instruction.
4. Devise an appropriate grading scheme based on equitable grading practices.

The desired outcome of a first programming course, for example, might be that students be able to write complete but fairly straightforward programs in a variety of problem contexts that require the use of any or all of: sequence, selection, iteration, modularization, and file I/O. Acceptable evidence for the capability would be to write a program that would demonstrate all the desired capabilities. East and Berns (2018) suggest also adding the bench-mark exam (Simon, et al, 2016). Before learning experiences and instruction can be developed we would need to further refine our overall goal for the course. This is where our experience in computer science is helpful as we decompose that large goal into subgoals or smaller capabilities. For each of those we would need to explicate outcomes/capabilities, determine acceptable evidence, plan learning experiences and instruction, and determine how grading in each of them would be integrated into the course grading scheme.

To continue the programming example, the breakdown of capabilities needed to achieve the overall goal might include: 1) using the development system specific to the course, 2) programming with input, output, and assignment statements (IPO) along with variables and common numeric and string operations and functions, 3) using Boolean variables and operators to develop expressions to achieve particular goals in varied contexts[2], 4) using selection (and IPO), and 5) using repetition (and IPO). The use of file I/O and modularization might be addressed as students practice applying basic knowledge to slightly larger problems.

For each of these units we would next need to determine the acceptable evidence of learning: that is, develop assessments. Doing so might require that we decompose further to identify finer-grained capabilities ultimately. We have mostly focused on competency demonstrations that were essentially quizzes addressing the capability and constituent capabilities. For example, the "using selection" assessment might address tracing a selection statement, creating an appropriate conditional or placing actions in the appropriate branch of a selection statement, and creating a snippet of code that uses input to produce output based on processing involving a selection statement.

Development of instruction follows relatively naturally once specific capabilities and their assessments are identified—devise activities that work toward those capabilities. With programming there are many examples of each activity that can be used for demonstration, practice, and/or assessment. The assessments can mirror the instructional/demonstration activities and learning/practice activities. Additionally, the students know what they need to be able to do to "pass" the assessments.

Managing the instruction from an equitable grading perspective involves a number of decisions. We suggest:

- Don't grade homework—separate learning and assessment activities. That necessitates developing an alternative mechanism for giving students corrective feedback when necessary.
- Offer students multiple opportunities to demonstrate competence. That necessitates determining any prerequisite structure between units and limiting the number or time frame for re-takes (and perhaps other situations, but equitable grading practices need to be kept in mind—no punishment).

---

[2] We believed that numeric and string contexts were familiar to students but more specific experience with Booleans was needed for competent use of selection and repetition. See Hughes & East, (2014)

- Omit all behavioral grading elements. Don't include attendance & participation, extra credit, etc. Do not penalize late work. If a points-based system is used, do not use zeros for missing assessments.
- Develop an equitable grading scheme for overall grades. Doing so necessitates re-examining what is included and how the ultimate grades are determined.

Revise the Overall Grading Scheme

For many of us (at least for we two), the course only comes together when we have detailed how the final/ overall grades will be determined. For the most part, the use of equitable grading practices has simplified that task. Since only assessments are graded—not homework, not behavior-related activity—there are fewer grading elements to consider. We must, however, account for retakes, determine whether assessments are binary or points-based, determine relative importance of assessments, and develop the mechanism for combining elements to determine a final grade. Additionally, we need to be able to keep students informed of their status with respect to individual outcomes and the overall course grade.

Normally we use can use grade book programs or spreadsheet application to accomplish both those goals. We are unsure as to whether grade book programs will work here but we know spreadsheets will. During the course performance and progress are indicated by the unit assessments only and depends on how many assessments were taken and passed. A mark (+ or -, 0 or 1) indicates performance; we need only determine how grades are determined. Perhaps, at the end, 5 of 5 are needed for an A; 4 / 5 for a B; 3 / 5, a C; 2 / 5, a D; and 1 or 2, an F. By tracking the number of assessments taken and the number passed we can readily calculate a current grade. Individuals can have an individual assessments-taken count and thus have an accurate picture of the current grade. The next step would be to determine how to combine the individual skills assessment grade with the final exam.

We suggest there be substantial consideration as to what actually should constitute A or B or C performance. Of course it is essential that we avoid thinking in any way other than actual academic capability—not rewards for effort, not penalties for cheating, not attitude (positive or negative)—only the identified outcomes.

5. Some of Our Experiences

To make sure our grades reflected only student capabilities, we had to revisit the things we were asking students to do and recognize that some things were meant to help them learn something, while others were meant to assess their learning. In our old system, these two different activities all went into the final grade. In the new system, our "homework assignments" were not graded at all, and were rather assigned simply to give students practice for the assessments they would be taking. The incentive for completing these assignments was an improvement in their ability which would hopefully be reflected in the assessments they would take.

To take into account the differences in when a student might learn the material, we allow students to retake assessments until satisfactorily completing them up until the end of the semester. For some assessments, we provide students with feedback and then have them revise and resubmit or retry the assessment. In other cases, we present them with a similar assessment but with different details.

Once we started using outcomes- or capability-based assessments and separated learning activity from assessment activity, it was a relative simple step to remove behavioral or bias-prone elements of grading.

To avoid having the weight of a "0" be disproportional, we have moved away from a 100 point grading scale and instead have used a scale with only a few values. In some instances, this have involved a simple pass/fail grading scheme. Some of our colleagues have found it more comfortable to have a little finer granularity and have used a 3, 4, or 5 point scale. The key idea in all these systems, of course, is to have the range for each grade be equivalent, instead of 60 values for an F as done in our traditional grading scheme.

As mentioned earlier, if we look to get an accurate grading system, we need to make sure that a 0 is reserved only for cases when a student has no competence in an area by the end of the semester. We have tried to move away from students thinking they have "failed" an assessment and instead think of it as an "in progress" assessment. We hope that this way of thinking will keep up their motivation to work at learning the skills we desire. Furthermore, of course, they are also allowed to retake the assessment to bring their score up. As a result of these policies, the only time a student will really have a "zero" is when they have not demonstrated any ability to perform an outcome by the end of the semester.

Another time-consuming part of our transition has been the development of assessments and learning activities for the outcomes we have identified. We of course want our assessments to be an accurate reflection of the outcomes, which by itself is no easy task. Furthermore, we need to build activities that can help students learn the skills required to complete the assessment. We have found this is not as simple as just taking our old homework and quizzes and calling them learning and assessment activities. Instead, for many cases it has required careful thought about a particular outcome, how we can assess it, and how we can help students learn it. This is of course not always a linear process, but when complete we should have clear plans about how to proceed in a class.

One area in particular we are working at is finding ways to provide students with feedback during their learning. In our old system, we used homework scores and the occasional written comment as our feedback mechanism. In our new system, however, we do not grade homework and therefore need a different way to provide feedback. We have thought about approaches including providing worked examples, using auto-grading software, and having students perform peer reviews of materials. We are still working on which might be best or better in particular contexts.

While the move to equitable grading has not always been easy, we have found it to be quite rewarding in several ways. First, students seem to overwhelmingly enjoy the new approach and seem to appreciate things such as the ability to retake an assessment, the simplified grading criteria, the avoidance of "busy work", and the focus on student learning over point allocations.

A second benefit has been we spend less time grading homework and more time "teaching", a welcome relief from our old grading systems. We now use our time to address student questions and design better learning activities. Furthermore, we have virtually eliminated the uncomfortable scoring disputes that seemed all too common in our traditional points-based grading system.

6. Closing Thoughts and Next Steps

We are still working at honing our grading system to adopt the principles of equitable grading. As with most large changes such as this, we have encountered several difficulties along the way. First, there is a large time commitment initially to identify the desired outcomes for a course. Even though we had outcomes listed before, we found doing a complete redesign of grading where the only thing we could assess was what was given as an outcome resulted in us revising some outcomes, removing others, and adding new ones. We imagine this will be an iterative process which may require several attempts before reaching a stable state.

We wholeheartedly encourage you to consider the ideas here. We expect it will substantially change, perhaps revolutionize, your practice and your attitude toward teaching and learning computer science.

References

Andrew Berns (2020). Experiences with Binary Grading Across the Curriculum. In *Proceedings of the Fifty-First SIGCSE Technical Symposium on Computer Science Education*, 1152-1157.

J. Philip East & Andrew Berns (2018). The Benchmarking Programming Exam and SOA in Introductory Programming Courses. In *Proceedings of the 51st Midwest Instruction and Computing Symposium*, April, Duluth, MN.

J. Philip East & J. Ben Schafer (2005). In-person grading: An evaluative experiment. In *Proceedings of the Thirty-Sixth SIGCSE Technical Symposium on Computer Science Education*, 378-382.

J. Philip East (2000). Teaching, assigning, and grading: Continuing efforts to enhance learning. In *Proceedings of the 33rd Midwest Instruction and Computing Symposium.* April, St. Paul, MN.

J. Philip East (1999). Providing feedback to CS students: Some alternatives. Presentation at *NECC '99*. June, Atlantic City, NJ.

Joe Feldman (2019). *Grading for EQUITY: What It Is, Why It Matters, and How It Can Transform Schools and Classrooms*. Corwin, Thousand Oaks, CA.

Stephen Hughes & J. Philip East (2014). Thinking about conditional thinking. In *Proceedings of the 47th Midwest Instruction and Computing Symposium*, April, Verona, WI.

Simon, Judy Sheard, Daryl D'Souza, Peter Klemperer, Leo Porter, Juha Sorva, Martijn Stegeman, & Daniel Zingaro (2016). Benchmarking introductory programming exams: Some preliminary results. *ICER'16*, September 8-12, 2016, Melbourne, Vic, Australia.

Grant Wiggins & Jay McTighe (2005). *Understanding by Design*. ASCD, Alexandria, VA.