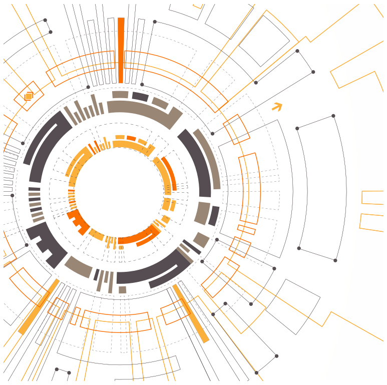


Abstract

Advanced practices for summative exam development and post-exam analysis are proven to be effective but aren't always practical, and, even when these are applied to some degree, exams remain inherently imperfect measures of student ability. Instructors may thus deem it necessary to adjust overall exam scores to account for aspects of an exam that may have been ill-suited to some or all students, and often these adjustments are made in an ad hoc and/or uninformed manner. This paper reviews reasons and methods for adjusting exam scores and proposes a new method that was developed organically from observation, reflection, and literature consultation. The scaling method considers that some of the underlying reasons for adjusting exam scores may affect certain sets of students more than others and seeks to incorporate this proposition while also avoiding weaknesses of other methods.

AUTHORS
Ryan K. Orchard, MASc
Macewan University

A Review Of Practices For Adjusting Exam Scores And A Proposed Nonlinear Scaling Method

Exams are inherently imperfect measures of student ability. Practices and tools exist to improve exam validity and reliability, but, even if these were convenient and accessible to the common higher education instructor and widely-used, the exams would still be prone to some degree of measurement error. It is postulated here that it is not uncommon for higher education instructors in ordinary exam settings, such as non-standardized exams developed and graded by the instructor, to make a post-exam adjustment to scores in a manner based on their own intuition. The current paper does not condone this practice, but rather it recognizes it as a reality, reflects upon reasons and methods for adjusting exam scores, and proposes a new method that was developed organically in response to years of observation while manually grading more than ten thousand exams (and counting).

CORRESPONDENCE
Email
orchardr@macewan.ca

The following will first provide context and scope, followed by a discussion of relevant considerations and best practices for exam development, particularly as they relate to exam validity and reliability in the given context and how they imply reasons for exam score adjustment and inform the choice of method. Next, methods for adjusting exam scores will be reviewed and critiqued. A new method will then be described and compared to the other methods, and implementation advice will be offered. A conclusion and discussion of limitations will end the paper.

Context and Scope

Grade-Curving Versus Grade-Scaling

Norm-referencing exam scores refers to the practice of referencing individual scores to those in a group or to a pre-defined grade distribution and is more common in situations where students are ranked for some purpose such as admission or awards (Kibble, 2017; Kulick & Wright, 2008), or for credentialing (Ben-David, 2000). *Grading on a curve* (*curving grades*, *grade-curving*) is a form of norm-referencing and generally refers to the practice of fitting exam grades to the normal distribution (the 'curve') to achieve a pre-determined proportion of students falling into each grade category (Tan et al., 2020). Grade-curving is a common practice in higher education (Kulick & Wright 2008), but is also somewhat controversial - see, for example, Grant (2016), who also noted that the practice may be used as a foil to grade inflation, and for this reason (and/or others) may be mandated by an institution. Kulick and Wright (2008) show that grade-curving can lead to the inclusion of luck as an unwanted factor in the partitioning of students into grade categories based on relative performance, which can have significant negative consequences for some students. Overall, there are strong arguments that grading on a curve tends to have more flaws than benefits (see, for example, Close, 2009), and, except for specific applications, should not be practiced.

It is still necessary to explore the why, as a means to inform the how.

For the current paper, the term *scaling*, as applied to exam scores, will refer to the general practice of applying a mathematical operation to adjust scores (usually upward) in a direct manner that does *not* compare individual scores to the group (as grade-curving does). Although the terms *curving* and *scaling* are used interchangeably in some of the literature, the current paper will differentiate between them as described and will focus only on grade-scaling.

Application

This paper assumes a context and scope in which the exams are developed, delivered, and graded (perhaps with assistance) by individual instructors in higher education, as opposed to settings such as large-scale standardized exams. This is an important stipulation given the realities of a potential lack of best practice components for exam development such as development blueprints, peer review processes for improvement, and detailed item analysis statistics (all discussed in the next section), but also perhaps due to time constraints experienced by individual instructors. Indeed, it seems difficult to deny that at least some exam authoring and administration in higher education may be performed using imperfect processes and without extensive training. Anderson (2018) notes that single-task grades (i.e., for an individual piece of work) tend to be very unreliable (p. 9), and that there can be significant disparity between teachers' opinions about what is fair grading and what is not (p. 18). As mentioned in the introduction, the current paper does not dismiss the value of using tools and best practices for exam quality improvement, but rather it assumes that an absence of these may be the case in many settings and that instructors are adjusting overall exam scores according to their judgement, and thus aims to better inform this inevitable process.

The score-adjustment methods presented in this paper are for adjusting *total individual exam scores*, not individual items (questions), although methods for exam score adjustment could be used in combination with (after applying) measures to address any issues with individual items. The discussion of exam scores will refer to the *numerator* as the total score for the responses on an individual exam and the *denominator* as the total (maximum) possible score for the numerator.

Background - Reasons For Adjusting Exam Scores

The topic of exam score adjustment (reasons or methods for) has very little explicit coverage in the literature, and thus most of the background provided here will be based on literature concerning topics such as item and exam development, as well as observations and reflections of the current author. Although the objective of the current paper is to explore *how* exam scores can be adjusted, it is still necessary to explore the *why*, as a means to inform the *how*; indeed, the new method proposed in this paper arose organically from years of observation and was developed to specifically address the perceived *why*. Embedded within the why-how relationship is the question of whether the underlying reasons affect some students differently than others, which could justify differences in score-adjustment depending on the raw score; the current paper will argue for a scaling method that adjusts the higher exam scores of students for whom the exam may have been better suited differently than it adjusts lower exam scores.

The reasons for exam score adjustment will be organized into a few general categories and discussed in turn below: individual items (questions); overall exam composition (distribution of question difficulty, and coverage of learning objectives and cognitive skill levels); and student interpretation and other student-related factors affecting performance. The terms *validity* (whether a test measured what it is supposed to measure) and *reliability* (whether a measurement result can be reproduced) may be used during discussion of individual items and/or exams (definitions provided are based on Kibble, 2017).

Individual Items

Malau-Aduli and Zimitat (2012) found that "there is considerable evidence that multiple-choice questions (MCQs) are poorly written" (p. 920) and went on to show that a peer review process for exams comprised of MCQs can have an improving effect on the quality of an exam, in many aspects. Kibble (2017) discovered that "faculty members cannot reliably judge the difficulty of individual items they write" (p. 111) and recommended faculty development, peer review in test development, and being deliberate in matching learning objectives and exam items (the latter two topics are discussed further in the next subsection). Indeed, exam authoring can be complex and time-consuming, particularly when it comes to writing MCQs that are of high quality, and increasingly so when higher cognitive abilities (beyond memorization) are to be evaluated, as described by Stevens et al. (2022). These authors provide a thorough framework to guide the authoring of MCQs; for example, questions that are based on a vignette (scenario) should be well-edited to be "unambiguous, concise, and readable" (p. 6), so as not to affect the student's comprehension of the question [which still might not account for student-specific factors, such as language differences, or even a propensity for some to overthink]. Adhi and Aly (2018) conducted a study in which they found that MCQs of the *one-best* type performed better than those of the *one-correct* type, in terms of student scores, reliability and discrimination ability. The depth and intricacies of the principles described by these (and other) authors implies that many exams, whether employing these practices or not, likely include at least some questions that are not sound and may affect the performance of some students.

Will argue for a scaling method adjusts the higher exam scores of students for whom the exam may have been better suited differently than it adjusts lower exam scores.

Depending on how an exam is administered and graded, there may be tools available for evaluating the *individual items* of an exam. Camenares (2022) encouraged using *item analysis*, which includes measures of question difficulty (percentage success rate for each question, over all students) and discrimination index (the relative success on a given question of students who were high-performing versus low-performing on the exam). If a specific unfair question is identified, it may be necessary to remove it entirely from the exam grade calculations, such as in the case where all students answered incorrectly, or remove it only for those students who answered incorrectly (which doesn't penalize those that did answer correctly and have it counted in the numerator and denominator, unless the grades are norm-referenced). Another option is to adjust student scores on individual questions using a "proportional bonus for questions that are too difficult and/or poor discriminators" (Camenares, 2022, p. 2). Rudolph et al. (2019) suggested that, in some cases, full credit for a poorly performing item could be awarded to all students, regardless of their answer on the question, or as a bonus for those that did get it correct (although in a later discussion of exam-level grade adjustments the current author will argue against providing a grade for something that doesn't deserve a grade). Rudolph et al. also contended that 'examinations are likely to contain quite a few flaws' (p. 1502), which (in the eyes of the current author) may warrant *exam-level* score adjustments in the absence of identification of and adjustment for specific poor items on the exam. It would also seem that an overall exam score adjustment may be less necessary for the students who answered most of the questions correctly, since they may have been less affected by any poorly constructed questions; this premise is modeled by the current proposed exam score scaling method.

Exam Composition - Distribution Of Question Difficulty

Most students correctly answer questions with low difficulty indexes (i.e., easy questions), while only the top students tend to correctly answer those with high difficulty indexes (Downing, 2009). Malau-Aduli and Zimitat (2012) alluded to "the theory that the most informative test items are those of middle difficulty and they provide higher discrimination between the high-scoring and low-scoring examinees" (p. 921). Thus, even with the help of detailed item analysis, it may be that no particular question(s) can be identified as poorly constructed, but rather the *overall exam* may have had too high a proportion of difficult questions and a low average score, which would not have disadvantaged the top achievers as much as the others (and thus may not warrant as much of a grade-adjustment). Further, assuming that there were at least *some* easy questions, any very low exam scores (relative to other exam scores) may be more attributable to a lack of preparation by some students than shortcomings of the exam; a consequence of the proposed method is that very low exam scores do not receive as large of grade increases as those in the middle (as will be described with the proposed method), which may have some justification.

Exam Composition - Coverage Of Learning Objectives And Cognitive Skill Levels

Exam blueprints are outlines and plans that "guide item writers to develop sufficient items that cover important content areas and objectives at the suitable cognitive level" (Eweda et al., 2020, p. 166). Abdellatif et al. (2024) promote blueprinting as a tool to combat two major threats to exam validity - construct underrepresentation (some course content not being appropriately represented on the exam), and construct irrelevant variation (question format, questions being too easy or difficult, or the test modality being inaccurate); Kibble (2017) adds the issue of "language cueing test-wise students to the correct answer and guessing from limited option sets" (p. 115), which speaks to the premise of the current paper that any overall score adjustment should be less necessary for those whom the exam was better suited (i.e., the 'test-wise' students who may not have been harmed even by poorly constructed items).

Using a blueprinting (or other) process doesn't mean perfect exams, however. Welch et al. (2017) found that "the accuracy and reliability of [faculty member's] ability to categorize" Bloom's Taxonomy to exam questions were low (p. 103). Omar et al. (2012) also recognized the importance of balancing lower and higher cognitive level questions, and the proper classification using Bloom's, and provided preliminary results indicating that an automated system may be useful in assisting with classification. Wellberg (2023) pointed out that mathematics teachers "tend to conflate difficulty and cognitive complexity" (p. 58). In the present context, these works suggest that if a good exam requires a good balance of question cognitive levels, among other factors, then the imperfect ability for any instructor to accurately categorize questions by cognitive level may mean that an exam in which the balance is deemed to have been questionable (based on instructor judgment, post-exam) may warrant score adjustments. Again, this would seem less likely to have adversely affected those students who had high raw test scores (before adjustment), and the proposed scaling method accounts for that.

Student Interpretation

During an exam that was developed and administered by the current author, a student annotated their thoughts on many of the multiple-choice questions (by their own accord). On a few questions it was clear that the student was knowledgeable about the concept but chose what the instructor considered an incorrect answer, including one particular question in which some of their annotations were evidence that they had been distracted by a single word in the question stem, but that they clearly understood the target concept. It was difficult to give a score for the question since a grader wouldn't be able to follow the same procedure on all of the student's incorrectly-answered questions (much of the annotation was difficult to follow) or for all students (most others didn't annotate their exam at all), or to analyze annotations on correctly-answered questions to confirm understanding of the target concept. The specific question had been used for years, and generally had a high success rate, and it seemed unlikely that any formal item analysis would have flagged it, despite the clear trouble that it gave one otherwise knowledgeable student. Another example (personal): once upon returning home

A consequence of the proposed method is that very low exam scores do not receive as large of grade increases as those in the middle.

from a walk the current author asked of their teenager “which of the following people that you know did we see at the dog park today – (a) your former pre-school teacher Ms. H, (b) the father of your friend L, (c) professional athlete RNH, or (d) all of the above?” The teenager answered ‘a’, and upon being told that the correct answer was ‘d’, they replied “but I don’t *know* RNH” [they only knew *of* them]. It can be surprising to a question author to find that a single and possibly extraneous word can be interpreted by one or more students in a very literal way and have unintended consequences, and this may even occur without the author ever knowing.

Noble et al. (2012) showed that a student’s answers don’t always reflect their knowledge; specifically, they interviewed students from different groups to evaluate target knowledge (whether the student understood the concept), which in some cases was clearly demonstrated despite the student having chosen an incorrect answer. They noted that “the difficulty [for the examinee]…was not in understanding [the target concept], but in interpreting the language of the test item and in creating a context in which the language of the item made sense to them” (p. 792). They also found that students from certain groups such as low-income households and English Language Learners were more likely to make an incorrect answer choice despite understanding the concept. They point to other studies in which answers differ between students not because of differences in knowledge or ability but because of the “interaction between students’ language and life experiences and the structure and content of test items” (p. 781). According to Fencl (2019), non-academic factors such as exam anxiety and language barriers can also affect exam performance. Thelk (2008) observed that students in one group can have a greater probability of correctly answering a question than students in another group, even after controlling for ability, due to bias at the item (question) level. Crawford and Fekete (1997) found that

A measured application of exam-score adjustment, when deemed appropriate, doesn’t seem out of line

> an instructor’s expectations of the skills needed to answer a question were sometimes inaccurate…[and in]…some cases the most important skill that determined a student’s grade on the question [was]…coping with distraction… much of the class did not even realize what key concepts were involved in the question, but instead they were distracted by irrelevant information (p. 188),

while Holmes (2021) conducted a study in which they found evidence that most students believe that some can simply be ‘bad test-takers’, even when they know the course material, and noted that identifying as such can have a strong association with exam anxiety (p. 297).

It seems, then, that although there may be many practices to avoid the shortcomings of exams, a measured application of exam-score adjustment, when deemed appropriate, doesn’t seem out of line, nor do minor differences in how it affects raw scores at different points in the overall distribution, as will be seen in the proposed method.

Other Motivations For Exam Score Adjustment

Finally, in many settings, instructors are free to use their own intuition and judgement to guide exam authoring and grading, and may judge an overall average exam grade to be too low for reasons of their own. Additionally, career pressure on instructors may motivate them to give high grades (and thus bump low exam results), as described by Wellberg (2023) as well as Jephcote et al. (2021), who noted that “the continued emphasis on the power of student evaluations may provide instructors with an incentive to…conform to grade leniency” (p. 549). While the current paper doesn’t endorse these as valid reasons, if one is going to adjust grades, they should at least be knowledgeable of the different methods.

Exam Score Adjustment Methods

The following will describe methods for exam score adjustment that are carried out *after* any adjustments for specific individual question deficiencies (i.e., methods for scaling *total* exam scores).

Three mock datasets of 1,000 exam scores were simulated using a random number generator to create desired distributions, one approximately normally-distributed, another with a bimodal distribution, and one uniformly-distributed. It was found that the distribution (normal, bimodal, or uniform) of raw exam scores had very little effect on the relative outcomes of the scaling methods, so for the sake of simplicity, the following analysis will only use the normally-distributed simulated exam scores. The mean and median of the mock exam scores were both 65%, with a standard deviation of 11.88%, a maximum grade of 100%, and a minimum grade of 28%. For the sake of a fair comparison, each of the scaling methods were calibrated so that the new (scaled) mean and median were 70% (up from 65% for the raw scores).

The simplest method for adjusting exam scores is to add a *constant value* to all scores (e.g., raise all scores by 5%, such that a 34% score becomes 39% and a 94% becomes 99%). This is obviously very simple to implement and for the exam-takers to understand, and may be considered by many to be fair since the amount of grade increase is equal across all students. Drawbacks include that *equal* and *fair* aren't necessarily the same under all circumstances (i.e., there may be reasons that some students deserve a different adjustment than others, as described previously). Further, 0% and 100% can be distorted - a blank exam would get 5% (i.e., grades are being *gifted*, despite no evidence of meeting learning objectives), and a student who did not answer all exam questions correctly could get an exam score of 100% (or higher). An example is provided in Table 1.

Another simple and common method is to *reduce the denominator* used for converting the scores to percentages, which effectively raises all exam scores (except in the case of a zero raw score). Aside from simplicity, this method has no other documented strengths while it does have some drawbacks, including that a score of 100% or higher is possible for a student that did not answer all questions correctly, and, in the words of Nelson et al. (1992), it "robs from the poor to give to the rich" (p. 463) in that already high exam scores will see a significantly larger increase (adjusted score minus raw score) than lower exam scores. For example, if raw scores are out of 40 and the denominator is adjusted to be 38, then 90% (36/40) becomes 94.7% (36/38), while 50% becomes 52.6% (20/38) and 10% becomes 10.5% (4/38). Although it seems difficult to imagine a scenario where the ultra-high scores should benefit the most from scaling, as happens with this method, it may remain common practice due to convenience, and because the quantitative implications aren't fully understood by some users. Table 1 provides scaled scores using this method for the mock exam scores, calibrated to achieve the desired mean by reducing the denominator from 100 (raw score) to approximately 92.86.

Equal and fair aren't necessarily the same under all circumstances.

Another method is *square root* scaling whereby the square root of the raw score out of 100 is multiplied by 10. 100% will still be 100% and 0% will still be 0%, but scores in between the extremes will be scaled in a way that will benefit "students who need it the most without removing incentive for higher performing students" (Page et al., 2018, p. 6). These authors propose the application of this method as a remedy for the disadvantaged underserved populations in STEM classes. The square root scaling method has a compelling property to it – it raises the scores in the middle portion of the distribution by more than those in the upper and lower sections (see Table 1), which may be justifiable (as described in a previous section of the current paper). However, this method does have a problem that may make it unusable – it raises the grades by too much: for example, 50% becomes 70.71% and 25% becomes 50%. This problem can be overcome by using a different degree (root), which was done for the mock dataset to achieve a 70% average by using a root of 2.14 instead of 2, but it required an optimization tool (Solver or Goal Seek in Excel) and it resulted in some raw scores at the higher end being adjusted negatively, which may make it ill-advised. Table 1 reports scaled values using a square root (and thus don't calibrate to the same mean as the other methods) for the sake of simplicity and to demonstrate the shape of the grade increases.

Maloy (1990) finds issue with adding a constant value to all scores and they propose a nonlinear strategy similar to the square root method in that it raises lower grades by more than it raises higher ones, except for very low grades. The equation is

$$S=100^{1-n} \bullet R^{n} \quad (1)$$

where S is the scaled score (out of 100), R the raw score (out of 100), and n a scaling parameter between zero and one which can be determined using a log-based equation and substituting a desired corresponding pair of values for S and R (for example with R being the *actual* raw score mean, median or passing score and S the *desired* mean, median or passing score). Results of this method are shown in Table 2, where n was calibrated to a value of 0.8222

Table 1
Amount of grade increase (scaled exam scores over raw scores) under three scaling methods – constant increase, denominator-reduction (provides larger score increases for higher scores) and square root (favours lower scores more than higher ones, except for very low scores)

	Constant		Reduce denominator		Square Root	
Raw Score (R)	Scaled Score (S)	Increase (S-R)	Scaled Score (S)	Increase (S-R)	Scaled Score (S)	Increase (S-R)
100	105.00	5.00	107.69	7.69	100.00	0.00
95	100.00	5.00	102.31	7.31	97.47	2.47
90	95.00	5.00	96.92	6.92	94.87	4.87
85	90.00	5.00	91.54	6.54	92.20	7.20
80	85.00	5.00	86.15	6.15	89.44	9.44
75	80.00	5.00	80.77	5.77	86.60	11.60
70	75.00	5.00	75.38	5.38	83.67	13.67
65	70.00	5.00	70.00	5.00	80.62	15.62
60	65.00	5.00	64.62	4.62	77.46	17.46
55	60.00	5.00	59.23	4.23	74.16	19.16
50	55.00	5.00	53.85	3.85	70.71	20.71
45	50.00	5.00	48.46	3.46	67.08	22.08
40	45.00	5.00	43.08	3.08	63.25	23.25
35	40.00	5.00	37.69	2.69	59.16	24.16
30	35.00	5.00	32.31	2.31	54.77	24.77
25	30.00	5.00	26.92	1.92	50.00	25.00
20	25.00	5.00	21.54	1.54	44.72	24.72
15	20.00	5.00	16.15	1.15	38.73	23.73
10	15.00	5.00	10.77	0.77	31.62	21.62
5	10.00	5.00	5.38	0.38	22.36	17.36
0	5.00	5.00	0.00	0.00	0.00	0.00

to increase the mean raw score of 65% for the mock data to a mean scaled score of 70%. (This model will be examined in more detail later when compared to the proposed method).

Becker (1991) responded to Maloy with a method that uses two scaling factors (a and b) that are determined based on the relationship between the highest raw score and desired highest scaled score, as well as the mean raw score and desired mean scaled score (i.e., *a* and *b* are found by simultaneously solving two equations); individual scores are then scaled using the equation

$$S=R \bullet a+b \quad (2)$$

where *S* and *R* are the scaled and raw scores (out of 100), respectively. The value of a will be less than one and will scale the raw grade downward before a constant b is added. This method was calibrated using the mock exam data and setting the maximum raw and scaled scores to both be 100 and the raw and scaled mean scores to be 65% and 70%, respectively, resulting in values of *a*=0.8571 and *b*=14.2857. Table 2 shows that this method benefits a zero score the most, with the difference between scaled and raw scores becoming less as raw scores increase, which avoids compensating already high scores by more than low ones, but results in the largest score increases being applied to those who may have been the least prepared for the exam.

Bailey (1992) provides a method that raises "lower grades more than higher ones but without advancing average students into the A range" (p. 221), using the equation

$$S=100–(W \bullet n) \quad (3)$$

where *S* is the scaled score and *W* is the total points possible (100, if a percentage) minus the raw score (*R*), which effectively makes *W* the number of lost marks. *n* is a scaling factor that can be determined by a formula and using the specific values of a desired *R*→*S* transformation (in the mock data, a 65%→70% transformation resulted in *n*=0.8571). Results from this method were identical to those of Becker (i.e., both models increase lower scores by more than higher scores and follow the same general shape, as per Table 2), which will be the case when the Becker model is calibrated using values for max *R* and max *S* of 100%.

Proposed Method

The proposed scaling method was developed in response to observations of general exam deficiencies (described previously), as well as perceived deficiencies in the basic scaling

Table 2

Comparison of amount of grade increase for Maloy, Becker, and Bailey models

	Maloy		Becker		Bailey	
Raw Score (R)	Scaled Score (S)	Increase (S-R)	Scaled Score (S)	Increase (S-R)	Scaled Score (S)	Increase (S-R)
100	100.00	0.00	100.00	0.00	100.00	0.00
95	95.87	0.87	95.71	0.71	95.71	0.71
90	91.70	1.70	91.43	1.43	91.43	1.43
85	87.49	2.49	87.14	2.14	87.14	2.14
80	83.24	3.24	82.86	2.86	82.86	2.86
75	78.94	3.94	78.57	3.57	78.57	3.57
70	74.58	4.58	74.29	4.29	74.29	4.29
65	70.17	5.17	70.00	5.00	70.00	5.00
60	65.71	5.71	65.71	5.71	65.71	5.71
55	61.17	6.17	61.43	6.43	61.43	6.43
50	56.56	6.56	57.14	7.14	57.14	7.14
45	51.87	6.87	52.86	7.86	52.86	7.86
40	47.08	7.08	48.57	8.57	48.57	8.57
35	42.18	7.18	44.29	9.29	44.29	9.29
30	37.16	7.16	40.00	10.00	40.00	10.00
25	31.99	6.99	35.71	10.71	35.71	10.71
20	26.63	6.63	31.43	11.43	31.43	11.43
15	21.02	6.02	27.14	12.14	27.14	12.14
10	15.06	5.06	22.86	12.86	22.86	12.86
5	8.52	3.52	18.57	13.57	18.57	13.57
0	0.00	0.00	14.29	14.29	14.29	14.29

methods used by some instructors, such as an adjustment that gives all students the same increase (which can result in a grade of over 100% and/or give free grades where they may not be earned or warranted), or an across-the-board reduction in the exam score denominator (which increases the grades of students that already had the highest grades by significantly more than it increases the lowest grades). The premise for the new method is that it is possible that an exam may have been *better-suited* to the students that had the highest scores (in addition to the fact that they may also have better met the learning objectives), thus a score-scaling method should scale scores in a nonlinear manner. At the same time, the scaling method should be such that the scaled grades maintain the same rank order as the raw scores, and two raw scores that have minimal difference should result in two scaled scores that also have minimal difference.

In the proposed method, the reduction in the denominator for an individual exam score is scaled according to the number of incorrect answers on the exam.

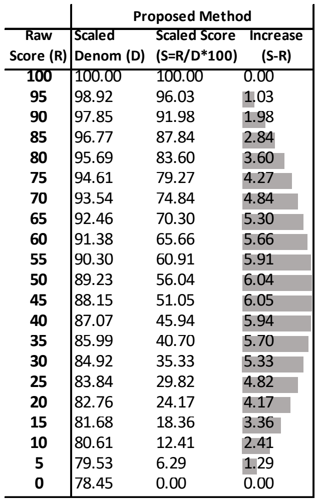
In the proposed method, the reduction in the denominator for an individual exam score is scaled according to the number of incorrect answers on the exam (i.e., a perfect score will not have a reduction in the denominator, scores with very few incorrect answers will see small reductions in the denominator, and scores at the bottom end will see the largest denominator reduction). Although the lowest raw scores receive the largest reduction in the denominator, they also have the fewest correct answers (in the numerator) and don't benefit as much from a denominator reduction, and so the largest increases in score (scaled minus raw) occurs for the scores near the middle. Described another way, the upper scores are helped less than the middle scores because the reduction in the denominator is less, which is justified by the premise that the exam may have been better suited for them, among other factors, while the lower scores are increased by less than the middle scores are increased, although they are helped by an even larger denominator reduction, because they didn't help themselves, so to speak, by having enough correct answers for the larger denominator reduction to make much difference. Grades are not bumped just to be bumped, but rather they are *scaled*, based on the number of correct and incorrect answers. The result is shown in Table 3 and has a shape similar to the square root and Maloy methods.

The formula is as follows, where S is a scaled score (out of 100), R is the raw score out of n total possible points, and n^* an adjusted total possible points that affects the magnitude of the scaling:

$$S = \frac{R}{n - (n - n^*)(\frac{(n-R)}{n})} \times 100 \qquad (4)$$

The denominator (which will simply equal n in the case of unscaled grades) is reduced from n by an amount (n-n^*) that is scaled by the proportion of questions that were incorrect ((n-R)/n). As described previously, although lower raw scores receive a larger denominator

Table 3
Scaled exam scores using the proposed method, which scales the amount that the denominator is reduced by a factor that incorporates the raw score



	Proposed Method		
Raw Score (R)	Scaled Denom (D)	Scaled Score (S=R/D*100)	Increase (S-R)
100	100.00	100.00	0.00
95	98.92	96.03	1.03
90	97.85	91.98	1.98
85	96.77	87.84	2.84
80	95.69	83.60	3.60
75	94.61	79.27	4.27
70	93.54	74.84	4.84
65	92.46	70.30	5.30
60	91.38	65.66	5.66
55	90.30	60.91	5.91
50	89.23	56.04	6.04
45	88.15	51.05	6.05
40	87.07	45.94	5.94
35	85.99	40.70	5.70
30	84.92	35.33	5.33
25	83.84	29.82	4.82
20	82.76	24.17	4.17
15	81.68	18.36	3.36
10	80.61	12.41	2.41
5	79.53	6.29	1.29
0	78.45	0.00	0.00

The square root method, Maloy's method, and the proposed method of the current paper are the only scaling methods that provide larger increases for scores toward the middle than for those at the high and low ends.

reduction, this won't always result in a larger adjustment to the total grade, since the scaled grade also depends on the number of correct answers (*R*, in the numerator). In other words, there must be incorrect answers for there to be a score adjustment, but there also must be correct answers to create mass for the denominator reduction to make a difference in the scaled score.

The value of *n** will be less than *n* (the maximum number of points available on the exam) and will not hold any intuitive meaning; it will be *approximately* double the difference between the raw score (R) and the scaled score (S) at the point where the difference between raw and scaled scores are largest. Lower *n** means that raw grades will receive a larger increase, and is set according to the desired scaling effect, which can be done a variety of ways, one being to specify the desired raw score (R) that should translate into a 50% scaled score, and solving the following equation (R is the raw score value out of *n* possible points or the raw percent grade where n would then be 100):

$$n^* = \frac{R \bullet n}{n - R} \quad (5)$$

For example, if n=100 and a 50% scaled score should be given to a raw score of R=45, then n*=81.82. A value for *n** can also be determined by using any numbers for *R* and *S* (in terms of a percentage out of 100) that give a desired *R*→*S* transformation and solving formula 6. Note that this method of calibrating *n** will result in a slightly different value than using equation 5, unless *S = 50* is used.

$$n^* = \frac{\frac{R}{S} \bullet 100 - R}{100 - R} \quad (6)$$

For example, using *R*=65% and *S*=70% gives a value of *n** of 79.59. (For comparison with the other scaling methods, the value for *n** was calibrated using the nonlinear solver in Excel so that the scaled mean of the mock dataset was 70%, which required using *n**=78.45). More advice for implementation is provided later in the paper.

Comparison and Discussion

As can be seen, the square root method, Maloy's method, and the proposed method of the current paper are the only scaling methods that provide larger increases for scores toward the middle than for those at the high and low ends. In looking at Maloy's method and the proposed method, the differences between the results are: (1) the magnitude of some of the adjustments (the largest score increases under the Maloy method are larger than those of the proposed method), and (2) which scores receive the most benefit for the mock dataset (scores in the 20-30% range for Maloy versus scores in the 40-50% range for the proposed method). These differences are shown graphically in Figure 1. Note that although the scaled scores of Maloy's method are quite a bit higher in some cases, the proposed method can achieve the same increase in the mean score (from 65% to 70%) with a smaller average increase in scaled versus raw scores since it has a slightly larger score increase for the higher frequency group (in other words, the Maloy method has significantly higher increases for scores at the lower end of the raw score scale, which in the mock data affected relatively few exams.

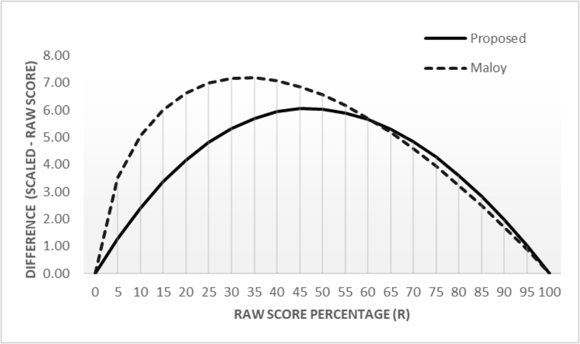
The proposed method is not being prescribed for all situations or claiming to be outright superior to all other methods.

Table 4
Comparison of scaled scores and increases under all scaling methods

	Constant		Reduce denom		Square Root		Maloy		Becker		Bailey		Proposed	
Raw Score (R)	Scaled (S)	Increase (S-R)	Scaled (S)	Increase (S-R)	Scaled (S)	Increase (S-R)	Scaled (S)	Increase (S-R)	Scaled (S)	Increase (S-R)	Scaled (S)	Increase (S-R)	Scaled (S)	Increase (S-R)
100	105.00	5.00	107.69	7.69	100.00	0.00	100.00	0.00	100.00	0.00	100.00	0.00	100.00	0.00
95	100.00	5.00	102.31	7.31	97.47	2.47	95.87	0.87	95.71	0.71	95.71	0.71	96.03	1.03
90	95.00	5.00	96.92	6.92	94.87	4.87	91.70	1.70	91.43	1.43	91.43	1.43	91.98	1.98
85	90.00	5.00	91.54	6.54	92.20	7.20	87.49	2.49	87.14	2.14	87.14	2.14	87.84	2.84
80	85.00	5.00	86.15	6.15	89.44	9.44	83.24	3.24	82.86	2.86	82.86	2.86	83.60	3.60
75	80.00	5.00	80.77	5.77	86.60	11.60	78.94	3.94	78.57	3.57	78.57	3.57	79.27	4.27
70	75.00	5.00	75.38	5.38	83.67	13.67	74.58	4.58	74.29	4.29	74.29	4.29	74.84	4.84
65	70.00	5.00	70.00	5.00	80.62	15.62	70.17	5.17	70.00	5.00	70.00	5.00	70.30	5.30
60	65.00	5.00	64.62	4.62	77.46	17.46	65.71	5.71	65.71	5.71	65.71	5.71	65.66	5.66
55	60.00	5.00	59.23	4.23	74.16	19.16	61.17	6.17	61.43	6.43	61.43	6.43	60.91	5.91
50	55.00	5.00	53.85	3.85	70.71	20.71	56.56	6.56	57.14	7.14	57.14	7.14	56.04	6.04
45	50.00	5.00	48.46	3.46	67.08	22.08	51.87	6.87	52.86	7.86	52.86	7.86	51.05	6.05
40	45.00	5.00	43.08	3.08	63.25	23.25	47.08	7.08	48.57	8.57	48.57	8.57	45.94	5.94
35	40.00	5.00	37.69	2.69	59.16	24.16	42.18	7.18	44.29	9.29	44.29	9.29	40.70	5.70
30	35.00	5.00	32.31	2.31	54.77	24.77	37.16	7.16	40.00	10.00	40.00	10.00	35.33	5.33
25	30.00	5.00	26.92	1.92	50.00	25.00	31.99	6.99	35.71	10.71	35.71	10.71	29.82	4.82
20	25.00	5.00	21.54	1.54	44.72	24.72	26.63	6.63	31.43	11.43	31.43	11.43	24.17	4.17
15	20.00	5.00	16.15	1.15	38.73	23.73	21.02	6.02	27.14	12.14	27.14	12.14	18.36	3.36
10	15.00	5.00	10.77	0.77	31.62	21.62	15.06	5.06	22.86	12.86	22.86	12.86	12.41	2.41
5	10.00	5.00	5.38	0.38	22.36	17.36	8.52	3.52	18.57	13.57	18.57	13.57	6.29	1.29
0	5.00	5.00	0.00	0.00	0.00	0.00	0.00	0.00	14.29	14.29	14.29	14.29	0.00	0.00

proposed method), and (2) which scores receive the most benefit for the mock dataset (scores in the 20-30% range for Maloy versus scores in the 40-50% range for the proposed method). These differences are shown graphically in Figure 1. Note that although the scaled scores of Maloy's method are quite a bit higher in some cases, the proposed method can achieve the same increase in the mean score (from 65% to 70%) with a smaller average increase in scaled versus raw scores since it has a slightly larger score increase for the higher frequency group (in other words, the Maloy method has significantly higher increases for scores at the lower end of the raw score scale, which in the mock data affected relatively few exams.

Figure 1
Increase in scaled score from raw score (S-R), for given raw score values, under the proposed method and the Maloy method



Implementation

The proposed method is not being prescribed for all situations or claiming to be outright superior to all other methods, but rather is being offered as an alternative that avoids some of the shortcomings of other methods, for consideration for use by an instructor based on their own reasons for adjusting exam scores.

Implementation requires determination of n^*. Presumably the instructor has analysed the exam results to conclude that grade-scaling is warranted in the first place. The instructor could then either decide on the raw score that should translate into a scaled score at the passing threshold (50%) and use equation 5, or determine R (raw score) and S (scaled score) values for a desired $R \rightarrow S$ transformation and use equation 6. Once the value of n^* has been determined, each grade can be individually converted using equation 4 and a calculator or spreadsheet (it may even be possible to set up a conversion formula in the LMS).

The practice of calibrating the scaling parameter (n^*) based on a raw score that is judged by the grader to be that which should convert to a passing (50%) scaled score has some similarities, albeit rudimental, to practices for *standard-setting* using cut-off points as described by Ben-David (2000) in a credentialling application. These methods, such as the Anghoff method which uses a panel of judges, consider the totality of the individual exam questions and identify a raw-score that would represent a passing grade for a "minimally competent examinee" (p. 122). Kibble (2017) also discusses best practices for standard-setting but acknowledges that these methods will not be practical in all situations (p. 117), and thus the relevant consideration for the current paper is that anchoring the scaling process with a thoughtful consideration of what raw score a student on the borderline of passing would achieve on the test (i.e., which becomes the scaled passing grade of 50%) is one straight-forward way to calibrate the proposed scaling method.

A final important aspect of implementation is transparency with students – a document or spreadsheet with the scaling formula can be posted for the students to see (and perhaps convert their own raw score into the scaled score), and/or the method can be explained at the time of review of the exam results in class. Student perceptions about the fairness (or other criteria) of this scaling method and others certainly seem a worthy area for further research.

Conclusion And Limitations

This paper began by discussing a variety of factors that can render an exam to be an imperfect measure of student ability, including some related to the construction of individual items (questions), some related to overall exam composition (including the distribution of question difficulty, and the coverage of learning objectives and cognitive skill levels), and some related to student interpretation. Although the current work does not suggest that efforts shouldn't be undertaken to improve exams and teaching, where applicable, instructors may nonetheless deem it necessary to adjust exam grades to account for the aforementioned inherent exam deficiencies and/or at the judgement of the individual instructor. It is hoped that any adjustments would be carried out in a careful and informed manner; however, some grade-adjustment methods may have drawbacks which may or may not be known to the user - for example, outright reducing the denominator of the exam by the same amount for all students will benefit higher grades more than lower grades, and can result in some grades exceeding 100%. One objective of the current paper is thus to collect and review documented processes for adjusting exam scores. Limitations include that only the methods available in the literature were reviewed (and this topic is sparsely studied), and that methods for *curving* grades according to a pre-defined distribution (e.g., the normal distribution) were not included, for reasons described in a previous section.

The proposed method is not being presented as a panacea; whether and how to use it is up to individual instructors, of course.

Based on the discussion of the underlying reasons that exams may be imperfect, which could affect some students more than others, as well as perceived deficiencies in grade-adjustment methods in the literature, the proposed method uses a reduction in the denominator that is scaled by the proportion of incorrect questions. This results in very high grades receiving small adjustments, due to having few incorrect questions and therefore minimal reduction to the denominator, very low grades receiving small adjustments despite having the largest reduction in the denominator, due to having few correct answers (causing a low numerator value), and grades towards the middle of the distribution receiving the largest adjustments, due to having a comparatively moderate reduction in the denominator as well as a large enough numerator for it to make a difference.

Limitations to this method include that it is based on the notion that high exam grades do not merit as much grade adjustment as grades towards the middle of the distribution, which reflects the intuition and experiences of the author, with implicit support by way of a discussion of reasons for exam deficiencies in the background section, but this basis may not align with the intuition and experiences of some examiners or examinees. Further, the degree of the correlation between raw grades and deserved adjustments is admittedly inexact (which is the nature of the beast in using examinations to measure student ability). However, the proposed method is not being presented as a panacea; whether and how to use it is up to individual instructors, of course.

References

Abdellatif, H., Alsemeh, A. E., Khamis, T., & Boulassel, M. R. (2024). Exam blueprinting as a tool to overcome principal validity threats: A scoping review. *Educación Médica*, *25*(3). https://doi.org/10.1016/j.edumed.2024.100906

Adhi, M. I., & Aly, S. M. (2018). Student perception and post-exam analysis of one best MCQ and one correct MCQs: A comparative study. *The Journal of the Pakistan Medical Association*, *68*(4), 570-575.

Anderson, L.W. (2018). A critique of grading: Policies, practices, and technical matters. *Education Policy Analysis Archives*. *26*(49), 1-27. https://doi.org/10.14507/epaa.26.3814

Bailey, L. C. (1992). Grade-scaling: A simplified approach. *Journal of Chemical Education*, *69*(3), 221. https://pubs.acs.org/doi/pdf/10.1021/ed069p221

Becker, C. E. (1991). Comparison of two methods for scoring examinations. *Journal of Chemical Education*, *68*(4), 309. https://pubs.acs.org/doi/pdf/10.1021/ed068p309

Ben-David, M. F. (2000). AMEE Guide No. 18: Standard setting in student assessment. *Medical Teacher*, *22*(2), 120-130. https://doi.org/10.1080/01421590078526

Camenares, D. (2022). Better Remedies For Bad Exams: correcting for difficult questions in a fair and systematic way. *International Journal for the Scholarship of Teaching and Learning*, *16*(3), 1-4. https://digitalcommons.georgiasouthern.edu/ij-sotl/vol16/iss3/4/

Close, D. (2009). Fair grades. *Teaching Philosophy*, *32*(4), 361-398. https://www.pdcnet.org/8525737F00588478/file/B08274525466752585257680005CC87/$FILE/teachphil_2009_0032_0004_0035_0072.pdf

Crawford, K., & Fekete, A. (1997, July). What do exam results really measure?. *In Proceedings of the 2nd Australasian conference on Computer Science Education* (pp. 185-190). https://dl.acm.org/doi/pdf/10.1145/299359.299386

Downing, S.M. (2009). Statistics of Testing. In S.M. Downing & R. Yudkowsky (Eds.), *Assessment in health professions education*, (pp. 107-109). New York: Taylor and Francis. https://doi.org/10.4324/9780203880135

Eweda, G., Bukhary, Z. A., & Hamed, O. (2020). Quality assurance of test blueprinting. *Journal of Professional Nursing*, *36*(3), 166-170. https://doi.org/10.1016/j.profnurs.2019.09.001

Fencl, H. S. (2019). Point of view: Focusing on learning as a marker of success for underrepresented students. *Journal of College Science Teaching*, *48*(5), 6-7. https://doi.org/10.1080/0047231X.2019.12290468

Grant, A. (2016, September 10). Why we should stop grading students on a curve. *The New York Times*. https://www.nytimes.com/2016/09/11/opinion/sunday/why-we-should-stop-grading-students-on-a-curve.html

Holmes, J. D. (2021). The bad test-taker identity. *Teaching of Psychology*, *48*(4), 293-299. https://doi.org/10.1177/0098628320979884

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. https://doi.org/10.1080/02602938.2020.1795617

Kibble, J. D. (2017). Best practices in summative assessment. *Advances in Physiology Education*, *41*(1), 110-119. https://doi.org/10.1152/advan.00116.2016

Kulick, G., & Wright, R. (2008). The impact of grading on the curve: A simulation analysis. *International Journal for the Scholarship of Teaching and Learning*, 2(2), n2. https://eric.ed.gov/?id=EJ1146678

Malau-Aduli, B. S., & Zimitat, C. (2012). Peer review improves the quality of MCQ examinations. *Assessment & Evaluation in Higher Education*, 37(8), 919-931. https://doi.org/10.1080/02602938.2011.586991

Maloy, J. T. (1990). A useful grade-scaling equation. *Journal of Chemical Education*, *67*(5), 414. https://pubs.acs.org/doi/pdf/10.1021/ed067p414

Nelson, J.E., Varma-Nelson, P., & Kloempken, T.A. (1992). A graphic grade-scaling method. *Journal of Chemical Education*, *69*(6), 462. https://pubs.acs.org/doi/pdf/10.1021/ed069p462

Noble, T., Suarez, C., Rosebery, A., O'Connor, M. C., Warren, B., & Hudicourt-Barnes, J. (2012). "I never thought of it as freezing": How students answer questions on large-scale science tests and what they know about science. *Journal of Research in Science Teaching*, *49*(6), 778-803. https://doi.org/10.1002/tea.21026

Omar, N., Haris, S. S., Hassan, R., Arshad, H., Rahmat, M., Zainal, N. F. A., & Zulkifli, R. (2012). Automated analysis of exam questions according to Bloom's taxonomy. *Procedia-Social and Behavioral Sciences*, *59*, 297-303. https://doi.org/10.1016/j.sbspro.2012.09.278

Page, R. B., Espinosa, J., Mares, C. A., Del Pilar, J., & Shelton, G. R. (2018). The curvy road to student success in underserved populations. *Journal of College Science Teaching*, *47*(5), 6-7. https://eric.ed.gov/?id=EJ1178262

Rudolph, M. J., Daugherty, K. K., Ray, M. E., Shuford, V. P., Lebovitz, L., & DiVall, M. V. (2019). Best practices related to examination item construction and post-hoc review. *American Journal of Pharmaceutical Education*, *83*(7), 7204. https://doi.org/10.5688/ajpe7204

Stevens, S. P., Palocsay, S. W., & Novoa, L. J. (2022). Practical guidance for writing multiple-choice test questions in introductory analytics courses. *INFORMS Transactions on Education*, *24*(1), 51-69. https://doi.org/10.1287/ited.2022.0274

Tan Yuen Ling, L., Yuen, 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), 7. https://digitalcommons.georgiasouthern.edu/ij-sotl/vol14/iss1/7/

Thelk, A. (2008). Detecting items that function differently for two-and four-year college students. *Research & Practice in Assessment*, *3*, 23-27. https://eric.ed.gov/?id=EJ1062685

Wellberg, S. (2023). Teacher-made tests: Why they matter and a framework for analysing mathematics exams. *Assessment in Education: Principles, Policy & Practice*, *30*(1), 53-75. https://doi.org/10.1080/0969594X.2023.2189565

Welch, A. C., Karpen, S. C., Cross, L. B., & LeBlanc, B. N. (2017). A multidisciplinary assessment of faculty accuracy and reliability with Bloom's Taxonomy. *Research & Practice in Assessment*, *12*, 96-105. https://eric.ed.gov/?id=EJ1168817