



Seattle

122nd ASEE Annual  
Conference & Exposition

June 14 - 17, 2015  
Seattle, WA

*Making Value for Society*

Paper ID #12746

## **Not all curves are the same: Left-of-center grading and student motivation**

**Dr. Joanna Wolfe, Carnegie Mellon University**

**Dr. Beth A Powell, Tennessee Technological University**

# Not all curves are the same: Left-of-center grading and student motivation

Joanna Wolfe  
Carnegie Mellon University

Elizabeth Powell  
Tennessee Tech University

## Abstract

Despite a substantial body of research that criticizes norm-referenced (i.e., “curved”) grading for fostering a competitive climate, the practice remains a staple in STEM education and is unlikely to change. One reason educational critiques of the practice may not have hit home is that not all norm-referenced grading is the same. There is likely a big difference between what we refer to as *left-of-center* grading, where exam means are in the 20 or 30 percentiles and a score of 40% can translate into an A, and exams where means are near 60% and a score of 80% translates into an A. This study tests the hypothesis that students will distinguish between different types of norm-referenced grading practices. One hundred and seventy seven engineering students at a private, research I university completed surveys asking about their perceptions of norm-referenced exams with means in the 20’s vs. those with means in the 60’s.

The results overwhelmingly show that students found exams with means in the 20’s—but not those with means in the 60’s—discouraging and as evidence of bad and uncaring teaching. Students receiving an “A” for exam scores in the 30’s were unlikely to feel proud of their accomplishment and were highly unlikely to feel that they had learned what the instructor expected. These same students, however, did feel proud when an “A” was based upon an exam score in the 80’s. Students were also more likely to consider cheating and were less motivated to study when the median score was in the 20’s.

Over 90% of students indicated that a primary purpose of exams should be to measure mastery of concepts, and nearly 80% indicated that measuring what a student had learned should also be a primary purpose. By contrast, only 12% of students indicated that “distinguishing exceptional students from others” should be a primary purpose. These results are at odds with the assumptions of *left-of-center* grading, which prioritizes distinguishing among different groups of students and only indirectly seems to measure a student’s mastery of course content or learning

## Introduction

In the course of interviewing students for a project on gender and interpersonal communication in engineering, we began to observe a trend of negative reactions to a common educational practice that we have come to call **left-of-center (LOC) grading**: exams with class means below 50 percent. Curious about this trend, we modified our interview protocol to systematically ask students to comment on the pros and cons of this practice. Over 60% of the women and 15% of the men we interviewed emphatically saw the negatives as outweighing the positives. This trend was particularly common among minority women, over three-quarters of whom described the practice as highly discouraging. The quotations below reflect some of their viewpoints.

We'll have like a 30 percent average [on exams]....When you take the exam, it makes you feel horrible. You come out of there like, "I answered a fifth of that right, at most." **It's sort of like, "Well, gee, what did I learn?"** (*Hispanic Female; elite private university*)

You don't feel like you learned it. I mean you get 50 percent on a test, and you get an A, **I mean that's horrible** because there were so many others that you didn't get right. (*African American Female; public research university*)

To me if I made a 30, **even if 30 is the highest grade in the class, I still failed.** I think **it's very demoralizing** (*White Female; public, non-research university*)

It's not necessarily about grades, because you can get a 40 percent and still get an A, so it's not really about the grade, but.... **you feel like you're failing even if you get an A on the exam.** (*Native American Female; elite private university*)

I ended up with over a 3.6 GPA, so I obviously didn't do that bad, but there were a lot of tests that I would end up like leaving in tears, frustrated...**It makes you feel like, "Why am I in engineering school? I don't understand what I'm doing. I'm not learning anything."** (*White Female; public, non-research university*)

As the quotes above indicate, tests in STEM classes frequently have means as low as 20 or 30 percent where a grade of 40 or 50 percent becomes an A. We term this practice left-of-center grading to distinguish it from exams in which the majority of students are able to complete the majority of problems. The practice is quite common: out of the 83 engineering undergraduates and alumni we interviewed, all but three had experienced the practice.<sup>1</sup> And, as we indicated above, our research has also found that female students are particularly troubled by left-of-center grading, suggesting that the practice may have major implications for the retention of diverse populations.<sup>1</sup>

LOC grading is a subset of norm-referenced grading. Norm-referenced grading, popularly known as grading on a "curve," involves grading students on the basis of their rankings within a particular cohort. It is typically contrasted with criterion-referenced grading, which involves comparing students' achievements with clearly stated criteria for learning outcomes and clearly stated standards for particular levels of performance.

Although most grading mixes norm-referenced and criterion-referenced components, there is a strong consensus among educational researchers that criterion-referenced grading is an ideal that should be aspired toward because it provides students meaningful feedback on their mastery of objective core competencies.<sup>2-8</sup> Criterion-referenced grading has been found to increase students' trust in the grading process,<sup>9</sup> increase use of effective learning strategies,<sup>10,11</sup> encourage intrinsic interest in what is being studied,<sup>4,7,10,12</sup> and discourage the counter-productive competition<sup>4,11</sup> that turns many students away from STEM fields.<sup>13</sup> In fact, ABET now specifies that programs should evaluate students based upon explicitly stated criteria.<sup>14</sup>

One reason we propose that students find LOC grading so frustrating is because it violates so many of the feedback principles of criterion-referenced grading. Because the range of scores tends to be narrow when means are very low, students do not receive meaningful feedback about what competencies they have and have not mastered. LOC grading measures students but does not provide information about learning. Since women are more likely than men to have a mastery-orientation (vs. performance-orientation) towards learning,<sup>15</sup> it makes sense that they find LOC particularly discouraging.

If the reason that students object to LOC exams is that they expect exams to provide feedback on what they have mastered, then we might hypothesize that students would be more positive about exam scores when the means are in the 60's. A student who achieves a score of 60 might infer that she or he at least partially mastered most of the core competencies the exam was designed to test, but still has substantial room to improve. At the same time, calibrating exams so that class means fall in the 60's would still provide plenty of opportunity for the strongest students in a class to distinguish themselves. (Previous research found that the need to distinguish and challenge the brightest students was the primary benefit instructors and students gave in support of LOC grading).<sup>1</sup> Instructors could arrange for approximately half of the items to test student mastery of core concepts covered in class (i.e., half of the items would be criterion-based) and the other half could test students' abilities to apply information in new ways.

In the course of studying LOC grading, we also became aware of other instructional strategies that seem to violate the assumption that exam grades should provide students with information about their progress in the course. For instance, students told us that they often did not have enough information about the instructors' curving practices—or about the means and ranges of exam scores—to know what their grade in the course was. In fact, early in their academic careers, students often assumed they were failing the course when in fact, their grades were above average. We also wondered, given the large number of items students miss in these exams, whether instructors would review these items in class so that the exam could be used to inform instruction rather than just serve to measure and classify students.

To find out more about LOC grading and related exam practices and their effects on students, we designed a survey to address the following questions:

- Assuming that their final grades remain the same, to what extent does the raw score of an exam affect students' motivation, self-efficacy, learning strategies, or perception of the instructor?
- To what extent do students believe that exams should be criterion-based? What other assumptions do students hold about the purpose of the exam?
- How common is it for instructors to provide students with information about exam scores and grading methods?
- How common is it for instructors to treat exams as a learning instrument by reviewing problems that the majority of the class missed?
- To what extent do student perceptions of LOC grading vary by gender, ethnicity, or achievement level (i.e., GPA)?

## Method

### *Participants:*

Participants include 177 engineering undergraduates at a private, research I university. All participants were U.S. citizens in their sophomore year or later. Approximately half of the participants (n=86) were women and half men (n=88). Under-represented minorities constituted just 16% (n=29) of the sample. Over three-fourths of the students (n=139) had GPAs of 3.0 or higher.

Participants were incentivized to participate with a \$10 credit loaded to their campus cash card. The survey had a response rate of slightly over 50%.

### *The survey:*

The survey asked four types of questions: students were first asked a series of four-point Likert-scale questions ranging from “not at all” to “very” about their perceptions of exams with means in the 20’s where the highest score is in the 40s and is curved to an “A” versus exams with means in the 60’s where the highest score is in the 80’s and is curved to an “A.” A complete list of survey questions for exams with means in the 20’s appears in Table 1.

**Table 1: Survey questions for exams with means in the 20’s**

- If you receive a score of 30 that ends up being an A, how proud do you feel about your accomplishment?
- If you receive a score of 30 that ends up being an A, how much do you feel you have learned of what the instructor expected?
- How much does an exam like this suggest the instructor cares about the students?
- How much does an exam like this suggest the instructor is a bad teacher?
- How much does an exam like this feel discouraging?
- How much does an exam like this encourage you to study harder for the next exam?
- How much does an exam like this encourage you to consider cheating on the next exam?
- How much do you enjoy the challenge of this kind of exam?
- If you receive a score of 20 that ends up being a C, how prepared do you feel to attempt the next level of this subject?

Next, students were asked about their perceptions of what should ideally be the purpose of exams: to measure their mastery of the core content of the course; to measure learning; to test their abilities to apply material in new ways; and to distinguish exceptional students from the rest. Students were asked to rate each of these as “should be a major purpose” “should be a minor purpose,” or “should not be a purpose.”

Students were also asked how common various post-exam practices were, including how often instructors reviewed exam questions the majority of the class missed and how often instructors gave students information about exam means and ranges.

Finally, students were asked to state a preference for exams with means in the 20's versus means in the 60's and provide reasons for their preferences in an open-ended response. Students had the option of stating "no preference."

**Analysis:**

The data was analyzed using paired t-tests to assess differences in perceptions of the two types of exams. Differences between gender, race, and GPA were analyzed using t-tests.

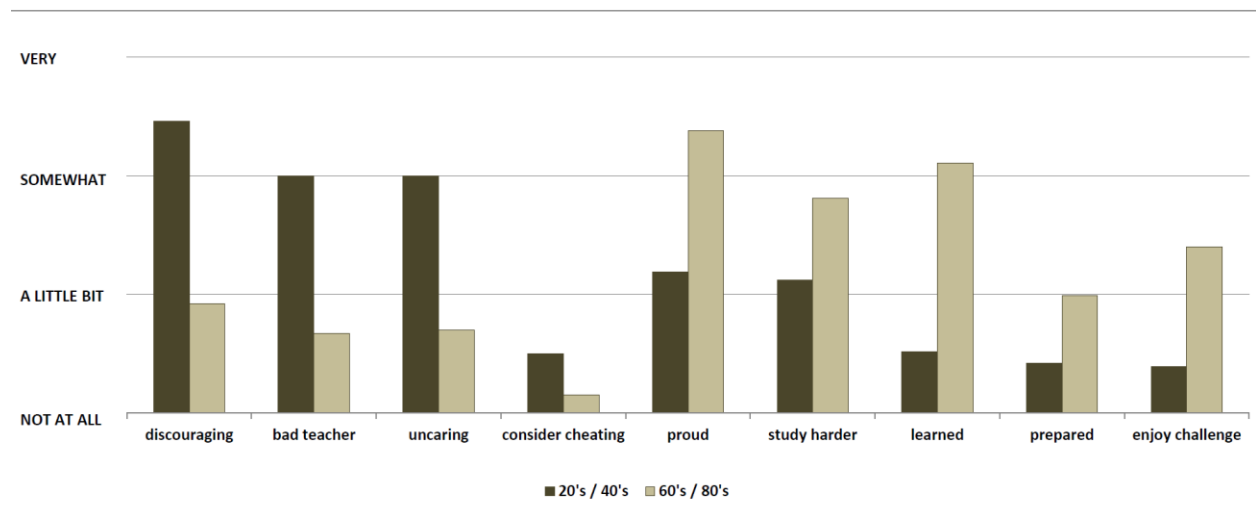
**Results**

***Student Perceptions of exams with means in the 20's versus 60's***

Students were asked to rate their attitudes towards exams with means in the 20's (where the highest score is in the 40s and a 40 is an A) with exams with means in the 60's (where the highest score is in the 80's and an 80 is an A). On all measures, students were more negative about exams with means in the 20's, even when their grades translated to an A.

Figure 1 shows that when exam means are in the 20's, students indicate they are more likely to be discouraged, consider the instructor a bad teacher, perceive the instructor as uncaring towards students, and more likely to consider cheating. If they receive a score of 30 on an exam and that score ends up being an A, they report that they are unable to take pride in their accomplishments, they feel little motivation to study harder for the next exam. Even if their grade is an A, students with low raw scores feel that they have not learned what the instructor expects of them.

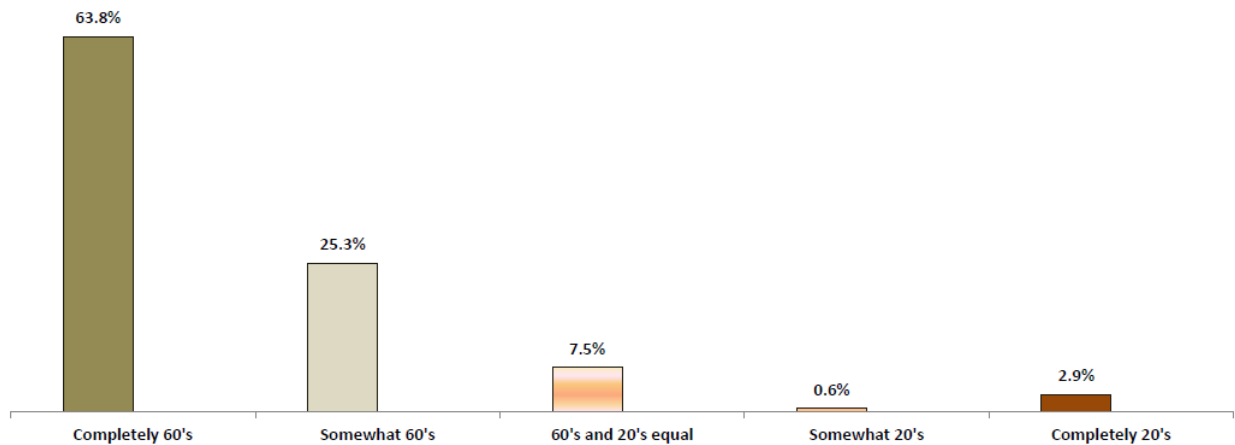
If students receive a score of 20 and that score ends up translating to a C, students feel less prepared to attempt the next level of the subject than they do if their score is a 60 and their grade is a C (though in both cases, students indicate a grade of C leaves them feeling unprepared). Students also generally do not find exams with very low scores to present an enjoyable challenge.



**Figure 1: Students motivation, self-assessment, study habits and impressions of the instructor when exams means are in the 20's (and a 20 translates to a C) versus when**

means are in the 60's (and a 60 translates to a C). All items are significant at  $p < .01$  or stronger.

Figure 2 shows that when students were asked which exam type they preferred, 89% stated a preference for exams with means in the 60's and where a score of 80 would translate to an A. Only 3.5% preferred exams with means in the 20's where a score of 40 would translate to an A. The remaining 7.5% of students expressed no preference.



**Figure 2: Students preferences for exams with means in the 20's versus means in the 60's. Results are significant at  $p < .001$ .**

Students' preferences for the two types of exams did not differ for students with high and low GPAs. However, we did find that students with high GPAs (3.0 and above) were more likely than those with low GPAs to report enjoying the challenge of both types of tests. Students with low GPAs were more likely than those with high GPAs to consider cheating when exam means were in the 20's.

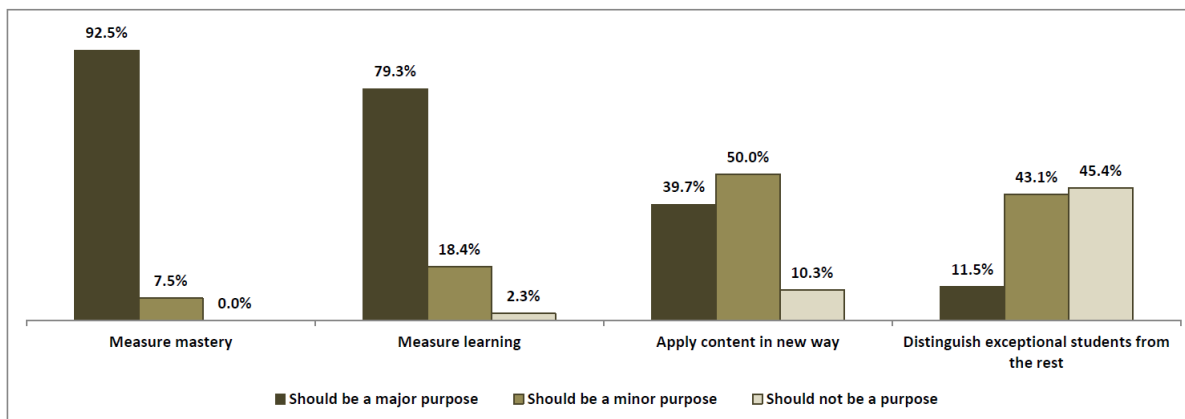
### ***Student beliefs about the purpose of exams***

Figure 3 demonstrates that students' beliefs about the purpose of exams are much more in line with the goals of criterion-referenced than norm-referenced grading. Students overwhelmingly agreed that exams should measure their mastery of core concepts in the course with 92.5% agreeing that it should be a primary purpose of the exam and the remaining students stating it should be a secondary purpose. No students indicated that assessing their mastery of core course concepts should not be a purpose.

Students also felt, although not as strongly, that exams should measure what they learned in the course. Over 79% of students agreed that measuring learning should be a primary purpose of exams and 18% saw this as a secondary purpose. Students who disagreed that measuring learning should be a purpose for exams generally stated that students could enter the class already knowing the material and it would be irrelevant to measure what they had learned.

Most students also felt that exams should assess their abilities to apply course content in new ways, although 50% claimed this should be a secondary, and not a primary purpose of exams.

Students generally felt that distinguishing exceptional students should *not* be a primary purpose of exams, although many felt it should be a secondary purpose. Over 45% stated that distinguishing exceptional students should not be a purpose at all.



**Figure 3: Percent of students who believe that various goals should be a major purpose, minor purpose, or not a purpose of exams**

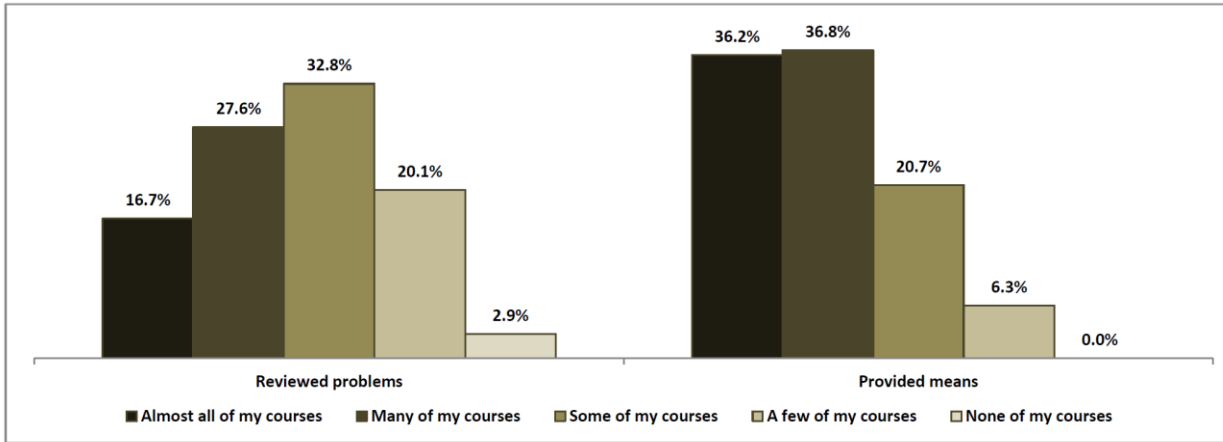
Students with high GPAs were more likely to state that distinguishing exceptional students should be a *minor* purpose of an exam, while those with low GPAs were more likely to state that this should *not* be a purpose.

### ***Post-exam experiences***

We were curious about how often instructors used exams as learning opportunities by reviewing problems the majority of the class struggled with and how often instructors gave students information about class means and ranges. Figure 4 indicates that it is not routine for instructors to review problems the majority of class missed on exams. While nearly 45% of students indicated that such reviews happen in “almost all” or “many” of their courses, the majority of students indicated that this is an infrequent experience, with 3% stating that such review never happens.

Instructors do provide information about class means and ranges, although this practice is not universal. Figure 4 shows that 72% of students indicated that in “almost all” or “many” of their courses, information about exams scores is provided that would help them gauge their performance relative to their peers.

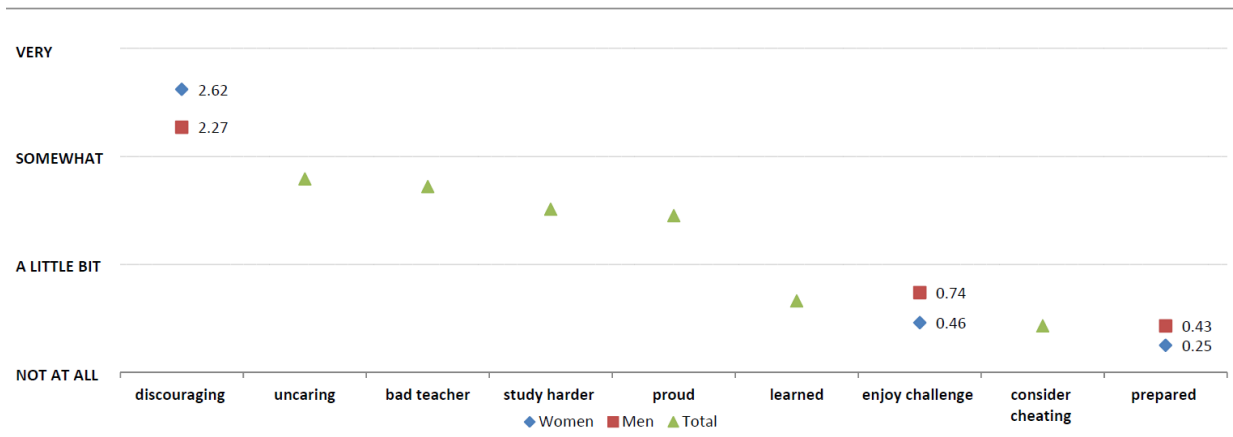




**Figure 4: Student perceptions of how frequently instructors review problems after exams and provide information about means and ranges.**

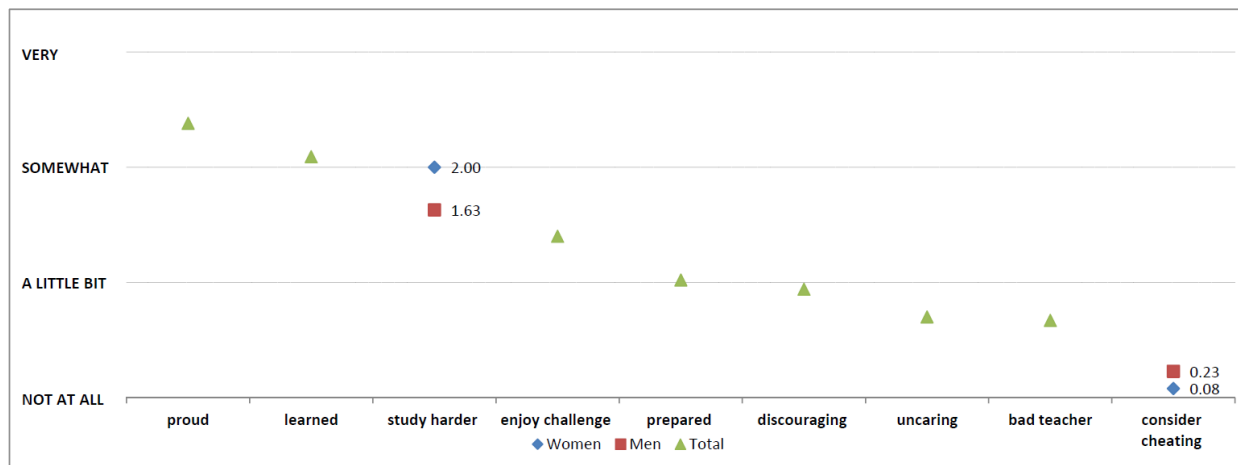
*Gender and ethnicity differences*

Figure 5 shows that female and male students differed in their perceptions of exams with means in the 20’s and the highest score was in the 40s. Women were more likely than men to feel discouraged by such exams and less likely to enjoy the challenge or report feeling prepared to attempt the next level of content.



**Figure 5: Gender differences in how students perceived exams with means in the 20’s. Scale = Very (3); Somewhat (2); A little bit (1); Not at all (0). Items with plot points for both men and women are significant at  $p < .05$ .**

Figure 6 shows there were also some gender differences in how female and male students perceived exams with means in the 60’s. Women were more likely to report feeling motivated to study harder, while men were more likely to consider cheating (although both genders were much less likely to consider cheating when means are in the 60’s than in the 20’s).



**Figure 6: Gender differences in how students perceived exams with means in the 60's. Scale = Very (3); Somewhat (2); A little bit (1); Not at all (0). Items with plot points for both men and women are significant at  $p < .05$ .**

Our analyses found one difference between minority students and non-minorities: minority students were less likely than majority students to feel that they had learned what the instructor expected when exam means were in the 20's.

### *Student Comments*

Nearly 90% of the students (n=156) provided comments to explain their choice for preferring one type of exam over the other. This high number of comments speaks to how strongly students felt about this topic. Over half of the students made comments that suggest they hold beliefs consonant with criterion-referenced grading. These students explicitly indicated that they expect exams to measure their mastery of content.

I want the exam to reflect the material I learned in the classroom. In super hard exams that isn't what I'm being tested on.

If the goal of exams is to test learning and mastery of course content, it seems unfair that some professors make the exams difficult enough that the median score is as low as 20%. Most of the exam, then, is above the ability the students and doesn't actually test knowledge of what has been taught.

[A mean in the 60's] provides some sort of confirmation telling us we learned something. With just a 20 its [sic] no different from day 0.

Why have "average mastery" be where a student can answer only 1/5th of the problems? In what real world setting does this apply?  
An exam with a median of 20 suggests to me that the professor had expectations that were not met, and that students did not actually learn much from the class.

Several students (8%; n=14) also commented that placing test means around 60 provided a good balance between testing mastery and challenging students:

[Median scores in the 60's] leaves room for hardworking students to excel, while at the same time it shows that the median of students were able to learn the majority of the material.

With a median score of 60, it means the test wasn't too hard that some people couldn't do very well in it. The test was an accurate judgment of how much we learned.

A few students (4%; n=7) commented that they felt that means should be higher than the 60's.

I feel that the median could also be a bit higher. I feel like you actually can apply what is learned and the questions are not completely out there.

However, most students seemed to feel that means in the 60's are reasonable.

## **Discussion**

Our findings strongly suggest that students hold beliefs about assessment that are consonant with criterion-referenced grading. Students believe that exams should assess them primarily on their mastery of core course content and secondarily on their ability to apply course concepts in new ways. Most students disagreed that comparing them to their peers—the basis of norm-referenced grading—should be a primary purpose of class assessments.

LOC grading violates the assumptions of criterion-referenced grading by having very few items on the test that the majority of students are able to answer. Our results suggest that students feel that these tests are unfair because there is little relationship between what was taught in class (or assigned as homework) and what they were assessed on. Students find LOC grading highly discouraging. They are unable to take pride in a grade of A when they missed the majority of items on the exam, and they feel that they have failed to meet the instructors' expectations. As previous research has found, LOC grading decreases student motivation to study<sup>1</sup> and makes students more likely to consider cheating.

It is not unreasonable to view LOC grading as one of the contributors to attrition in engineering. Students in our study strongly perceived LOC grading as evidence of bad teaching and uncaring instructors, which are among the most common factors cited in students' decisions to leave STEM fields.<sup>13,16</sup> We also found evidence that LOC grading disproportionately discourages female and minority students.

However, our findings also suggest that the negative consequences of LOC grading can be ameliorated if instructors shift exam means upward by including more criterion-referenced items that test student mastery of key course concepts. Students did not report widespread discouragement when means were in the 60's and the top of the range was in the 80's. Many felt that this kind of distribution was a good compromise between assessing them on what they had mastered while challenging them in new ways. It is important to note that calibrating exams so

that means are in the 60's will not contribute to grade inflation. There would still be a large range of scores, and most exams would not max out the top of the grading scale. Instructors could achieve means in the 60's by ensuring that half of their items were tied to specific course concepts while the other half could present students with new challenges.

Of course, this survey was only of students at one institution. But it seems likely that students elsewhere would feel similarly. What remains the biggest question about LOC grading is understanding why faculty engage in this practice and what it would take to persuade them to change their assessments.

### **Acknowledgements**

This material is based upon work supported by the National Science Foundation under Grant No. 1262274. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.”

We also gratefully acknowledge the support and assistance of the Carnegie Institute of Technology and the Office of Institutional Research and Analysis at Carnegie Mellon University.

### **References**

1. Wolfe, J. & Powell, E. in *American Association of Engineering Educators Conference Proceedings* (2014).
2. Gronlund, N. E. *Assessment of student achievement*. (Pearson/Allyn and Bacon, 2006).
3. Astin, A. W. & Antonio, A. L. *Assessment for Excellence: The Philosophy and Practice of Assessment and Evaluation in Higher Education*. 380 (Rowman & Littlefield Publishers, 2012). at <<http://books.google.com/books?hl=en&lr=&id=jEsRVTbukyMC&pgis=1>>
4. Covington, M. V. *Making the grade: A self-worth perspective on motivation and school reform*. (Cambridge University Press, 1992).
5. Sadler, D. R. Interpretations of criteria- based assessment and grading in higher education. *Assess. Eval. High. Educ.* **30**, 175–194 (2005).
6. Mintzes, J. J. & Leonard, W. H. *Handbook of College Science Teaching*. (National Science Teachers Association, 2006).
7. Ames, C. Classrooms: Goals, structures, and student motivation. *J. Educ. Psychol.* **84**, 261–272 (1992).
8. Weimer, M. E. *Improving your classroom teaching*. (SAGE Publications, 1993).
9. Covington, M. V. & Omelich, C. L. Task-oriented versus competitive learning structures: Motivational and performance consequences. *J. Educ. Psychol.* **76**, 1038–1050 (1984).
10. Bergin, D. A. Effects of a Mastery Versus Competitive Motivation Situation on Learning. *J. Exp. Educ.* **63**, 303–314 (1995).

11. Anderman, E. M. & Midgley, C. Changes in self-reported academic cheating across the transition from middle school to high school. *Contemp. Educ. Psychol.* **29**, 499–517 (2004).
12. Wilkinson, T. J., Wells, J. E. & Bushnell, J. A. What is the educational impact of standards-based assessment in a medical degree? *Med. Educ.* **41**, 565–572 (2007).
13. Seymour, E. & Hewitt, N. M. *Talking About Leaving: Why Undergraduates Leave the Sciences*. (Westview Press , 1997).
14. ABET - Criteria for Accrediting Engineering Technology Programs, 2014 - 2015 . at <<http://www.abet.org/etac-criteria-2014-2015/>>
15. Hutchison-Green, M. A., Follman, D. K. & Bodner, G. M. Providing a Voice: Qualitative Investigation of the Impact of a First-Year Engineering Experience on Students' Efficacy Beliefs. *J. Eng. Educ.* **97**, 177–190 (2008).
16. Marra, R. M., Rodgers, K. A., Shen, D. & Bogue, B. Leaving Engineering: A Multi-Year Single Institution Study. *J. Eng. Educ.* **101**, 6–27 (2012).