Why grading on a curve is bad and how to make open book exams. C. Wieman 3/20/30

Response to request from David Goldhaber-Gerson
"I would really welcome your sharing what is known on this when you have time. My sense was also that it is detrimental to create a sense of competition between students rather than a feeling that each one should learn as much as possible (and show that) against a fixed standard. That said, when teaching a course for the first time it can be difficult to guess how well students will do on an exam. I grappled with that this quarter, after a midterm on which students did more poorly than I expected (and I decided to make an adjustment to grades accordingly.)

Even more urgently, your advice on writing open-book, open-internet exams would be valuable to me and likely others. We've been rewriting the exam for Physics 63 for "search engine de-optimization", but we may be missing key insights."

There are two fundamental problems with grading according to a fixed distribution ("curve").
1. There are basic ethical questions involved. The justification for use of a distribution like this is based on two flawed assumptions, that the students and the quality of instruction are the same every year. On short time scales, at all the universities I know of, which is probably about 20, barring some obvious change in admissions or requirements, there is a high degree of reproducibility in the incoming student populations on average in intro STEM courses. This has been checked with multiple tests. Faculty often perceive differences that are not supported by tests. However, as Mark notes, this consistency of population is not true on time scales of several to many years. The quality of instruction and hence resulting learning can vary on much shorter time scale. There are a large number of studies where switching from traditional instruction to research-based "active learning" improves performance on most outcome assessments that are kept consistent across offerings. This is true when the switch in methods involves the change in instructor or the change in teaching method by the same instructor. At both Colorado and UBC, I had to deal with several conflicts over the fact the department had set grade distributions, but the instructors who changed their teaching through the program I ran were seeing that their students could now perform the equivalent of a full grade higher on essentially identical exam questions compared to previous years. They understandably felt that it was morally wrong to stick to a set distribution and thereby fail a student who, had they done the same work on the exam in the previous year, would have received a low B, or to give students Bs when they were performing what was A level work in past years. In addition, the data typically show differential impacts of the teaching methods and grading policies on various demographic groups, with grading on a fixed distribution likely leading to relatively lower performance by historically underrepresented groups, although in most cases this is only one of several factors are changed. So that is the ethical argument against grading according to a fixed curve.

2. Probably more importantly is the impact of this grading policy on learning. Almost everyone is personally aware of the learning benefits of discussing material with peers and trying to teach to others, and this is an area well studied by cognitive psychologists. (See *ABCs of How We learn* for concise readable summary). What they have seen is that the brain processes material differently when it is called upon to explain it to others, or even when it just thinks will be having to explain it to others. The different cognitive processes this induces results in learning beyond what can be achieved by solitary thought. Those of us who pose challenging clicker questions to our classes are familiar with this phenomena, as it is evident in students answering individually and no one in a group getting it correct, but then after the group discusses the question, the entire group gets it correct. In a study my group did, we find that they did not just get it correct during class, but it sticks. When tested individually some

days later, nearly all students still give the same correct answer as the group figured out. As described in ref (Science. 2009 Jan 2;323(5910):122-4. doi: 10.1126/science.) these peer discussions actually result in greater learning than does an explanation by expert instructor. It is comforting however that, as shown in ref the best learning is produced by peer discussion followed by instructor explanation.

What does this have to do with the grading distribution? When a course is graded on a fixed curve, it clearly inhibits student discussion of the material and this "social learning". Now students are not competing to master the material, they are competing to do better than their fellow students. So anything they do that might help their fellow student understand the subject better, or figure out a homework problem, is directly hurting their grade. That downside is more apparent to them than the modest increase in their own learning that may result. Students are quite rational in their analysis of this cost of helping fellow students, and it is generally reflected in their subsequent behavior. When courses are graded on a curve there is less interaction between students, hence less learning. We have measured that there is also generally overall lower motivation. That decrease in motivation likely arises from the achievement goal being changed from "master this new and potentially interesting subject" to "do better than other students on whatever arbitrary grading hoops we have to jump through". This does not lead to a universal decrease in motivation. There are some students that are quite motivated by the sense of competition, but our data indicates that there are more that respond negatively to it and are demotivated by that competitive grading, at least in intro STEM courses. URM populations tend to be predominantly in the latter category.

As a practical question, how does one set fixed mastery levels, when it is almost impossible, particularly in teaching a course for the first time, to create exams, and sometimes problem sets, that will produce a given desired distribution (such as top students get above 90%, solid students with moderate weaknesses 80-90 %, etc.)? It improves consistency by having very detailed learning goals and linking test questions to them, but, as discussed below, this almost always gives me exam averages in the 70%s, which is usually a bit low.

I have found a way to "cheat" that works pretty well to send the right messages to the students and allows me to compensate for the shortcomings of my exams. I set the grading based on the absolute standard percentages, and I have a lot of things that contribute to that, all being things I know will contribute to their learning: problem sets, quizzes or midterms, doing worksheets and answering questions in class, reviewing and reflecting on errors on psets and exams, as well as score on final exam. I am fairly conservative in how difficult I try to make my exams, so a smaller fraction than I think deserve As will get above 90%, but in giving the grading rubric in the syllabus, I say, "The course will be graded on this absolute scale and my goal is for everyone to get above 90% and get an A." "But, I will reserve the right to move the grade cutoffs lower at the end of the term, to raise the overall grades, if I think that is merited." And then, if I want more As and fewer Cs, I just move the respective cutoffs down. No one complains if I increase grades overall in this manner, and they still feel like they are being judged according to how well they master the material, not how much better then do than other students. Of course there is still some sense of competition, they are all interested in their midterm score compared to class distribution.

2. I already talked with David and gave him examples of how to create reasonable exams that are open book, notes, and internet, but for others if interested.
I start with a very detailed list (order 60 for course) of learning goals that are in operational terms, "what a student should be able to do". Then every question just maps onto one or more learning goal(s). The average score on my open everything exams always turn out to be in the 70's per cent range.
Some of these goals are being able to carry out a specific calculational procedure for some defined range of contexts. For those questions I just make computer graded multiple choice questions, but with

my own scenarios so cannot directly look up precise situation on internet. If students have basic understanding of what procedure to use and how, and just need quick reminder of details and constants involved which they know where to look up, I am fine with them learning to that level of mastery—matches my own. If they have no idea how to go about doing the calculation, they will never have time to learn it during the final exam. Students overwhelmingly recognize that and seldom even try.

The majority of the questions involve long answers, that, unlike the multiple choice, do not focus on a simple calculation or final number or equation. The solution explicitly involves explaining the process, reasoning, and decisions involved. Questions that require significant explanations in their own words of some phenomena or the justification of the structure of an equation or criteria for choosing different solution approaches are particularly useful for catching when students are collaborating on answers. It has happened but almost never, and if they are cheating it is clearly not very effective judging by the exam scores. On occasion, if grading issues require it, I will turn these long answer questions into very complex multiple-choice questions that call on them to make decisions and give reasons, but that is too detailed to get into here.