Better Remedies For Bad Exams: Correcting for Difficult Questions in a Fair and Systematic Way

Devin Camenares <camenaresd@alma.edu>

Alma College

Received: 23 March 2022; Accepted: 5 September 2022

Balancing assessment of learning outcomes with the expectations of students is a perennial challenge in education. Difficult exams, in which many students perform poorly, exacerbate this problem and can inspire a wide variety of interventions, such as a grading curve. However, addressing poor performance can sometimes distort or inflate grades and demotivate students. Here, I review this problem and propose solutions which are fair and based upon item analysis of an assessment. To correct for exams with questions that were more difficult than intended, item score and discrimination is used to determine a bonus value in a way that is limited and continuous. Educators that utilize item analysis to improve exams for future iterations can easily use this approach to retroactively lessen the burden of some questions. This solution expands the repertoire of tools for achieving a transparent, fair, and equitable means of assessment.

INTRODUCTION

Across many disciplines and institutions of higher learning, educators are occasionally faced with the undesirable scenario wherein many students perform poorly on an exam. There are a variety of reasons for such an outcome: for example, the test might not have been constructed properly, the students were not instructed on relevant material, or the students did not prepare themselves adequately. There are perhaps an equal variety of responses and remedies from educators, aligning with a diverse set of teaching philosophies (Brookhart et al., 2016). Some allow the results to stand, perhaps making adjustments to future iterations of the test or other assignments in that semester. Yet others apply some form of an active remedy: either scaling grades upwards, removing difficult questions from calculation of the total, or matching the class performance to a predetermined distribution. The latter is typically referred to as 'curving' the grade, and is usually done based on the exam average (Kulick, 2008). Traditionally, the grades are fit to a normal bell curve distribution, although other approaches with logistic curves and bimodal distributions have been tried (See Fig S2 for an illustration of a normal distribution).

Determining if a test was improperly constructed is almost always a useful exercise, for this can help inform the design of a better assessment in the future (Bibler Zaidi et al., 2018). This determination is usually carried out using item analysis, looking at each question on the test individually against different metrics. This analysis, derived from Item Response Theory, usually involves determining the difficulty and the discriminatory power of a given question (Adroher, Prodinger, Fellinghauer, & Tennant, 2018; Benedetto, 2020; Himelfarb, 2019; Towns, 2014). Question difficulty is often assessed as the average score achieved across the entire class. The discriminatory power or discrimination index is a measure of how well the question separates high performing students from low performing students. If a question was answered correctly more often by students that performed well on the test overall, it will have a positive discrimination index. Questions in which low performing students did equally well or even better than high performing students may have a low or negative discrimination index. This situation may arise for very difficult questions in which all students resorted to guessing, or ones in which there was a 'trap' answer or distractor.

Since the overarching goal of higher education institutions is for students to learn, tests and exams should act as assessments for their mastery of the material. Moreover, students are the immediate beneficiaries of this knowledge. An improper assessment, with exceedingly difficult questions, will not give students appropriate feedback on the progress of their own learning. It may also result in a lower overall grade than is warranted based on their command of the material. Attempts to correct for this unwanted result, however, may create additional problems. Curves or scaling that operates on the test average may give a student a skewed perspective on their own progress, and may affect their motivation for future coursework. Curving an exam may also result in grade inflation, grade compression, or the perception of an unfair learning environment, as suggested both by empirical observations and simulation (Bailey & Steed, 2012; Dubey & Geanakoplos, 2010; Grant, 2016; Kulick, 2008). A more surgical approach, such as the removal of questions that are deemed to be poor by item analysis, may be better but can suffer from arbitrary cutoffs.

As the preceding paragraphs suggest, the issues surrounding grading and problematic assessments have not escaped the attention of experts in the scholarship of teaching and learning (SoTL). On the contrary, there are a variety of perspectives on this topic and considerable debate regarding solutions which have been summarized elsewhere (Anderson, 2018). This is not surprising, given the importance these practices to teaching and student outcomes. Indeed, students and faculty often form different opinions as to whether grade inflation exists, if a curve is appropriate, or if a test is administered fairly (Baglione & Smith, 2022; Chowdhury, 2018). The reliability of individual grades, the impact they have on student self-esteem, and their overall meaning have all been systematically explored in the SoTL field without definitive conclusions or the emergence of a single best practice (Goldhaber & Ozek, 2019). In addition, there is considerable interest in the larger implications that grades and assessments have on institutions of higher education and equity (Jephcote, Medland, & Lygo-Baker 2021). It is against this backdrop that an educator must decide how to proceed when the class underperforms on a given assignment.

Rather than curving or scaling total exam grades, this paper argues for the use of functions to be applied to each question, to

provide students a proportional bonus for questions that are too difficult and/or are poor discriminators. These functions can be easily applied after carrying out a typical item analysis, and they do not feature any arbitrary cutoff values but are instead continuous. By applying a significant bonus only for very difficult questions, it will address problems of an improperly constructed test, while allowing for a wide grade distribution if students had not prepared well for the exam. Herein are described two such functions – a simple model that only takes into account item difficulty, and a logistic model that also incorporates the discrimination index.

USING A POWER FUNCTION TO CORRECT FOR OVERLY DIFFICULT QUESTIONS

Item analysis can help reveal the difficulty of a given question. While most educators purposefully vary the difficulty of the questions on their exams, sometimes the apparent difficulty of a question is greater than was intended. Anecdotal evidence, from interactions with colleagues online and in person, show a preference to deal with questions beyond a particular difficulty threshold by removing them from the calculation of total score, effectively awarding their full credit to all students. However, this does not deal effectively with borderline cases, in which a question was just slightly less difficult than the threshold. Since this arbitrary threshold cannot perfectly distinguish between a challenging and unfair question, it indicates the need for a continuous function that can be applied to each question to determine an aggregate bonus to apply to every students test (Fig. 1A). Such an approach can be based upon a power function, raising a measure of student performance on a question to a set exponent (See Fig. S1 for a general illustration of a power function). It was designed to produce an increasingly large bonus for questions more difficult than the threshold, and an exceedingly small bonus for questions that were easier. The threshold (S_t) should be set to a level that is the lowest score you would expect for a question on your exam. It is common and reasonable to include several difficult questions on an exam, for which you might expect only a third of the students to get it correct – hence, in the examples shown, the threshold was set near this level. The degree to which this function is switch like depends on the value or factor in the exponent, ("F") and can be easily modulated for a more or less gradual application of the bonus. The bonus for each question can be calculated in this way, summated across the entire test, and then applied to the total grade of everyone in the class.

USING LOGISTIC FUNCTIONS THAT INCORPORATE BOTH DIFFICULTY AND DISCRIMINATION

The classic three parameter model of Item Response Theory utilizes a logistic function which incorporates item difficulty, discrimination index, and some measure of intrinsic student ability (Adroher et al., 2018; Benedetto, 2020; Himelfarb, 2019; Lim, Lee, Ahn, Lee, & Im, 2016). This model served as inspiration for modifications to the function described in the preceding section. The new function will apply bonus points based on two of these values as a pair of logistic functions (Fig. 2A). A logistic function features a switch-like response to a changing variable (See Fig. S1 for a general illustration of a logistic function). There are two key features to this approach. First, questions in which the students did well can receive only an infinitesimal bonus, since the maximum value will be governed by the nominator of 1 – S, where S is the score ranging from 0 to 1. If the entire class got the question correct, the value of S is 1 and the nominator, and hence the bonus, is 0. Most scores well above the desired threshold will likewise be very small. Questions for which the class scored at the threshold will receive half of the maximum bonus for that question. Like the simple model, if the question was more difficult than was intended, and the score was below this threshold, this bonus will be even greater.

The other key feature is that the bonus is modulated by the discrimination index, using another logistic function. This index can help distinguish between difficult questions that are merely challenging from those that are unfair. Using a threshold score of 0.35 and a threshold discrimination index of 0.15, which is a moderate to poor performing value, we can see the effect of different values for a range of questions (Fig 2.) For a question in which only 20% of the class gets the correct answer, if this 20% was mostly high performing students, then this represents a challenging yet fair question. Accordingly, the bonus applied only raises the class average for this question to about 30%. However, if the question was unfair or otherwise improper, featuring a negative discrimination index, the class average for this question may be raised to about 50%. Like the simple model, this can also be made more sensitive by changing a constant value in the exponent (Fig 2B).

SIMULATION OF BOTH MODELS

Previous studies have analyzed the impact of a traditional curve on student grades through simulation (Kulick, 2008). These studies approximated student performance by assigning each student a value from a normal distribution that represented their level of preparation for the exam – this then determined if they succeeded on questions from a uniform distribution of difficulty. However, subsequent analysis showed that this simulation doesn't align with some real world examples, and relies on the unrealistic assumption of uniform difficulty across an exam (Bailey & Steed, 2012). Nevertheless, the simulation is a useful tool and starting point for investigating the impact of the models suggested above.

Here, this simulation was repeated with two important alterations. First, instead of curving the grades of the class, the bonus specified by either the simple or logistic model was applied. Second, question difficulty was not uniform. Indeed, the bonus models presented here are designed specifically to deal with exams that feature questions that were more difficult than anticipated or intended. Here, question difficulty was generated the same way that student preparation was simulated: by selecting values from a normal distribution with a specified mean and standard deviation. This way, the amount of challenging, unfair, problematic, or trivial questions could be varied. The results of this simulation can be found in Table 1.

One clear, yet unsurprising trend emerges from this simulation: the more unfair questions there are on the exam, the greater the bonus applied to the student's grade. A striking example is when the distribution of difficulty matches that of student preparation exactly, meaning most questions are near the limit of most student's abilities. In this case, the results do not fit well to a normal distribution, yet the low class average may suggest that some curve or remedy is required. However, when the variance in difficulty is low (Table 1, row B), item analysis of this exam would reveal that a majority of exam questions were neither problematic

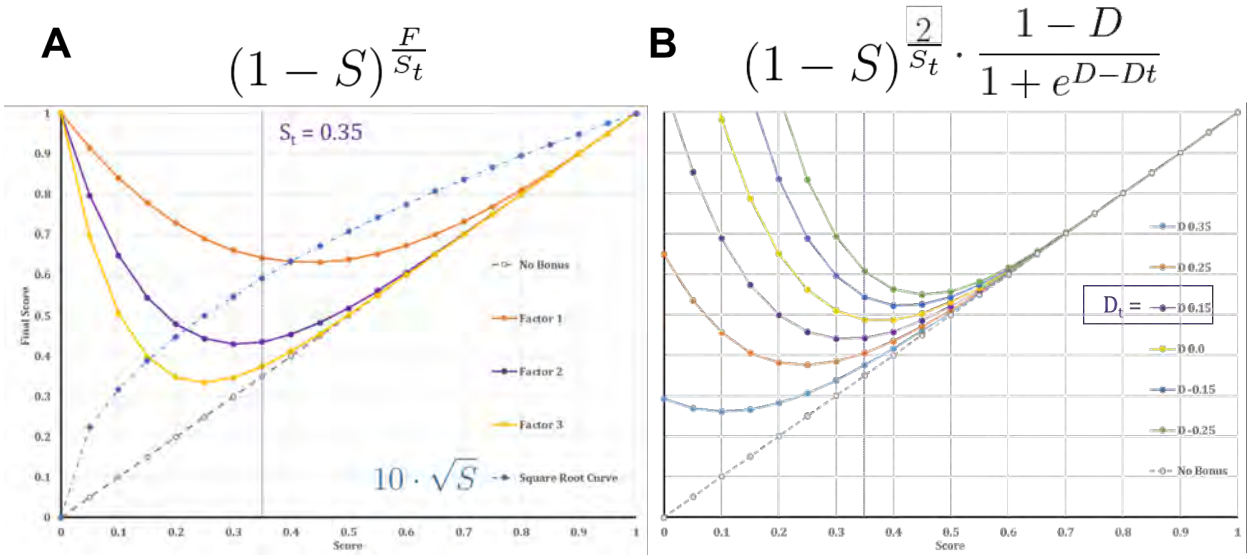

Figure 1. The final score of each question, after bonus points are applied using the simple model (A), or the modified simple model (B), under a range of different conditions. The student performance on a question (S) is shown on the x-axis, and the final score after the bonus is shown on the y-axis. Bonus points are applied for the three colored, solid lines, using the score threshold (St) of 0.35, as indicated by the dashed purple line. The value, nominator, or 'factor' (F) in the exponent is different among the three colored lines – the purple solid line matches the header equations, which calculates the bonus as follows: for each question, the fraction of available points awarded is calculated based upon the score of the class (S), ranging from 0 to 1, and a threshold score (St), representing the intended maximum difficulty of all the questions. By way of contrast, the performance a commonly employed method, the 'square root curve', is shown as an inset in the graph. This method takes the square root of the student's raw score, multiplies it by 10, and uses this as the new score. In (B), the simple model using a factor of 2 is modified by multiplying it with a logistic function that incorporates the discrimination (D) for a question item. For a given threshold value (Dt, here 0.15), the bonus for questions with a discrimination ranging from 0.35 (good) to -0.25 (poor) are shown as different colored lines. The simple model is based upon a power function, while the modified one is based upon a logistic function. The general principles and behaviors of these types of functions are illustrated in Fig S1.

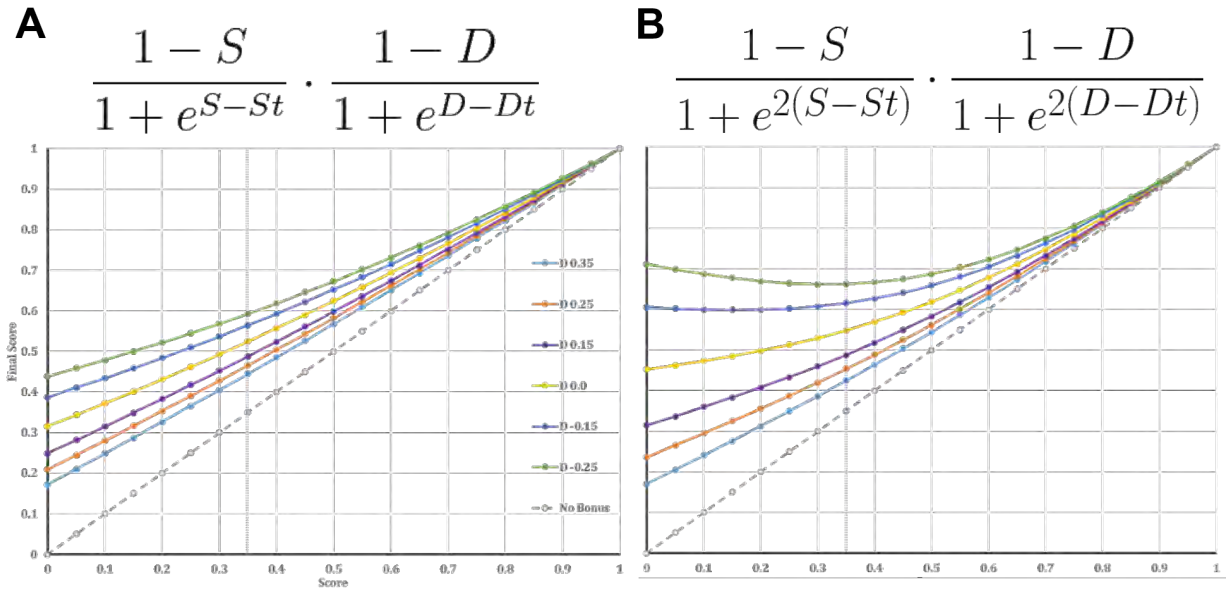

Figure 2. Example of the score obtained for a question after bonus points are applied using the logistic function (y-axis) for questions with a range of initial scores (x-axis) and discrimination index values. This was calculated using the equations shown above each graph. Using these equations, for each question, fractions of the available points are awarded based on the product of two logistic functions, shown above. The left function features the item difficulty or score (S) as a fraction between 0 and 1. This is compared to a target or threshold score (St). The right function similarly features the discrimination index (D) or the threshold for the desired discrimination index (Dt). A crude discrimination index value was obtained by first splitting students into two smaller groups – those that were in the top 50% of the class with respect to performance on the test, and those in the bottom 50%. The difference between the number in the former group getting the question correct from the number in the latter group that answered the question correctly is multiplied by two and then divided by the size of the entire class. Negative values result when more poor-performing students answered the question correctly than did the higher achieving students. Bonus points were calculated for a series of 20 different questions, each with 6 possible discrimination index values. The difficulty threshold (St) was 0.35, as indicated by the purple dotted line, and the discrimination threshold (Dt) was 0.15, which matches the purple series line. The discrimination index for a question is represented by different colored lines and the inset legend. As a comparison, the points that are obtained with no bonus applied is shown as a gray dotted line.

Table 1. Simulation of student performance and the utility of the simple and logistic bonus models. Following previous studies(Kulick, 2008), 400 students were simulated answering 60 dichotomously scored questions across 3 exams. Student performance was modelled primarily by two factors: preparation, and item difficulty. In all simulations, preparation for each student was randomly selected from a normal distribution with a mean (u) of 0.75 and a standard deviation (d) of 0.083. Difficulty varied across simulations as indicated in the first column: (A) item difficulty was randomly generated as a number between 0 and 1. (B-F) item difficulty was randomly selected from a normal distribution with the indicated means (u) and deviations (d): B = 0.75u, 0.083d; C = 0.75u, 0.166d; D = 0.65u, 0.083d; E = 0.85u, 0.083d; F = 0.5u, 0.25d (See Fig S2 for an illustration of the distribution of questions). All other columns show the average of 5 separate simulations with those conditions, and the standard deviation shown in parentheses. Class averages are reported as raw score (R), plus either the bonus from the simple model (R+S) or the logistic model (R+L). Questions were categorized as either trivial (T), moderate (M), challenging (C), unfair (U), or problematic (P) when <100%, <80%, <60%, <40%, or <20% of students answered them correctly, respectively. The actual bonus amounts for the simple model (B-S) or logistic model (B-L) are also reported.

	Class Averages			Question Categories					Bonus Amounts	
Simulation	**R**	**R+S**	**R+L**	**P**	**U**	**C**	**M**	**T**	**B-S**	**B-L**
(A) Random Difficulty	74.96 (3.6)	83.62 (1.1)	78 (2.8)	11 (3.5)	2 (1.6)	3.6 (1.5)	2.6 (1.1)	40.8 (1.3)	8.66 (3.1)	3.04 (0.9)
(B) Difficult, Uniform	49.48 (4.3)	57.62 (3.3)	52.94 (3.9)	12.8 (5.1)	11.8 (4.3)	12.6 (3.6)	10.6 (2.1)	11.6 (4.5)	8.14 (2.3)	3.46 (0.9)
(C) Difficult, Varied	51.14 (3.4)	66.92 (2.1)	56.8 (2.8)	19 (2.2)	7.6 (3)	6.4 (1.5)	6.8 (2.4)	20.2 (2.9)	15.78 (2.6)	5.66 (0.7)
(D) Moderate, Uniform	69.88 (3.5)	78.16 (2.3)	72.82 (3.1)	10.2 (2)	4 (1.2)	5.4 (1.9)	6.2 (3.1)	33.8 (4.1)	8.28 (1.4)	2.94 (0.5)
(E) Very Difficult, Uniform	18.68 18.68)	48.56 (48.56)	29.46 (29.46)	38.8 (38.8)	13 (13)	3.2 (3.2)	3.4 (3.4)	1 (1)	29.88 (29.88)	10.78 (10.78)
(F) Easy, Highly Varied	83.48 (1.3)	88.22 (1.5)	85.16 (1.4)	5.2 (0.4)	1.8 (1.1)	4 (2.8)	2.8 (1.1)	46 (1.9)	4.74 (0.8)	1.68 (0.2)

nor unfair. Accordingly, the simple and logistic models only afford a small bonus to such a performance.

This can be contrasted with an exam that has the same average level of difficulty of questions, and nearly the same class average, but a wider distribution of question difficulty (Table 1, row C). In this case, student performance suffered due to a greater amount of problematic or unfair questions. This is not the result of some students focusing on the wrong material or being unlucky – if it was, more of their peers would have done better. These types of questions, especially those in which less than 20% of the class answer correctly, are more likely due to problems with question construction or delivery of the instructional content. In these cases, a larger bonus is warranted, and granted by both the simple and logistic models. Thus, it is possible to have two exams with nearly identical class averages, but different levels of student preparation and different remedies applied.

CONCLUSIONS AND FUTURE DIRECTIONS

While the goal of enrolling in a college class is ostensibly to learn the material, many students are also highly focused on making sure they perform well on the assessments, mostly in service of earning a particular grade for the course and their transcript. Thus, how these assessments are constructed, graded, and adjusted can have serious impacts on a student's perspective of the class and their ability to learn (Marini, Shaw, Young, & Ewing, 2018). Exams and tests that are exceedingly difficult can be discouraging, especially if only a select set of students performed well (Hernández-Julián, Peters, Hernández-Julián, & Peters, 2022). Low class averages correlate with poor student evaluations of faculty, thus creating pressure on faculty to remedy these situations. Some methods of curving or scaling grades are likewise considered unfair by some students, such as those that are in a homogenous class of high achievers (Kulick, 2008; Tan Yuen Ling, Yuen Pui Lam, Loo, Prinsloo, & Gan, 2020). It can create a hyper-competitive environment that adds stress, inhibiting learning and incentivizing cheating, and can exacerbate inequality of outcomes between different groups of students (Johnson et al., 2006; Kashyap, 2019; Kaustubh, 2020).

These problems demand a solution; methods to provide students with a systematic, fair, and informative way to assess their own learning, and the ability to adjust these metrics in likewise fashion. It should include approaches that help to ameliorate low class averages when necessary, while avoiding the specter of grade inflation and grade compression. This is important not only for student perception of their course, but for helping to bolster the eroding public perceptions of higher education (Babcock, 2010; Caruth, Author, & Caruth, 2013). The models for bonus points presented here are designed as tools for instructors in the event that one or more of the questions on their exam were problematic, while avoiding pitfalls such as potentially unfair and arbitrary cutoffs. By joining the pantheon of other tools and remedies for a problematic exam, they expand the practice of teaching and the possibilities an instructor has at their disposal. They also offer fertile ground for SoTL inquiry as to their effectiveness in enhancing the student experience, compared with more conventional curves.

These functions were chosen because they should be relatively easy to implement for faculty that already routinely carry out item analysis (An example of its application is provided in tables S1 and S2). More sophisticated approaches, such as using the three-parameter model from Item Response Theory, may be even better but more difficult to calculate. For example, equipped with the item difficulty and discrimination index across all questions, one could estimate the 'ability' or 'preparation' parameter for each student individually. Then for problematic items, this 'ability' can be used in the model together with a desired difficulty and discrimination index threshold to determine if the student would have gotten the item correct if it was properly constructed. Another possible, and easier improvement would be to combine the simple model with the logistic equation for the discrimination index value. This hybrid function provides the desired switch like response of the simple model, but one that is modulated by item discrimination index (Fig. 1B). Finally, these methods can be used in a variety of ways, such as determining the number of points available on an extra credit assignment following a difficult exam.

In addition to being easy for faculty to calculate, the presented methods can be easily communicated to students.

Doing so may be a step towards reducing student anxiety regarding difficult exam questions, competitive grading curves, and their final grade. Giving students this additional peace of mind may help them focus on what really matters in the course: mastering the content (Chassay, Kenney, & Brase 2019; Connell, Donovan, & Chambers, 2016; Gilbert, 2021).

REFERENCES

Adroher, N. D., Prodinger, B., Fellinghauer, C. S., & Tennant, A. (2018). All metrics are equal, but some metrics are more equal than others: A systematic search and review on the use of the term 'metric.' *PLOS ONE*, *13*(3), e0193861. https://doi.org/10.1371/journal.pone.0193861

Anderson, L. (2018). A Critique of Grading: Policies, Practices, and Technical Matters. *Education Policy Analysis Archives*, 26(4). https://eric.ed.gov/?id=EJ1176557

Babcock, P. (2010). Real costs of nominal grade inflation? New evidence from student course evaluations. *Economic Inquiry*, *48*(4), 983–996. https://doi.org/10.1111/J.1465-7295.2009.00245.X

Baglione, S.L. and Smith, Z. (2022), "Grade inflation: undergraduate students' perspective", Quality Assurance in Education, Vol. 30 No. 2, pp. 251-267. https://doi.org/10.1108/QAE-08-2021-0134

Bailey, G. L., & Steed, R. C. (2012). The Impact of Grading on a Curve: Assessing the Results of Kulick and Wright's Simulation Analysis. *International Journal for the Scholarship of Teaching and Learning*, *6*(1). https://doi.org/10.20429/ijsotl.2012.060111

Benedetto, L (2020) Item Response Theory for assessing students and questions (pt. 1) *Medium*. (n.d.). Retrieved January 12, 2022, from https://medium.com/@bndl22/advantages-in-using-item-response-theory-for-assessing-students-and-more-4a9665258863

Bibler Zaidi, N. L., Grob, K. L., Monrad, S. U., Holman, E. S., Gruppen, L. D., & Santen, S.A. (2018). Item Quality Improvement: What Determines a Good Question? Guidelines for Interpreting Item Analysis Reports. *Medical Science Educator*, *28*(1), 13–17. https://doi.org/10.1007/s40670-017-0506-1

Brookhart, S. M., Guskey, T. R., Bowers, A. J., McMillan, J. H., Smith, J. K., Smith, L. F., … Welsh, M. E. (2016). A Century of Grading Research: Meaning and Value in the Most Common Educational Measure, *86*(4), 803–848. https://doi.org/10.7916/D8NV9JQ0

Caruth, D. L., Author, C., & Caruth, G. D. (2013). Grade Inflation: An Issue for Higher Education?. *Turkish Online Journal of Distance Education*, *14*(1), 102–110.

Chassay, L., Kenney, K. L., & Brase, G. L. (2019). Moving to the head of the class: Exam study decisions when courses grade on a curve. https://newprairiepress.org/ksuugradresearch

Chowdhury, F. (2018) Grade Inflation: Causes, Consequences and Cure. *Journal of Education and Learning*, *7*(6) pg86-92 https://eric.ed.gov/?id=EJ1191199

Connell, G. L., Donovan, D. A., & Chambers, T. G. (2016). Increasing the Use of Student-Centered Pedagogies from Moderate to High Improves Student Learning and Attitudes about Biology. *CBE—Life Sciences Education*, *15*(1), ar3. https://doi.org/10.1187/cbe.15-03-0062

Dubey, P., & Geanakoplos, J. (2010). Grading exams: 100,99,98,… or A,B,C? *Games and Economic Behavior*, *69*(1), 72–94. https://doi.org/10.1016/J.GEB.2010.02.001

Gilbert, J. (2021). Mentoring in a Cooperative Learning Classroom. *International Journal for the Scholarship of Teaching and Learning*, *15*(2). https://doi.org/10.20429/ijsotl.2021.150202

Goldhaber, D., Ozek, U. (2019) How Much Should We Rely on Student Test Achievement as a Measure of Success? *Educational Researcher* 48(7) pg479-483 https://doi.org/10.3102/0013189X19874061

Grant, A (2016) Why We Should Stop Grading Students on a Curve. *The New York Times. (Opinion)*. https://www.nytimes.com/2016/09/11/opinion/sunday/why-we-should-stop-grading-students-on-a-curve.html

Hernández-Julián, R., Peters, C., Hernández-Julián, R., & Peters, C. (2022). Why Try? The Superstar Effect in Academic Performance. *Eastern Economic Journal*, *48*(1), 147–165. https://doi.org/10.1057/S41302-021-00197-5

Himelfarb, I. (2019). A primer on standardized testing: History, measurement, classical test theory, item response theory, and equating. *Journal of Chiropractic Education*, *33*(2), 151–163. https://doi.org/10.7899/JCE-18-22

Jephcote, C., Medland, E., Lygo-Baker, S., (2021) Grade inflation versus grade improvement: Are our students getting more intelligent?, *Assessment & Evaluation in Higher Education*, 46:4, 547-571, DOI: 10.1080/02602938.2020.1795617

Johnson, M. D., Hollenbeck, J. R., Humphrey, S. E., Ilgen, D. R., Jundt, D., & Meyer, C. J. (2006). Cutthroat Cooperation: Asymmetrical Adaptation To Changes In Team Reward Structures. *49*(1), 103–119. https://doi.org/10.5465/AMJ.2006.20785533

Kaustubh (2020). Implications Of Cheating In a Relative Grading System. *Cantor's Paradise*. Retrieved January 12, 2022, from https://www.cantorsparadise.com/implications-of-cheating-in-a-relative-grading-system-6435163be88b

Kashyap, R. (2019). For Whom the Bell (Curve) Tolls: A to F, Trade Your Grade Based on the Net Present Value of Friendships with Financial Incentives. *The Journal of Private Equity*, *22*(3), 64–81. https://doi.org/10.3905/JPE.2019.22.3.064

Kulick, G. R. (2008). The Impact of Grading on the Curve: A Simulation Analysis. *International Journal for the Scholarship of Teaching and Learning*, *2*(2). https://doi.org/10.20429/ijsotl.2008.020205

Lim, H. S., Lee, Y. M., Ahn, D. S., Lee, J. Y., & Im, H. (2016). Item Analysis of Clinical Performance Examination Using Item Response Theory and Classical Test Theory. *Korean Journal of Medical Education*, *19*(3), 185–195. http://www.koreamed.org/SearchBasic.php?RID=2306716

Marini, J., Shaw, E., Young, L., & Ewing, M. (2018). Getting to Know Your Criterion: Examining College Course Grades and GPAs over Time. *College Board*.

Tan Yuen Ling, L., Yuen Pui Lam, B., Loo, W. L., Prinsloo, C., & Gan, M. (2020). Students' Conceptions of Bell Curve Grading Fairness in Relation to Goal Orientation and Motivation. *International Journal for the Scholarship of Teaching and Learning*, *14*(1). https://doi.org/10.20429/ijsotl.2020.140107

Towns, M. H. (2014). Guide to developing high-quality, reliable, and valid multiple-choice assessments. *Journal of Chemical Education*. https://doi.org/10.1021/ed500076x

Supplemental Information

Calculating Difficulty. Difficulty for a question, or Score (S), is calculated simply by taking a sum of the classes points obtained for that question and dividing it by the product of class size and question value.

Calculating Discrimination Index. A crude discrimination index value was obtained by first splitting students into two smaller groups – those that were in the top 50% of the class with respect to performance on the test, and those in the bottom 50%. The difference between the number in the former group getting the question correct from the number in the latter group that answered the question correctly is multiplied by two and then divided by the size of the entire class. Negative values result when more poor-performing students answered the question correctly than did the higher achieving students. Other formulas and approaches to getting this index value are possible, and this is sometimes reported without additional calculation. Here excel was used in the following way: first, by finding the class average, and then setting up two columns using an =IF formula. In one column, a 1 was recorded if the student was in the top 50%, and a 0 if they were not. The second column had the inverse pattern. To find the number of students in the top 50% that got the question correct, a =SUMPRODUCT formula was used, multiplying the first column array by the column containing student answers or scores for a given question. A similar approach and formula was repeated for the second column to find the number of students in the bottom 50% that got the question correct.

Simulation The simulation study by Kulick and White, 2008, was used as a basis for the simulation of student performance. For simplicity and greater distribution, the calculations were carried out in Excel. The model depends primarily on simulating two factors: student 'preparedness' and item 'difficulty'. Following the Kulick and White study, I simulated 400 students taking 60 multiple choice questions across 3 exams. The scoring of these questions was dichotomous. Student preparation was randomly selected from a normal distribution centered around a mean of 0.75, with a standard deviation of 0.083, using the =NORM.INV function in Excel (See Figure S2 for an illustration of the output and this distribution). Item difficulty was generated in two different ways. For the 'random' difficulty, a number between 0 and 1 was generated via the =RAND function in Excel. If the question had a difficulty rating above that of the student's preparation, the question was answered incorrectly; otherwise, the student got the answer correct. For the 'normal' difficulty questions, the difficulty factor was generated by randomly selecting from a normal distribution, using the same function used for student preparation. The mean and deviation were varied and were used as specified in Table 1. Questions were categorized as either trivial, moderate, challenging, unfair, or problematic when <100%, <80%, <60%, <40%, or <20% of students answered them correctly, respectively.

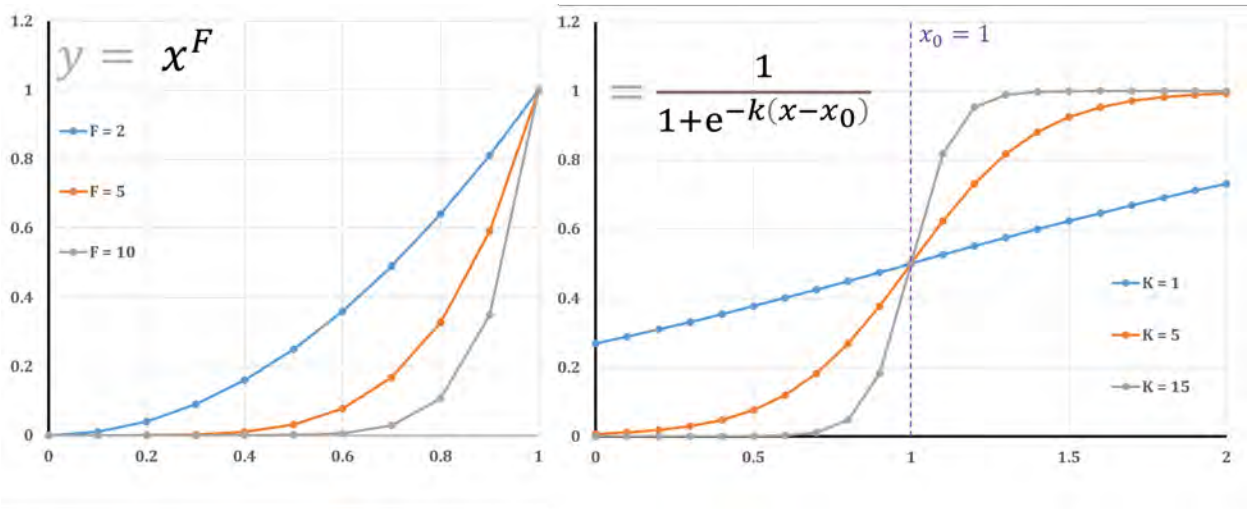

Figure S1. The generalized behavior of power and logistic functions. Power functions (left) take an x-value and raise it to a set exponent. The larger the exponent, the more drastic the difference in output (y-value) for a given input (x-value). Logistic functions (right) are more complex, and provide a switch-like behavior around a given mean (x_0, purple dotted line here set to 1). They accomplish this by raising euler's number (i.e. a natural logarithm) to the difference between the x-value and the mean. The constant k determines how steep or switch-like the curve is, as shown by the variety of values in the graph; the higher k value gives a steeper curve. These functions form the basis of the functions presented in the paper, with relevant values replacing the constants.

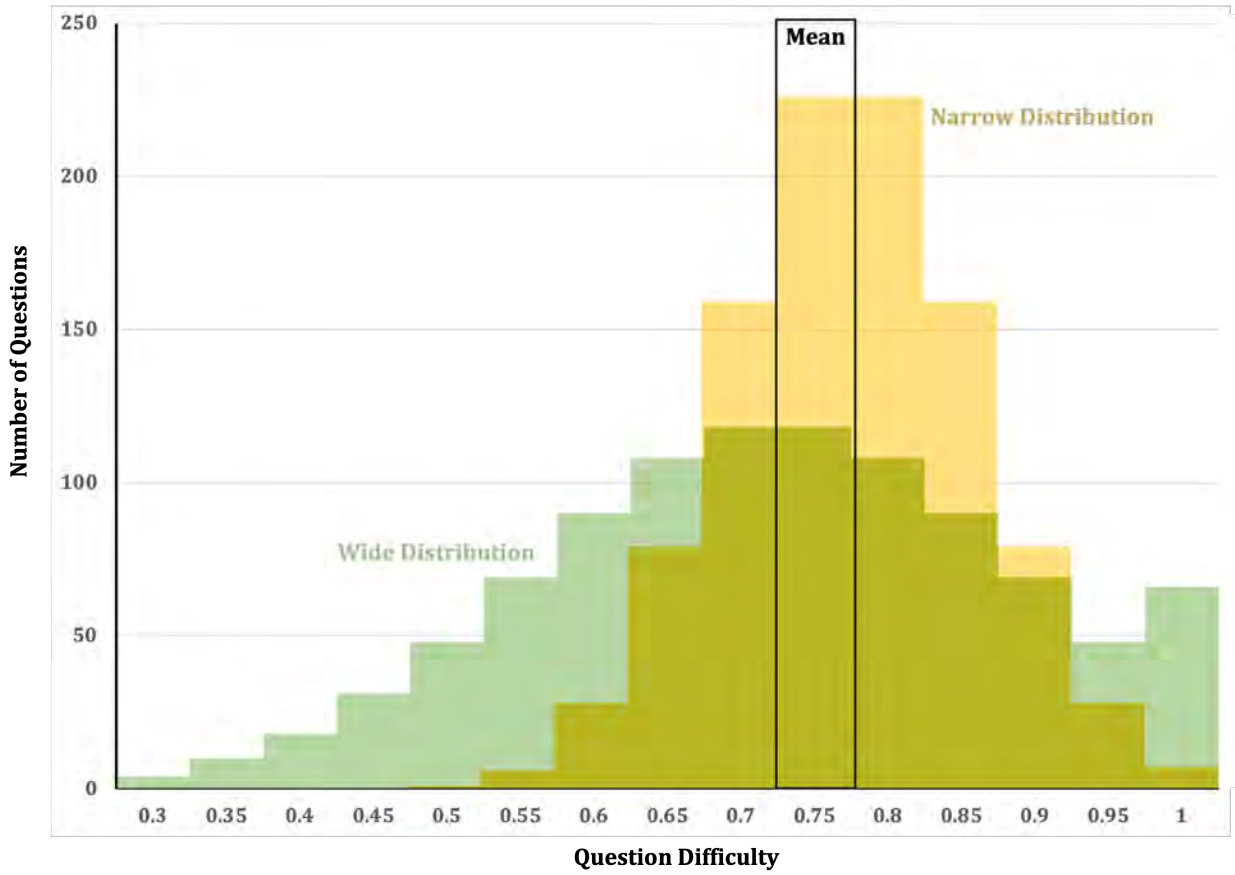

Figure S2. Behavior of a normal distribution for varying question difficulty. A normal distribution features a concentration of items around a mean, with the width of the distribution being described by the standard deviation. It can be used as the basis for curving a class grade, or for simulating student preparation or question difficulty. Here, two overlapping distributions are shown, both generated by using the NORM.INV function in Excel to create 1000 questions with an average difficulty of 0.75 (the mean). The graph shows how many of these questions falls within a certain difficulty rating, either less than 0.3, to each 0.05 interval (For example, the "0.35" category shows all questions that falls between 0.3 and 0.35). The narrow distribution (yellow) was generated using a standard deviation of 0.083, which is the same used for the questions in Table 1, row B. The wide distribution (green) was generated using a standard deviation of 0.166, which is the same used for the questions in Table 1, row C. The distributions are shown overlapping to facilitate the comparison between wide and narrow. Note that the right edge of the distribution has been compressed, showing all values greater than 1 in a single column.

Table S1. Example data for 33 students taking a 20 question quiz. For some questions, there is one correct answer, worth 1 point. Other questions allow for partial credit, awarding a fraction of the point. The average score for all students across the test was a 13.54 (67.7%). Students that scored above this average are bold and shaded in green.

	Question																			
Student	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
1	1	0	0	0.67	1	1	1	0	0.5	0	1	1	1	1	0.33	0	1	0.75	0	0.5
2	**1**	**1**	**1**	**1**	**0**	**0**	**1**	**1**	**0**	**0**	**1**	**1**	**1**	**1**	**0**	**1**	**1**	**1**	**1**	**0.25**
3	**1**	**1**	**1**	**1**	**0**	**1**	**1**	**1**	**1**	**1**	**0**	**1**	**1**	**1**	**0.33**	**1**	**1**	**0.5**	**0**	**0.75**
4	1	1	1	1	1	0	0	1	0.5	1	0	1	0	0	1	1	1	0.25	1	0
5	1	1	1	0.67	1	1	1	1	0.5	1	1	0	0	0	0	0	0	0.25	0	0.25
6	**1**	**0**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**0**	**1**	**1**	**0**	**1**	**1**	**1**	**1**	**0**	**1**
7	**1**	**0**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**0**	**1**	**1**	**0**	**1**	**1**	**1**	**1**	**0**	**0.75**
8	**0**	**1**	**0**	**0.67**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**0.33**	**1**	**1**	**0.5**	**0**	**0.5**
9	**1**	**0**	**0**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**0**	**1**	**0**	**1**	**1**	**1**	**1**	**1**	**1**	**0.75**
10	**1**	**0**	**0**	**0.67**	**0**	**1**	**1**	**1**	**0.5**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**0.75**
11	1	0	0	0.67	1	1	1	0	0	1	1	0	1	1	0	0.67	1	0.5	1	0.75
12	**1**	**1**	**1**	**0.67**	**1**	**1**	**1**	**1**	**1**	**1**	**0**	**1**	**1**	**1**	**0.33**	**1**	**1**	**0.25**	**0**	**0.5**
13	1	0	1	0.67	0	1	0	1	0	1	0	1	1	1	1	0.67	1	0.75	0	0.75
14	1	0	1	0.67	1	1	0	1	0	1	0	1	1	1	0.33	0.67	1	0.25	0	0.75
15	0	0	1	0.33	0	1	0	0	1	1	0	0	0	0	0.33	1	1	0	0	0.5
16	0	0	1	1	1	1	0	1	0.5	0	1	1	0	1	0	1	1	1	1	0.5
17	1	0	1	1	1	1	1	1	0.5	1	0	1	1	0	0	1	1	0.5	0	0.75
18	**0**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**0.33**	**1**	**1**	**0.75**	**0**	**0.25**
19	**1**	**0**	**1**	**1**	**1**	**1**	**0**	**1**	**1**	**1**	**0**	**1**	**1**	**1**	**0**	**1**	**1**	**0.5**	**0**	**0.75**
20	1	0	0	0.67	0	1	0	1	0	0	0	0	1	0	0	0.67	0	0	0	0.75
21	**1**	**0**	**0**	**1**	**1**	**1**	**1**	**1**	**0.5**	**1**	**1**	**1**	**1**	**1**	**0.33**	**1**	**1**	**0.75**	**1**	**0.75**
22	**1**	**0**	**1**	**1**	**0**	**1**	**1**	**1**	**1**	**1**	**0**	**1**	**1**	**1**	**0.33**	**1**	**1**	**0.75**	**0**	**0.5**
23	**1**	**1**	**1**	**0.33**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**0.33**	**1**	**1**	**0.75**	**1**	**1**
24	1	0	0	1	1	1	0	1	0	1	0	1	1	0	0.33	1	1	0.5	0	0.5
25	0	0	1	1	1	1	0	1	1	1	1	1	0	1	0.33	0	1	0.5	0	0
26	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**0**	**1**	**0**	**1**	**1**	**1**	**1**	**1**	**0**	**0.75**
27	1	0	1	0.67	1	0	1	1	0	1	1	1	0	1	0.33	0	1	0.75	0	0.75
28	1	1	1	0.67	1	1	0	1	0	1	1	1	0	0	0	0.33	0	0	1	0.25
29	0	0	0	0.67	0	0	1	0	1	1	1	1	1	0	0.33	1	0	0.5	0	1
30	1	0	1	0.33	1	1	0	1	1	1	0	1	1	1	0	0	1	0.5	0	0.75
31	1	0	1	1	0	1	0	1	1	1	0	1	0	1	1	1	1	1	0	0.75
32	**1**	**0**	**1**	**0.67**	**1**	**1**	**0**	**1**	**1**	**1**	**0**	**1**	**0**	**1**	**0.67**	**1**	**1**	**0.75**	**1**	**0.75**
33	**1**	**0**	**1**	**0.67**	**1**	**1**	**1**	**1**	**1**	**1**	**0**	**1**	**1**	**1**	**0**	**1**	**1**	**0.75**	**0**	**0.75**

Table S2. Calculations for class performance on the first 10 questions from the example quiz. For each question, the average score (S) was found by totaling all of the points obtained by the class and dividing it by the class size. This is usually reported as the difficulty metric in item analysis. The power function uses the exponent F/St; here, F was 2 and St was set at 0.35. These values are displayed in scientific notation due to their small size; 6.E-05 indicates 6 x 10-5, or 0.00006. Note that the bonus obtained for most questions is small, since class performance was good. However, the class collectively struggled with question 2, which yields a relatively higher bonus of 0.1. Discrimination is calculated as described elsewhere in the supplementary text. The discrimination logistic function here used a threshold value (Dt) of 0.15. The hybrid bonus is a multiplication of the power function by the discrimination logistic function, as shown in Figure 1, panel B. Only the first 10 questions are shown here for clarity. However, if the entire dataset is processed, a total hybrid bonus of 0.33 should be obtained, raising the class average to 13.88 out of 20 (69.4%).

	Question									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Total Points; T	27	10	24	26.33	24	29	20	29	21.5	29
Average Score; S = T/33	0.82	0.3	0.73	0.8	0.73	0.88	0.61	0.88	0.65	0.88
Power Function Bonus; = (1-S)$^{F/St}$	6.E-05	1.E-01	6.E-04	1.E-04	6.E-04	6.E-06	5.E-03	6.E-06	2.E-03	6.E-06
Top Performing Student Total Score	14	7	12	13.67	12	15	14	16	14	15
Bottom Performing Student Score	13	3	12	12.67	12	14	6	13	7.5	14
Discrimination	0.06	0.24	0	0.06	0	0.06	0.48	0.18	0.39	0.06
Discrimination Logistic Function	1.01	0.84	1.07	1.01	1.07	1.01	0.63	0.89	0.71	1.01
Hybrid Bonus	6.E-05	1.E-01	6.E-04	1.E-04	6.E-04	6.E-06	3.E-03	5.E-06	2.E-03	6.E-06