Grade Scaling: Issues and Approaches[*]

Keir G. Armstrong

Department of Economics
Carleton University
Ottawa, ON K1S 5B6

karmstro@ccs.carleton.ca

September 5, 2003

[*] This research has benefited from financial support by the Social Science Research Council of Canada.

Abstract

The need for grade scaling typically emerges in the context of an assessment of student work based on relatively objective or fixed subjective criteria that produces a distribution of results that the instructor believes to be problematic in some sense; e.g., a multiple-choice test in which a large enough proportion of the students did so poorly that, left unscaled, it is likely to deter them from putting further effort into the course. Even instructors who believe that grade scaling is pedagogically unsound may, from time to time, be faced with the practical reality that, all things considered, it is a necessary evil. As such, a good understanding of the options that instructors have open to them in this matter would seem to be essential.

In this paper, I discuss issues surrounding the scaling of grades as well as the relative merits of different approaches to doing so. The main issue dealt with concerns the justification for grade scaling on pedagogical grounds. This takes us some distance in establishing a set of axioms that inform the choice of a general approach to grade scaling. Next, I show that, among seven different approaches, including five that are fairly well known and one that is entirely new, only the latter satisfies all of the axioms. Finally, I show that the new approach can be used as a "self-scaling" technique for adjusting course grades to reflect class participation in a manner that is non-detrimental to students who reach a minimum standard and differentially beneficial to students who are closer to the pass-fail boundary (relative to those who are further from it, in either direction) on the basis of the other required elements of the course.

JEL Classification Numbers: A20, C65.

Key Words: grade scaling; assessment of student learning.

1. Issues

At the beginning of my academic career, I sought out the advice of a senior colleague who had close to thirty years of teaching experience and had been recognized, both officially in the form of a teaching achievement award and unofficially in the form of general acclaim, for his effectiveness as a first-year undergraduate instructor in my discipline. On the matter of how to structure the evaluation of student performance, he made what I thought at the time to be a surprising admission: He found it necessary to include an explicit "fudge factor," as he called it, with a weight of ten per cent, that gave him some latitude to adjust each student's final grade as he saw fit. What surprised me about this admission was not so much the fact that my colleague scaled his students' grades—this was precisely what I had been directed to do on several occasions in graduate school as a teaching assistant following an unexpectedly disastrous class performance on a mid-term or final examination—but that he anticipated and explicitly took account of the need to do so, even after teaching the same course twenty-odd times and being recognized as an outstanding instructor. Up until that point, my impression had been that grade scaling was a sort of fail-safe measure, hidden deep within the instructor's bag of tricks, to be invoked, nay spoken of, only with regret. I have since come to realize that my colleague's approach was in fact justifiable on solid pedagogical grounds.

How can this be so? An important reality of today's classroom (in the Western world, at least) is that the students therein have a greater variety of backgrounds and learning styles than ever before. Moreover, we can expect the nature of student diversity to continue to change in unpredictable ways into the foreseeable future. The students themselves are aware of this situation and, more or less as a result, are not as tolerant of monolithic teaching styles as they used to be. Consequently, a sound teaching philosophy must admit a diversity of approaches that are continually evaluated and modified to better cope with the changing diversity of students. An inevitable consequence of such experimentation is that, at least occasionally, most of the students in a class will perform below the instructor's expectations in an assessment of learning, regardless of how carefully it has been designed. When this happens, "it [must be] recognized that the

instructor and students may have had a different understanding regarding the importance of specific content or that the assessment tool was in some way inadequate or confusing to students. [Scaling] the scores reduces or eliminates the penalty to students when the failure to succeed may, in part, have been the fault of the instructor" (Royse, 2001, p. 193).

At the core of the pedagogical justification for grade scaling, then, is fairness. To be fair to the diversity of students in a class, the instructor must experiment to some extent with different ways of teaching the required material. To be fair to the students when subjected to the possibly invalid assessments of learning that may result from such an approach, the instructor must scale the scores when they fall substantially short of his or her expectations.

While not the only remedy for this sort of unfairness, scaling is preferable to the alternatives of re-grading and re-assessment because it is much less costly to implement. Scaling grades can be accomplished in a matter of minutes using a computer with spreadsheet software. By contrast, re-grading can take on the order of hours or even days to complete and requires that the evaluation criteria be made less stringent—something that may be either impossible (e.g., multiple-choice questions) or undesirable (e.g., if specific standards must be maintained for pedagogical reasons). Re-assessment (e.g., giving a replacement test) takes even more time and has additional costs associated with throwing the course off schedule.

2. Axioms

How scaling should be carried out is clearly an important issue in the light of the fairness justification. To start, let's treat an approach to scaling grades as a mathematical formula that converts any number x between zero and one to a possibly different number y between zero and one—zero and one corresponding to the lowest and highest possible scores (zero and one hundred per cent) that a student can receive on an assessment of his or her learning. Assuming that every student has been assessed under equivalent conditions and in accordance with objective or consistent subjective standards, there should be no question about grades reflecting *relative* learning attainments in the context

of a particular class. Furthermore, there should be little doubt that similar grades reflect similar degrees of learning. These two conclusions correspond to asserting that a grade-scaling formula should be *rank-order preserving* and *continuous*.

A1. *Rank-Order Preservation*: For any two students, if the raw score of one exceeds that of the other ($x' > x$) then the scaled scores should be ranked the same way ($y' > y$).

A2. *Continuity*: For any two students, if x and x' are close in value then so should be y and y'.

Since the fairness justification interprets scaling as compensation for possibly invalid assessments of learning, a grade-scaling formula should be *non-detrimental* in the sense of not reducing any student's score.

A3. *Non-Detrimentality*: For any student, $y \geq x$.

Assuming that the assessment tool is relatively broad in scope and that the associated standards of evaluation do not demand perfection, any student who receives a raw score of zero can easily be seen to have earned it. Consequently, a grade-scaling formula should be *non-beneficial to zero scores* or "*zero stationary*."

A4. *Zero Stationarity*: For any student, if $x = 0$ then $y = 0$.

Many critiques of scaling are actually critiques of a particular approach to scaling known as *curving* or *grading on a curve*. In a nutshell, curving means that the raw scores are adjusted so that they fit a pre-determined distribution pattern. For this approach to be valid, it must be assumed that the students in a given class constitute a random sample drawn from a population with a distribution of aptitudes for learning that is consistent with the aforementioned pattern. Except, perhaps, for very large classes, this assumption is unlikely to be realistic. And even if it were realistic, I suspect that most educators would chaff at "[t]he notion that grades, and the learning they supposedly represent, are a limited commodity dispensed by the teacher according to a statistical formula" (Walvoord and Johnson Anderson, 1998, p. 100). Accordingly, a grade-scaling formula should be *flexible* in the sense of being able to reflect adequately different instructors' judgements about what a specific distribution of grades should look like.

A5. *Flexibility*: The grade-scaling formula admits the possibility of effecting independent changes to two or more characteristics of the distribution of scores.

3. Approaches

There are five approaches to scaling that are fairly well known. The simplest of these, the *common increment method*, increases each raw score by the same amount; i.e.,

$$y = x + a \ ,$$

where a is a positive number that is no larger than the difference between one and the highest raw score. This formula satisfies every axiom except A4 (since $y = a > 0$ when $x = 0$) and A5 (since it increases the mean score by a but has no effect on the dispersion of the scores), and is usable only if there are no perfect scores.

The *linear* or *base reduction method* multiplies each raw score by a common factor; i.e.,

$$y = Px \ ,$$

where P is a number between one and one divided by the highest raw score. This formula satisfies every axiom except A5 (since it increases the dispersion along with the mean), and is usable only if there are no perfect scores.

The *affine method*, also known as *Welch's* (1992) *method*, multiplies each raw score by a common factor and then adds one minus that factor to each result; i.e.,

$$y = (1 - P) + Px \ ,$$

where P is a number between zero and one. This method satisfies every axiom except A4 (since $y = 1 - P > 0$ when $x = 0$) and A5 (since it decreases the dispersion while increasing the mean).

The *standard-* or *z-score method* divides the difference between each raw score and the mean raw score by the standard deviation of the raw scores[1] (yielding the associated "z-scores"), and then adds the *desired* mean to each result multiplied by the *desired* standard deviation; i.e.,

$$y = s^* z + \bar{x}^* \ ,$$

where $z = (x - \bar{x}) / s$ is the z-score, $\bar{x}$ is the mean raw score, s is the standard deviation of the raw scores, and $\bar{x}^*$ and s^* are the desired mean and standard deviation, respectively, chosen by the instructor. This formula satisfies every axiom except A3 (unless either $\bar{x} / \bar{x}^* \leq s / s^* \leq 1$ or $(1 - \bar{x}) / (1 - \bar{x}^*) \geq s / s^* > 1$) and A4 (unless $\bar{x} / \bar{x}^* = s / s^*$), and is usable only if all the raw scores lie between $\bar{x} - s\bar{x}^* / s^*$ and $\bar{x} + s(1 - \bar{x}^*) / s^*$.

The curving method discussed above (in the paragraph between the formal statements of A4 and A5) orders the raw scores from highest to lowest and then partitions them into grade categories based on a pre-determined distribution pattern; "for example, 10 percent *A*s, 10 percent *F*s, 20 percent *B*s, 20 percent *D*s, and 40 percent *C*s" (Walhout, 1997, p. 84). Assuming that such grade categories have associated numerical values, curving fails to satisfy every axiom—unless the lowest category has a value of zero, in which case it satisfies A4; or unless the grade categories are numerous enough to form a quasi-continuum, in which case it satisfies A1 and A2.

A sixth approach to scaling, the *power method*, is much less well known than the preceding five. It involves raising each raw score to a common power; i.e.,

$$y = x^b ,$$

where b is a number between zero and one. This formula satisfies every axiom except A5. The power method's lack of flexibility derives from the fact that, depending on the value chosen for b, it gives a particular raw score the largest increment by a particular amount.[2] As b rises from zero towards one, the raw score receiving the largest increase rises and the amount of that increase declines. The graphs of these relationships are shown in Figure 1.

The seventh and final approach to scaling that I consider is entirely new: The *generalized power method* is a flexible variant of the power method defined by

$$y = x + P\left[x^b(2 - x) - x\right] ,$$

where P is a number between zero and $1 / (2 - b)$, and b is a number between zero and two. This formula gives the instructor the options of choosing which raw score gets the largest adjustment (via the b parameter) and by how much (via the P parameter). The

relationship between the b parameter and the target raw score x^* receiving the largest adjustment is described by a non-linear function that I will refer to as $b = \varphi(x^*)$.[3] It turns out that a good approximation to $b = \varphi(x^*)$ in the range of target scores between 26 and 62 per cent is given by the straight-line formula[4]

$$b = 2.5x^* - .25 .$$

Figure 2 illustrates this approximation.

Once x^* has been chosen and the b parameter has been determined, the instructor needs to decide how much he or she wishes to increase the target raw score ($y^* - x^*$). The appropriate value of the P parameter can then be found using the formula

$$P = \frac{y^* - x^*}{(x^*)^b (2 - x^*) - x^*} ,$$

which comes from solving the generalized power formula for P with x and y set equal to x^* and y^*, respectively. As indicated above, the nature of this formula requires that $y^* - x^*$ be such that

$$0 \le P \le \frac{1}{2 - b} .$$

Substituting for P in this expression and rearranging terms yields

$$0 \le y^* - x^* \le \frac{(x^*)^b (2 - x^*) - x^*}{2 - b} .$$

Since $b = \varphi(x^*)$, the upper bound on $y^* - x^*$ (given by the right-hand side of the preceding expression) can be treated as a function of x^* alone. This function, which is shown in Figure 3 to be closely approximated by

$$(x^*)^{\frac{x^*}{1-x^*}} - (x^*)^{\frac{1}{1-x^*}} ,$$

decreases at a decreasing rate from a value slightly less than one at x^* slightly greater than zero to a value slightly greater than zero at x^* slightly less than one. Consequently, the greater the chosen target raw score, the smaller the maximal amount by which it can be raised.

4. Examples

Example 1: Suppose the instructor wishes to target a raw score of 50 per cent and raise it to 59 per cent; i.e., $x^* = .50$ and $y^* = .59$. Since 50 is between 26 and 62, the requisite value of b is given by the straight-line approximation formula as

$$2.5(.50) - .25 = 1 .$$

The requisite value of P is given as

$$(.59 - .50) / \left[(.50)^1 (2 - .50) - .50\right] = .09 / .25 = .36 .$$

Note that this value does not exceed the upper bound of 1 / (2 – 1) = 1. Plugging the parameter values $b = 1$ and $P = .36$ into the generalized power formula yields

$$y = x + .36\left[x^1 (2 - x) - x\right] = 1.36x - .36x^2 .$$

As shown by the upper curve in Figure 4, application of this formula to a collection of raw scores does nothing to any score of zero or 100 per cent, raises any score of 50 per cent by 9 percentage points, and raises scores that are closer to 50 per cent by more than those that are further away.

Provided that $0 < x^* \leq .50$ (which implies that $0 < b \leq 1$), the generalized power method scales raw scores in a manner analogous to that of the preceding example. Furthermore, unlike any of the other considered methods, it satisfies all five axioms discussed above. As x^* is increased beyond .50 and towards 1, however, the method becomes unusable over an growing range of low-end raw scores and violates axiom A3 over a wider range of low-end raw scores. If, for instance, $x^* = .62569$, the upper bound on the domain over which the straight-line approximation formula for b is applicable, then the generalized power method is unusable whenever there are raw scores below 1.21 per cent and reduces any scores that lie between 1.21 and 11 per cent (by tiny amounts not in excess of 1.3 percentage points). Similarly, if $x^* = .7$, the method is unusable whenever there are raw scores below 3.23 per cent and reduces any scores that lie between 3.23 and 29.7 per cent (by at most 5.5 percentage points).

Example 2: Suppose the instructor wishes to target a raw score of 74 per cent and raise it by as much as possible. Reading off the $b = \varphi(x^*)$ curve in Figure 2, $b \approx 1.5$ at $x^* = .74$. Setting $P = 1 / (2 - 1.5) = 2$ will result in the target score being raised by the

maximum possible amount. Plugging the parameter values $b = 1.5$ and $P = 2$ into the generalized power formula yields

$$y = x + 2\left[x^{1.5}(2 - x) - x\right].$$

Application of this formula is limited to collections of raw scores with every element in excess of 6.69 per cent. As shown in Figure 5, it does nothing to any score of 38.2 or 100 per cent, raises any score of 74 per cent by 12.4 percentage points, raises any scores between 38.2 and 100 per cent that are closer to 74 per cent by more than those that are further away, lowers any score of 14.3 per cent by 8.52 percentage points, and lowers any scores between 6.69 and 38.2 per cent that are closer to 14.3 per cent by more than those that are further away.

Implementation of the generalized power method is obviously best accomplished using a computer. For concreteness, let us assume that we have recorded twenty-five raw scores in rows 1 through 25 of column A of a spreadsheet program (e.g., Excel or Quattro). Assuming further that these scores are numbers to be taken out of 20, we need to convert them to percentages before doing anything else. To do this, we enter 20 in cell A26 and +A25/A$26 in cell B25, copy the latter cell, and then paste it to cells B1 through B24.[5] Rows 1 through 25 of column B now contain the raw scores expressed as percentages (the *x*s).

Suppose we decide to target a raw score of 30 per cent and raise it to 45 per cent. To reflect these choices in our spreadsheet, we enter .30 in cell B26 and .45 in cell B27. Since 30 is between 26 and 62, the requisite value of *b* is given by the straight-line approximation formula. Accordingly, we enter +2.5*B26-.25 in cell C26 and see that

$$b = 2.5(.30) - .25 = .5.$$

We then find the requisite value of *P* to be

$$(.45 - .30) / \left[(.30)^{.5}(2 - .30) - .30\right] = .23767$$

by entering +(B27-B26)/(B26^C26*(2-B26)-B26) in cell C27. In order to verify that this value is not too large, we compare it with the upper bound 1 / (2 – .5) = .66667 found by entering +1/(2-C26) in cell C28. Finally, we apply the generalized power formula by entering +B25+C$27*(B25^C$26*(2-B25)-B25) in cell C25, and then copying and

pasting it to cells C1 through C24. Rows 1 through 25 of column C now contain the scaled scores (the *y*s) that result from $x^* = .30$ with $y^* = .45$. The impact of this exercise is shown in Figure 6.

The advantage of having b, P, and the upper bound on P entered as formulas in our spreadsheet is that we can easily re-scale the raw scores until we are satisfied with the results. All we have to do is enter new values for one or both of x^* and y^* in cells B26 and B27, respectively, and the computer will automatically re-calculate the parameter values and the scaled scores.

5. Self-Scaling

The final aspect of grade scaling that I would like to discuss is the use of the generalized power method as the basis for a "self-scaling" mechanism that reflects a student's participation in class. The principal objectives of this approach are to reward such participation in a purely non-detrimental manner on the one hand, and to penalize habitual absence from class on the other. In my experience, the "carrot and stick" that these objectives represent cannot be implemented to satisfactory effect under the usual approach of dedicating some fraction of the final course grade to in-class participation. The problem I have encountered with this approach is that, for those students who are relatively shy or sufficiently disengaged, the expected costs of showing up and trying to avoid attention outweigh the expected benefits of the mediocre marks they are likely to receive in exchange. Furthermore, many such students appear to believe that they can offset a low participation mark through effort expended on other aspects of the course.

I have come to assess the participation of the students in my first-year seminar course on a week-by-week basis with up to one point being awarded for simple attendance and an additional point for making a "basic contribution"—either individually or as part of a team, depending on the requirements of the associated seminar assignments. If earned, the latter point is subject to a "contribution quality adjustment" of –½, –¼, 0, +¼, +½, +¾ or +1 points awarded at my discretion. At any time during the course, a student's "personalized adjustment factor" can be calculated by adding up his or her seminar participation marks, after subtracting one point from each, and then

multiplying the result by .0075, and his or her raw score can be calculated as a weighted average of marks awarded for other required elements. A student's course-grade-in-progress and, ultimately, his or her final grade in the course, is generated by applying the generalized power method to his or her raw score with b set equal to one and P set equal to his or her personalized adjustment factor.

I explicitly warn the class at the outset of the course that, because each of a student's participation marks is reduced by one point in the calculation of his or her personalized adjustment factor, habitual absence from the seminars will render this factor *negative* thereby making his or her final grade *less than* his or her raw score. More encouragingly, I also point out that consistently active seminar participation will result in a positive personalized adjustment factor thereby making the associated final grade greater than the corresponding raw score.

For a given personalized adjustment factor, the mark-up (or mark-down) of the final grade over the raw score under my self-scaling mechanism is larger the closer the raw score is to 50 per cent (since $b = 1$ as in Example 1 above) and zero for raw scores of either zero or 100 per cent. Consequently, the largest impacts of seminar participation on final grades accrue to those students whose raw scores wind up close to the pass-fail boundary. In a 24-week course such as mine, the upper and lower bounds on these impacts are as shown in Figure 4.[6]

The differential impacts of a given personalized adjustment factor over the range of possible raw scores can be seen to be fair because they give the biggest boost when a student needs and, in a sense, deserves it most, while at the same time offering a measure of insurance against a falling grade when his or her needs are less (since the impact rises as the raw score falls towards 50 per cent). And since the size of the impact on a given raw score depends on the frequency and quality of seminar participation, every student has a self-interested motivation to try and make worthwhile contributions and thereby enhance the learning experience of the entire class. Without such an incentive, there is a tendency for many students to free ride on the contributions of others or, worse still, to opt out of the seminars altogether.

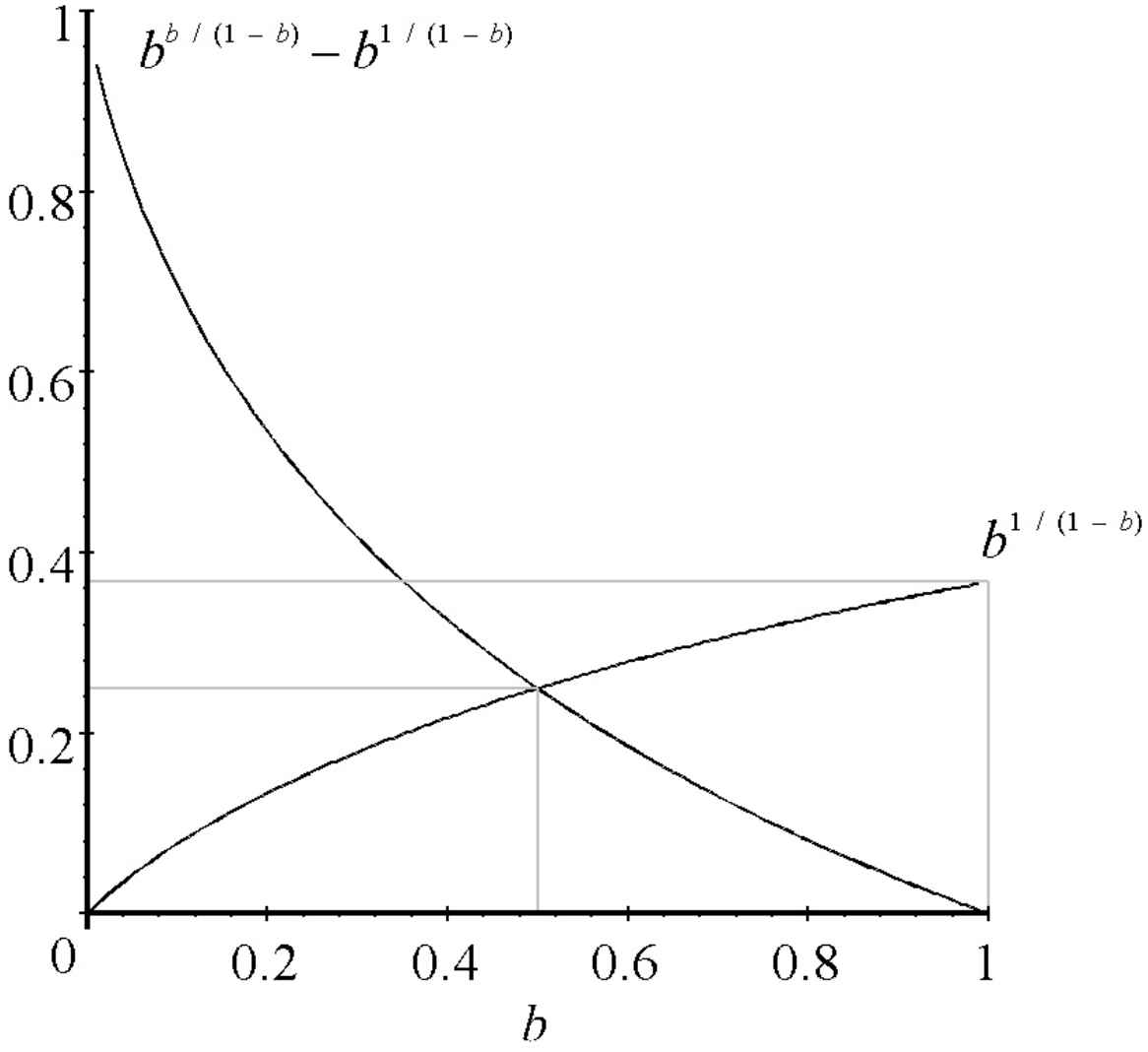

Figure 1. The raw score receiving the largest increase ($b^{\frac{1}{1-b}}$) and the amount of that increase ($b^{\frac{b}{1-b}} - b^{\frac{1}{1-b}}$) under the power method.

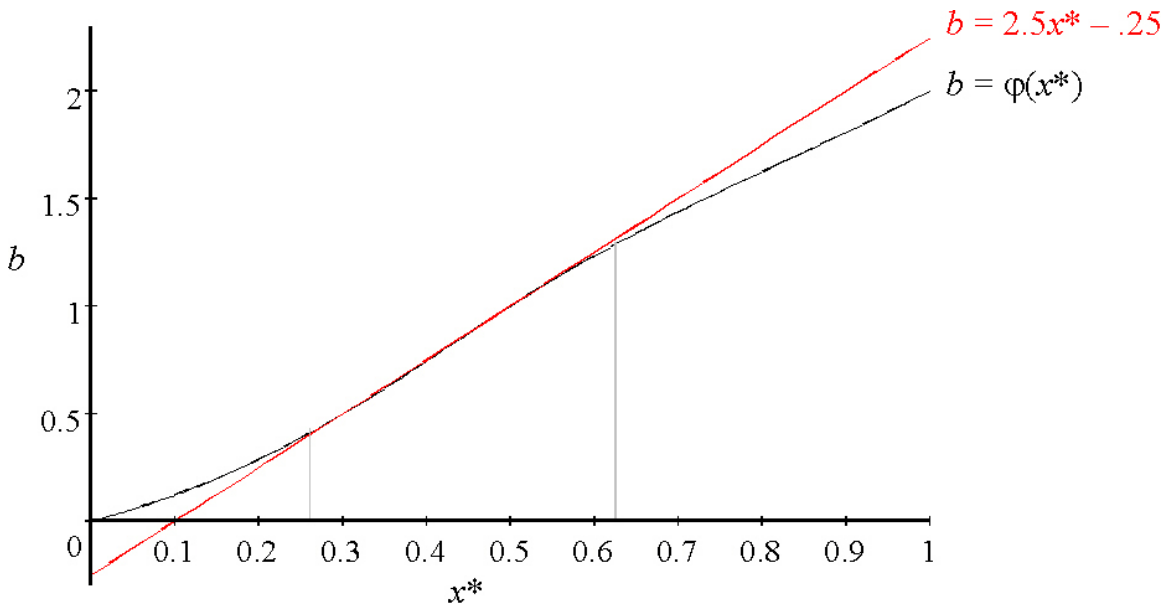

Figure 2. The actual ($b = \varphi(x^*)$) and approximated ($b = 2.5x^* - .25$) relationships between the b parameter and the target raw score x^* under the generalized power method.

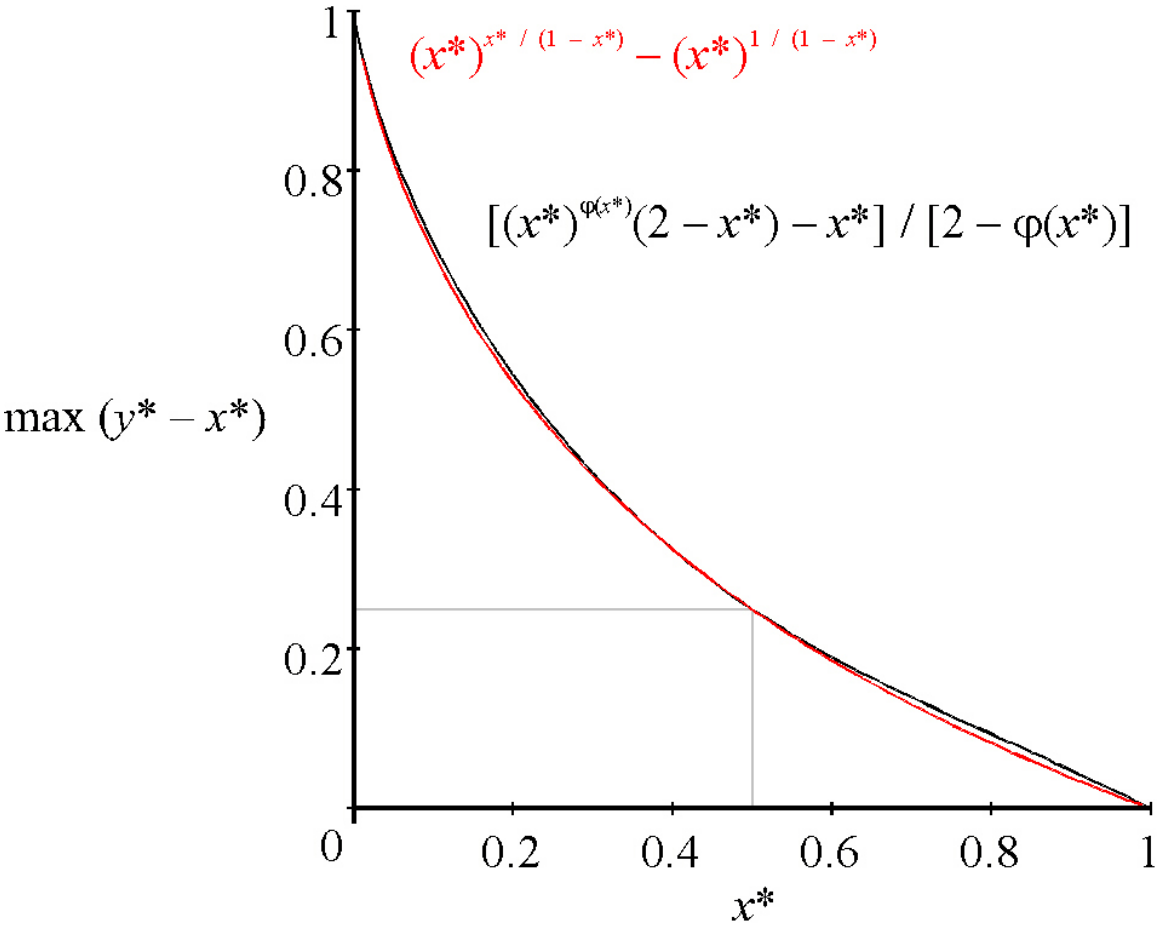

Figure 3. The actual and approximated relationships between the upper bound on $y^* - x^*$ and the target raw score x^* under the generalized power method.

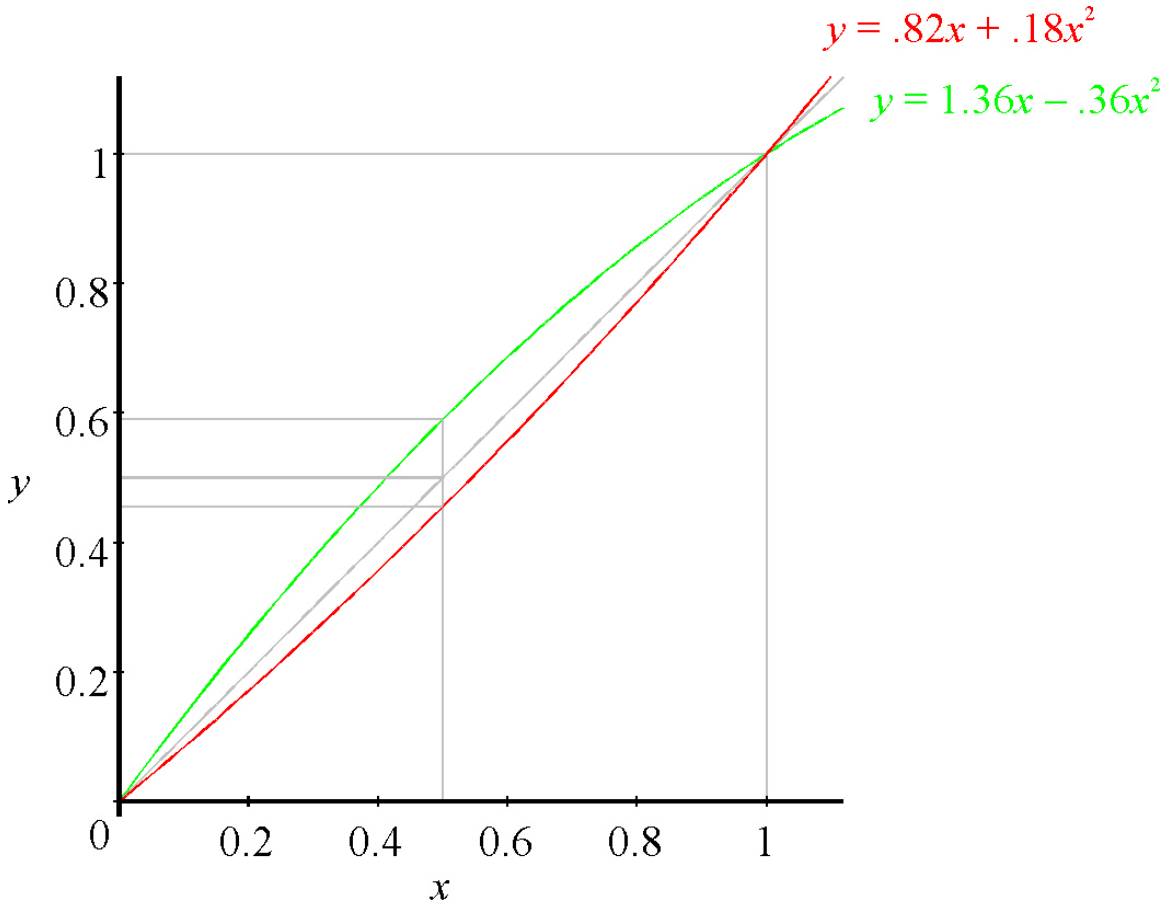

Figure 4. The impact of the generalized power method when $b = 1$ and $P = .36$ ($y = 1.36x - .36x^2$), and when $b = 1$ and $P = -.18$ ($y = .82x - .18x^2$).

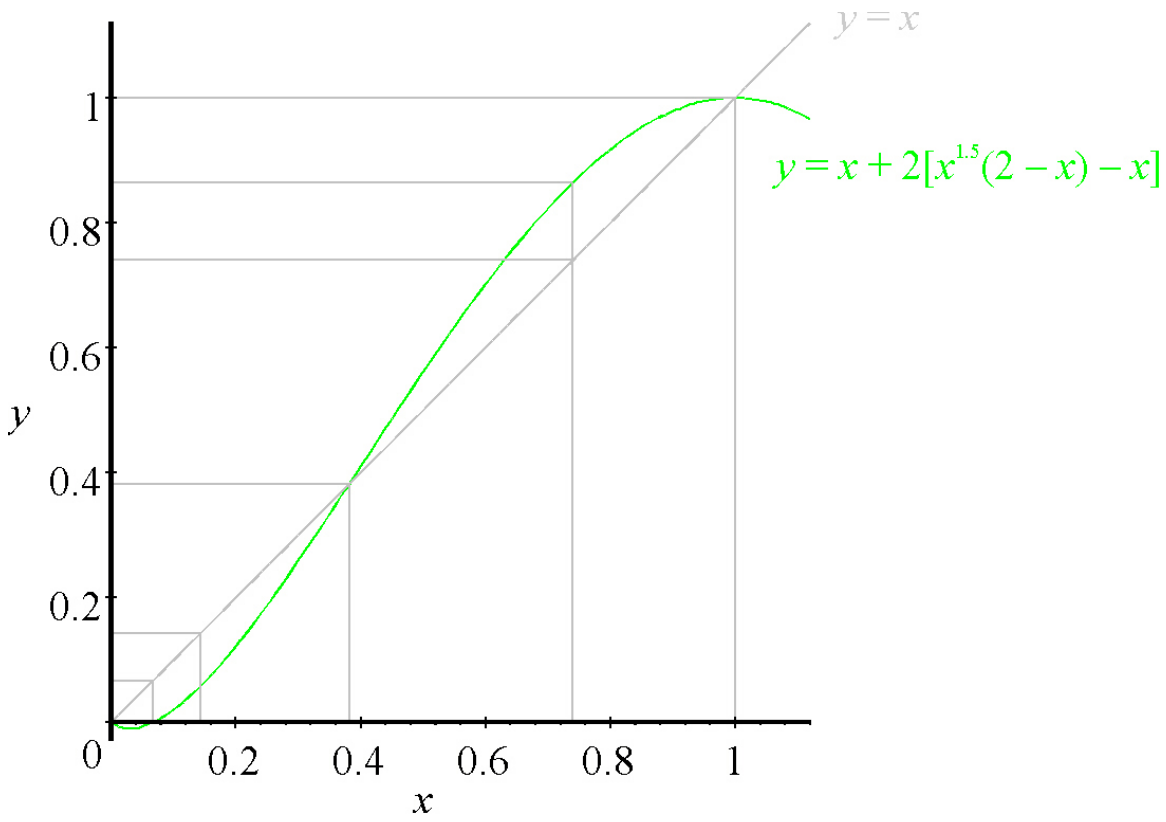

Figure 5. The impact of the generalized power method when $b = 1.5$ and $P = 2$.

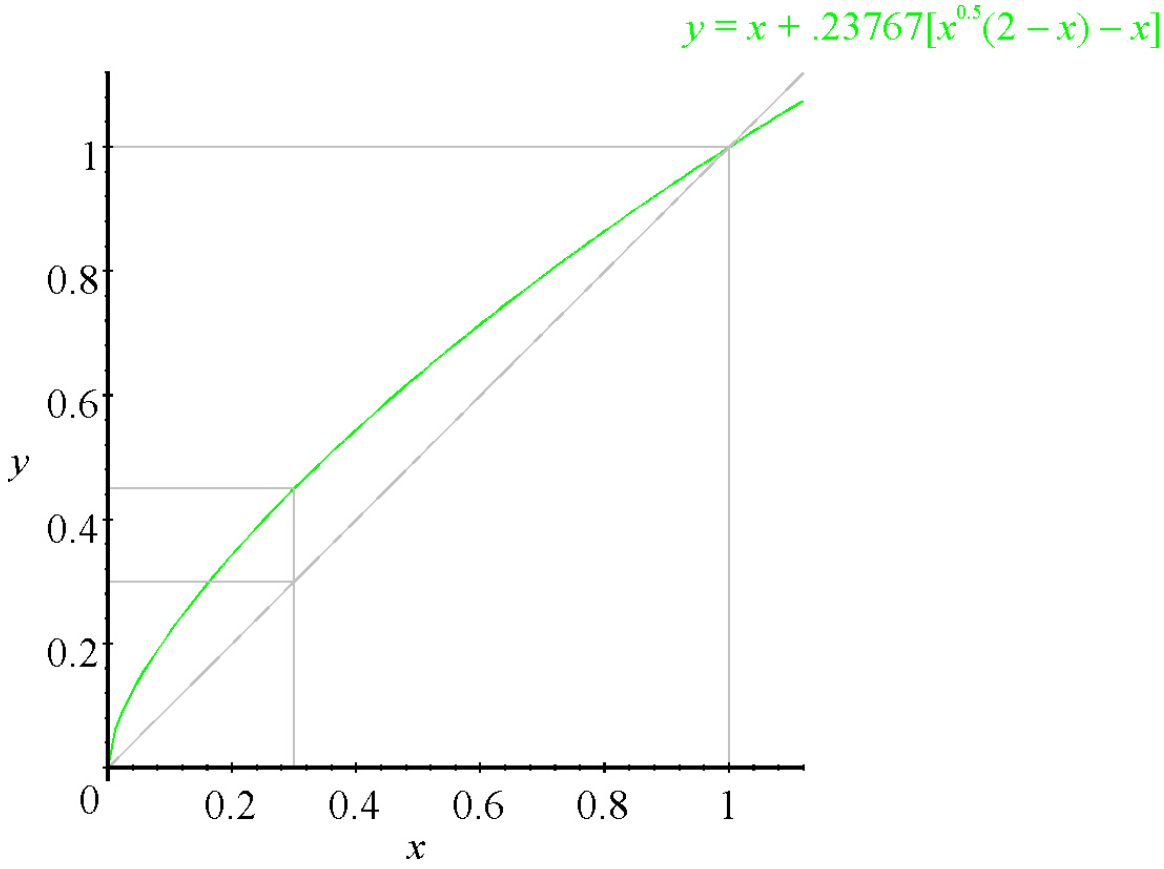

Figure 6. The impact of the generalized power method when $b = .5$ and $P = .23767$.

References

Royse, David. *Teaching Tips for College and University Instructors: A Practical Guide*. Boston: Allyn & Bacon, 2001.

Walhout, Donald. "Grading Across a Career." *College Teaching*, 45, No. 3 (Summer 1997), 83–87.

Walvoord, Barbara E., and Johnson Anderson, Virginia. *Effective Grading: A Tool for Learning and Assessment*. San Francisco: Jossey-Bass, 1998.

Welch, Gary G. Letter. *Mathematics Teacher*, 85, No. 8 (Nov. 1992), 608.

---

[1] Standard deviation is the most commonly used measure of dispersion.

[2] To see this, take the derivative of y with respect to x, set it equal to one, and solve for x.

[3] Taking the derivative of y with respect to x in relation to the generalized power formula, setting it equal to one, and simplifying the result yields

$$2bx^{b-1} - (b+1)x^{b} - 1 = 0 \ .$$

Given b, one of the solutions to this equation is the raw score that gets the largest adjustment under the generalized power method. Taking $x = x^*$ as given instead, the equation yields the value of b that results in the target raw score x^* receiving the largest adjustment; i.e., $b = \varphi(x^*)$.

[4] More rigorously, the absolute log-difference between the value of this formula and $\varphi(x^*)$ at any x^* such that $.26099 \leq x^* \leq .62569$ is less than two per cent; i.e., $|\ln(2.5x^* - .25) - \ln \varphi(x^*)| < .02$ for all $x^* \in [.26099, .62569]$.

[5] Note that the $ sign fixes the reference to row 26 in each pasted formula while the reference to row 25 becomes a reference to the row number of the target cell.

[6] Respectively, the bases for these bounds were calculated as $P = 24(3 - 1)(.0075) = .36$ and $P = 24(0 - 1)(.0075) = -.18$.