Grading for Equity: Background

J. Philip East, Andy Berns, J. Ben Schafer
University of Northern Iowa

This document was developed to provide introductory information about Grading For Equity to participants in a workshop at SIGCSE 2021 in an effort to maximize workshop time on discussion and doing activity rather than presentation activity. Thus, examples and rationale arguments are present to a lesser degree than might be anticipated.

The Process

Feldman (2019)[1] says grading practices should be accurate, bias-resistant, and motivational. These "pillars of our vision" provide the basis for analyzing grading practice and are actually mutually reinforcing as equitable grading practices are generally supported by all three. Our discussion below of Grading for Equity (GFE) identifies the process for achieving GFE, describes the guiding principles (mathematically accurate, bias-resistant, and motivational), explains how they apply to traditional grading practices, and identifies suggested practices that will make grading truly equitable. A core idea of GFE is that learning activities need to be separated from grading activities.

Identify Outcomes

The critical first step to implementing GFE is the identification of expected learning outcomes from the instruction. Outcomes must be amenable to direct assessment. It is reasonable that some be conceptual knowledge assessed by asking for definitions and examples. The bulk of the outcomes, however, would likely involve knowing when and how to apply knowledge to achieve some task.

For example, in programming, we want students to be able to "use counting loops". Certainly knowing the syntax and semantics of a *for* loop in the language being used is desirable but we probably do not need to worry about assessing such "knowledge". The desired outcome is that students be able to use such a loop when appropriate. Thus, asking students to identify when such use is and is not appropriate (and why that is so) and to develop code for an indicated problem seem a more appropriate assessment. One might also wish to ask students to trace, debug, or describe the purpose of sample code, as those capabilities are necessary to truly be able to "use" counting loops. Including such component outcomes is useful for identifying potential student difficulties and/or to more comprehensively identify the extent of the outcome.

Clearly identified outcomes are absolutely essential to the development of assessments that can be used for equitable grading. Computer scientists inherently understand this concept

[1] *Grading For Equity: What It is, Why It Matters, and How It Can Transform Schools and Classrooms*. Corwin, 2019. See also https://gradingforequity.org/.

though they may not have applied it to instruction and grading. To develop a software system (or a piece of it) requires the careful delineation of specifications. Without carefully identified specifications, correctly functioning programs cannot be developed efficiently, if at all. The same is true of instruction and grading.

Develop Assessments

Once outcomes have been identified, assessments need to be developed for each outcome. Initial efforts at assessment development will be less than perfect, perhaps substantially so. As outcomes lead to assessments, however, assessments lead to instruction and learning activities. Without knowing how students will demonstrate outcome capabilities, one cannot design the instruction and learning activities needed to allow students to demonstrate competency. So, a good idea of the ultimate assessment(s) is needed even if the actual assessment(s) are not produced at this time. The “good idea” of an assessment must be explicit. An idea in one’s head is too ephemeral to guide the development of instruction.

An alternative to actual, specific assessments for outcomes might be a comprehensive study guide. The study guide would include exemplars of all the desired capabilities for each unit of instruction (a unit is a set of closely related outcomes, e.g., counting loops, the instruction cycle, etc.). The study guide can be shared with students to indicate what they need to be able to do. It can also be used to guide outcome assessments or even as the source of assessments.

Prepare Instruction

With a clear vision of desired outcomes, instructional activity can be planned. The instructional activity includes developing lectures or other knowledge transmission material (e.g., videos for a flipped classroom), planning activity to clarify student understanding (e.g., code walkthroughs, responding to student questions, peer review of work, etc.), producing the learning activities (homework), and finalizing assessments. At least the first, and perhaps several other, instance(s) of instruction preparation will likely involve cycling back and forth between outcomes, assessments, and instruction to ensure their correspondence.

A part of instruction is providing feedback and grading. In traditional instruction, these activities are closely related. That becomes much less so with GFE. Thus with GFE, instructional planning needs to include the provision of feedback separate from grading as well as identifying how final grades will be determined.

Devise a Grading Scheme

The process for determining final grades is an essential element of instruction and GFE. It needs to be clear and transparent to the students while embodying the precepts of GFE. Some instructors use the grading system to assist in developing course plans. We discuss grading scheme possibilities below after explaining major elements of GFE.

Grading for Equity: Pillars and Practices

As noted above, GFE has three essential qualities or pillars. Grading should be **accurate** in that it accurately and consistently reflects a student's capability and matches the instructor's professional judgment of capability. Furthermore, grading should reflect the capability that exists at the end of instruction, not the beginning or some other in-process point. Second, grading should be **bias-resistant**. The grading practice itself should not provide advantage or disadvantage to any particular subset of students--those inclined vs not inclined to speak up, those with much vs little prior background with the material, those with good vs poor academic skills and background, those with good vs poor self-concept, those with lots of vs limited study time, those whose highest or lowest priority in school is your course, those with no or myriad personal issues, etc. Again, we should not knowingly allow any advantage or disadvantage in any such respect to exist in our grading practice. Finally, grading should, as much as possible, **motivate** students rather than discourage or demotivate them. While few of us are experts or even knowledgeable of motivation theory and research, we should, when presented with such knowledge revise our grading practice to be in line with that understanding.

The difference-making aspects of GFE are the grading practices. Feldman recommends altering or discontinuing some practices and adopting others. Our discussion below identifies problematic practices and suggested practices, all with rationales for their use or change/non-use in terms of the three pillars--that grades should be accurate, bias-resistant, and motivational.

Problematic Practices

The Percentage Scale

Traditional practice usually involves awarding points for various elements of student work, totalling and/or weighting scores or categories of scores, and then assigning grades using percentages such as 0-59: F, 60-69: D, 70-79: C, 80-89: B, and 90-100: A. Several issues arise from this practice.

The most obvious issue is the disparity in the grade ranges. Though we tend not to think about it this way, the grading scale above suggests there are 60 gradations of failure and 40 gradations of success or, going further, 60 ways to get an F and 10 ways to get a D (or C or B or A). From that point of view it doesn't seem to be a position we would really espouse. It also is likely not **accurate**. Surely it is not the case that there are 10 (or 11) ways to get an A but 60 ways to get an F. To be accurate when we grade, all the grade ranges should be the same size.

Using the percentage grading scale is not **motivational**. A very low score is hard to overcome when the failing range is so much larger than the others. While some students might be able to overcome such a low score most will not. Students will realize it, making the hoped-for motivation to "work harder" actually end up suggesting to students that getting a good grade is hopeless. One way to overcome these issues with the percentage scale is to have a minimum grade of 50, thus, making all the ranges the same size.

Even if we equalize the grade ranges, however, at least one issue remains. In our experience, it is difficult to consistently discriminate between 10 gradations of work for any grade range. Can a grader consistently and **accurately** distinguish scores of 96 and 95, both of which are an A rather than A+ or A-? We suspect not. These minute differences in points awarded that can be hard to defend to students end up motivating students to argue over points rather than motivating them to learn the material. One way to reduce the gradations of the various grade ranges is to have a 15 point scale (0-14 probably) that goes from A+ (14) to F- (0). The gradations could be further reduced to a five point scale (0-4 for F, D, C, B, A). We have used both two and three point scales. A two point scale could indicate demonstration (or not) of competency and a three point scale could indicate unsatisfactory, satisfactory, and good performance.

Weighted Averages

Another problematic aspect of using the percentage scale is when different elements of course work are weighted to produce an overall score/grade. For example, we might include homework, class participation (including attendance), quizzes and exams, and extra credit in our grades. The tables below (adapted from Feldman, 2019) show why we believe this can be problematic.

In our first example (the table immediately below) two students were subject to the same category weightings, Student 1 is a very conscientious student who did almost everything that was expected but had difficulty doing quizzes and exams. This might have been because the student had test anxiety or because the student while doing all the work did not really learn sufficiently to do well on quizzes and exams. Student 2 did very well on the quizzes and exams but perhaps only did homework and attended class when the material was new to her/him. The students got essentially the same scores and presumably the same grade. Should they have gotten the same grade? Does their learning appear to be the same? (We suggest not.)

Category	Category Weight	Student 1		Student 2	
		Score	Weight	Score	Weight
Homework	30%	80%	.24	60%	.18
Quizzes/Exams	50%	60%	.30	96%	.48
Participation	15%	100%	.15	60%	.09
Extra Credit	5%	100%	.05	0%	.00
Weighted Percentage		74%		75%	

In the second example (below) we see two different weightings of the point categories. This might occur with two different faculty or with a single faculty member who changed the weightings from one semester to the next. The student scores were exactly the same, only the weightings changed. In Class X a variety of activities contribute fairly strongly to the overall grade. In Class Y, quizzes and exams make up almost the entire grade. In this case the scores differ by at least one grade level. Is it desirable for a student this semester to get exactly the same scores as a student last semester but receive a lower grade?

Category	Student Score	Class X		Class Y	
		Weight	Score	Weight	Score
Homework	80%	30%	.24	5%	.04
Quizzes/Exams	60%	40%	.24	85%	.51
Participation	90%	20%	.18	5%	.045
Extra Credit	100%	10%	.10	5%	.05
Weighted Percentage		76%		64.5%	

These examples allow us to see that the percentage scale and weighting of scores can be problematic and lead to grades that do not actually result in the grades we might wish. When we have time to consider particular examples like this outside our normal hectic days of syllabus preparation, instructional planning, teaching, and grading we can reflect on the **accuracy** of our grading schemes.

The Zero Score

When you see a zero in the gradebook, what do you know about the capability of the student with the zero? … Several possibilities exist. … Often the zero means nothing has yet been submitted by the student. It might mean that the student work appears to be copied from another student or be the object of copying. Occasionally, a zero is an accurate reflection of the student's actual capability. We suspect that most uses of the zero are for the first two of the above possibilities. Thus, the score/grade used is not an **accurate** indicator of student capability.

When used in conjunction with the percentage grading scale the zero score becomes highly problematic. Not only is it likely inaccurate in itself, it makes it nearly impossible to provide an accurate indicator of overall capability. A zero when combined with a perfect score of 100 averages out to a score of 50 which is still a failing grade. Do we really want to say that a student with no understanding of concept 1 and perfect understanding of concept 2 should

receive a lower grade than another student who has poor (D level) understanding of both concepts?

Using Grades to Reward (Attendance, Participation, Extra Credit, etc.) or Punish (Lateness, Cheating, etc.) Behavior

Traditional grading practice often uses grades to encourage or discourage certain behavior in students--almost always with the best of intentions. Appropriate student behavior will be beneficial to learning. It can also provide poorer students an easy way to lift their grades. Following class rules will prepare students for "work". Academic rigor includes behavior in addition to content. All students should be treated the same (regardless of their context). We "know" this because it is the way we were taught and have taught for tens of years.

But, students are not all the same. Some possess substantial course-related background knowledge, some have none or perhaps have misconceptions. Some have good academic skills, some poor. Some have cultural mores, personality traits, or gender experience that encourages or discourages speaking up in the presence of the teacher. Some have few or no life issues affecting the time and energy available for school work while others must work, care for family members, travel to school, have medical conditions, spend more time studying, etc.

Behaviors like those above reflect nothing about student capability with respect to course outcomes. Including behavior as part of a grade automatically makes grades **inaccurate**.

More importantly, including non-academic behavior like those above when grading will inadvertently introduce **bias** into the grades. Grading these behaviors will advantage some students and disadvantage others. Students with better academic backgrounds and skills will automatically engage in the desirable behaviors while those with poorer backgrounds may not see their utility. Including class participation will advantage extroverts and those with positive self-efficacy while disadvantaging introverts and those with less academic confidence. Culture and life experience will cause disparate behavior in students. Proper academic behavior may be viewed or experienced as "acting white". Some cultures' views of student-teacher relations may discourage questioning the teacher. Similarly, finding answers on-line may be perfectly in tune with some cultures. Less able students (with respect to academics in general or content) may not be able to consistently get work done on time. Generally speaking it is the "better" students who benefit most from including academic behavior in grades--not what most of us really intended. It seems to us that once confronted with these ideas it is unprofessional and immoral to continue including "appropriate" academic behavior in grades.

Finally, including these behaviors in grading **demotivates** most students. They recognize that the behaviors say nothing about their capability, particularly those that actually develop/possess the desired capability. When faced with behavior deterrents (penalties for cheating and lateness), students are likely to become even less engaged.

Many of us include the grading of behavior in an effort to train students to follow the rules of workplaces. GFE would not disallow that. However, it would require that such behavior be included in the desired course outcomes. GFE requires that all outcomes be assessed directly

and have instruction designed to develop the capability. It seems likely that including workplace behavior as an instructional outcome would necessitate surveying industry to identify appropriate and inappropriate behavior and the importance afforded each in order to accurately reflect the desired capability then developing instruction and assessment that would develop it and produce evidence of success.

Using Group Work Products in/for Individual Grades

GFE notes that applying the grade for the work of a group to individual students is **inaccurate**. It seems likely that it is also demotivating as some students will be inclined to allow others to do the work and students doing the work will be frustrated. On the other hand group work can be a very effective learning tool. The key to using group work as an instructional strategy is to carefully identify the desired outcomes/capabilities and develop *individual* assessments for them. Clearly doing so will require substantial rethinking of grading practice and should be done with careful consideration[2].

Grading Homework

The purpose of homework is to provide students with an opportunity to learn. Providing feedback on homework is good, but actually "grading" the homework is inconsistent with the goal of enhancing learning. For example there are at least three different examples of student learning and homework: 1) a student already has the desired knowledge/capability (so doesn't do the homework), 2) a student develops the desired knowledge/capability while doing the homework, and 3) a student requires doing the homework, perhaps poorly, and receiving feedback on it to develop the desired knowledge/capability. All three students developed the desired capability but would generally receive widely varying scores. Their grades should be the same but would not be under traditional grading practice and would, therefore, be **inaccurate**.

Additionally, grading homework can introduce **bias** into grading. Bias could arise from a variety of differences in students. General academic background/skill will (dis)advantage some students. Familiarity with course content will (dis)advantage some students. Time available to students as a result of issues with work, study, family, psychological well being, etc. will (dis)advantage some students.

Logically, the existence of bias and inaccuracies arising when homework is graded would lead to demotivation of students. It is likely, however, that homework grading is so ingrained in educational practice that the issue may not occur to many students. However, some students will recognize the reality and become less motivated. Additionally, the use of GFE practices will gradually spread, more students will become aware, and **motivation** will become more of a problem.

---

[2] When Philip discovered *Grading For Equity*, he was teaching a course in which almost all student activity involved group work. The plan was to apply scores on the group work to individual grades. He became mired in a quandary--it was too late to change course plans but following the plans would yield inaccurate grades and seemed unethical as he had quickly bought into GFE.

To be equitable, instruction needs to separate learning and assessment activities. Substantial thought as to how to accomplish that task and still provide feedback to students will be required. Key to doing so is the identification of desired outcomes/capabilities and the development of appropriate assessments.

Suggested Practices

Feldman (2019) suggests the modification or replacement of the above problematic practices with others that are more equitable--accurate, unbiased, and motivating. Doing so requires the clear identification of desired outcomes/capabilities and the development of appropriate assessments of those outcomes as a part of instructional design/planning. Suggested practices are discussed below similarly to Feldman's presentation.

Practices Enhancing Accuracy

Using practices that are mathematically accurate should ultimately enhance both motivation and lack of bias. The practices below are primarily motivated by accuracy.

Avoid Using Zeros

Students almost never have absolutely no capability with a given outcome. Therefore to be accurate zeros should not appear in the gradebook. That likely means having some other indicator of missing work. It will also entail devising some means of informing students of their current grades. Several discussions below provide insight into practices that help avoid the use of zeros.

Grading Scale

Feldman (2019) suggests two alternatives for a grading scale. We describe those and suggest some additional alternatives.

For accuracy's sake it is important that the grading scale have equal-sized ranges for all grades. One way to accomplish this is to use a minimum grade. For example, the F range would be 50-60% rather than 0-60%. That prohibits a very low failing grade from having a larger influence on the overall grade than a low D or C would have. Alternatively, Feldman suggests using the 0-4 point scale instead of the 0-100 scale with a minimum score of 50. The 0-4 point scale lets us think in terms of letter grades and is amenable to the application of rubrics or conceptual categories such as:

- Minimal, Marginal, Satisfactory, Good, Excellent
- Inadequate, Approaches Expectations, Meets Expectations, Exceed Expectations, Demonstrates Mastery
- No Reasonable Attempt, Little Skill, Some Skill, Substantial Skill, Exceptional Skill
- No Evidence re Standard, Some Evidence re Standard, Key Gaps in Meeting Standard, Met Standard, Exceeded Standard
- No Key Elements, Some Key Elements, Essential Key Elements, Most Key Elements, Complete re Key Elements

For those more in tune with letter grades equivalences the 15 point scale (i.e., F-,F,F+, …, A-,A,A+) or 13 point scale (no F- or F+) can be used.

To us, the five point scale allows or requires the exercise of professional judgment when grading. We believe that all grading is subjective (since the instructor decided on the assessment) and getting rid of the 0-100 point system recognizes the need for professional judgement. We have experimented with scales other than the 5, 15, or 13 point scales that also entail professional judgement. One possibility is the binary scale--implying competency demonstrated or not. In some cases student competency was uncertain and the student was asked to come in and discuss the assessment result. A 3 point scale can allow differentiation between levels of competency, e.g., did not meet expectations, met expectations, and exceeded expectations. In using professional judgment and the smaller (2 to 5 point) scales we have had no students argue about the score/grade assigned (perhaps after hearing our rationale).

More Recent Achievement and Retakes

As alluded to in some previous discussions, the ultimate goal of instruction is the development of desired capability. It should not matter at what point during instruction the capability was developed so long as it had been demonstrated at some point. An **accurate** grading system will reflect student capabilities developed during the course, not just those evidenced at the initial assessment. An **unbiased** grading system would allow those with less background or academic skills more time to learn, i.e., another chance at the assessment.

Assessments occurring later in courses sometimes examine concepts/capabilities that combine or include several early concepts/capabilities, in essence, reassessing the earlier work. However, some outcomes might not be doubly covered in this manner or it might be advisable to catch the difficulty prior to the later cumulative assessment. In this case, a retake is appropriate. We know all students don't "get it" the first time and often tell students, "If at first you don't succeed, try, try again". If we believe this we are obligated to provide students additional opportunities to succeed. Both these approaches will require some upfront work on the part of instructors. For capabilities covered by cumulative assessments the instructor would need to be aware of such relationships and prepare a grading scheme that maps out the connections and notes improvements in performance. For retakes, multiple versions of the assessments must be developed.

Instituting a grading scheme that allows later work to revise earlier assessment results ends up being **motivational**. Students no longer experience the hopelessness of early failures lowering their final grade. They recognize that it is normal to not get it the first time and begin to develop a growth mindset.

Benefits of GFE

The result of using practices in tune with GFE should be that grades are more accurate, less prone to bias, and motivational to students. But, in our view, GFE is more than just a system of grading. It is an instructional design system that affects both students and instructors.

At the center of GFE is the initial step of identifying the desired learning outcomes--what students should be able to do. Moving from instruction intended to cover a body of knowledge to instruction that assesses actual capabilities allows instructors to place the focus of learning on student capability instead of points needed to pass (or get the desired grade).

As test-driven programming clarifies problem specifications for a developer/programmer, GFE's assessment-driven design clarifies instructional goals for the instructor (and for students). Then the development of assessments leads naturally to the development of better instruction since the goal is clear to the developer/instructor. The careful identification of desired student outcomes that can be reliably assessed builds confidence in the instruction.

No longer including behavior in grading makes grading more equitable. It also makes the grading and instruction more culturally responsive. The advantage or disadvantage built into traditional instruction arising from student background no longer exists. This is a relatively easy first step toward being more culturally responsive (though we will wish to eventually take additional steps as we gain insight into the issue).

One of the more noticeable advantages of GFE arises from not grading homework and having assessment retakes--we no longer need to worry about copying on homework. Students soon learn how much of the homework needs to be done to be able to demonstrate their competency with instructional outcomes and, with assessment retakes, they are not penalized as they come to this realization. Since the homework affects only learning, not grades, it does not matter whether the student used the work of others in their learning. This saves time and energy on the instructors part and removes a whole class of negative interactions between students and instructors.

Similarly, with the emphasis on demonstrating learning rather than accumulating points, students no longer argue over points or the grading of their assessments. We have used competency/mastery demonstrations/quizzes for many outcomes and have had essentially zero arguments about our judgement of student capability. (We are occasionally unsure of student understanding and ask students to explain it.)

The time previously devoted to grading homework is replaced with activity providing more professional gratification. Students still need to know if their work was correct which can be done with a combination of autograders, assignment keys, class discussion of instructor or student solutions, etc.

Transitioning to GFE

In his CSEd podcast[3] Feldman says a piecemeal approach to GFE is plausible. We think that might work if the piece(s) selected are substantial and consistent. It seems hard to explain the change to students in an effort to get them on board if the practices that are used seem incompatible.

[3] https://sites.duke.edu/csedpodcast/2021/02/15/season-2-episode-4-grading-for-equity/