Elise M. Stevens *graduated from the University of North Carolina - Chapel Hill School of Media and Journalism in May 2016. She is currently a post-doctoral fellow in the University of Oklahoma's Health Sciences Center, where she studies cognitive science and neuropsychology and how these fields can inform effective message design for improving health outcomes. Her research has been published in* Communication Research. **Rhonda Gibson** *joined the University of North Carolina - Chapel Hill School of Media and Journalism faculty in 2001. Her teaching areas include journalism writing and reporting, media ethics, sexual minorities and the media, and mass communication pedagogy. Her research focuses on the effects of exemplification in journalism on issue perception and the effects of images of sexual minorities in the media. Her research has been published in* Journalism & Mass Communication Quarterly, Communication Research, Newspaper Research Journal, *and* Journalism & Mass Communication Educator, *among other publications, and she is currently working on a book examining the changing communication strategies and public opinion regarding same-sex marriage.*

Kuhlthau, K., Ruscio, J., Luce, C., & Furey, M. (2017). What's in a grade? A professor's guide to adjusting scores on student assessments. *Journal on Excellence in College Teaching, 28*(2), 81-110.

What's in a Grade? A Professor's Guide to Adjusting Scores on Student Assessments

Kaitlin Kuhlthau
John Ruscio
Christine Luce
Matthew Furey
The College of New Jersey

Grades serve many important purposes, and educators agree it is important to assign them accurately and fairly. Interviews with 100 professors across a broad range of academic disciplines at a mid-sized state college revealed little consensus on whether, when, or how to adjust scores when assigning grades. Their responses informed a review of reasons to adjust scores under certain circumstances as well as concerns about doing so. The authors review the characteristics of many score adjustment methods, including both quantitative and qualitative techniques. Professors are encouraged to consider when score adjustments may be warranted and to thoughtfully select methods that help to achieve their grading goals.

A student's course grade is an indicator of rank, quality, proficiency, intensity, or value. Grades serve as a key informational and motivational communication tool among students, professors, and third parties such as future employers, graduate or professional school admissions committees, and peer counselors or advisors. In an academic setting, grades may assess an understanding of course material, measure strengths and weaknesses, reflect effectiveness in presentation and assessment of material, or signify approval or disapproval of the level of performance in a course. In graduate or professional schools or in the workplace, grades may be used to predict work and study habits, knowledge, skills, abilities, effort, and future success in educational pursuits or a career (Rojstaczer & Healy, 2012; Svinicki & McKeachie, 2011).

For grades to be an effective tool, a professor must decide how to accurately assess, evaluate, and interpret student performance. The criteria for earning grades should reflect the purposes and functions they are designed to serve in various contexts. In other words, the relationship between the assessment of student performance and the assignment of grades should be clear. Grades should reflect the quality of students' performance based not only on the expectations and learning objectives set forth by the professor, but also by the ways that interested third parties are likely to use and interpret these grades (Walvoord & Anderson, 1998).

Grades bear past, present, and future implications. They reflect historical competencies, measure student achievement and knowledge, and operate as predictors of future academic and career successes (Davis, 1993; Mavis, 2014). Because grades have considerable consequences for students, it is critical for professors to assign them as accurately and fairly as possible (Piontek, 2008; Walvoord & Anderson, 1998). Accuracy issues include face, construct, and criterion-related validity, with grades reflecting students' comprehension and mastery of course objectives and serving as useful predictors of their future performance. Fairness issues include transparency and objectivity in the grading process as well as awarding grades that reflect students' true performance levels regardless of the difficulty level of an assessment tool.

In this article, we highlight one aspect of the grading process that we believe to be of underappreciated potential for helping to achieve important goals: adjusting scores. Though this is sometimes referred to loosely as "curving" or "scaling" scores, we prefer the more general term "adjusting" because it encompasses a broad range of techniques that are not limited to quantitative methods (for example, linear or nonlinear score transformations). Furthermore, informal terms such as "curving" scores can be misleading. For example, one common score adjustment method involves adding the same number of points to all students' scores. Referring to this particular method as "curving" is inaccurate, because it involves a linear transformation without any kind of curve. Additionally, students may assume that "curving" scores means that grades are forced into a distribution that follows a normal, or bell-shaped, curve, which is not true for most methods of adjusting scores. Thus, we believe it is important to communicate clearly the nature and justification for any score adjustment method, and precise language can be helpful in this regard.

Score adjustment is a fairly common practice in higher education (Kulick & Wright, 2008). For example, observed scores on an assessment might diverge significantly from what a professor expected, perhaps because he or she inadvertently constructed an assessment instrument with too low or too high of a difficulty level. In such a case, the professor may decide to adjust scores before assigning grades. Because few professors have had much explicit training in grading in general, let alone in score adjustment methods, personal experiences or trial-and-error often contribute to subjective judgments about what "feels right" when evaluating student performance (Strashny, 2003). Grading standards and expectations likely differ across and within colleges, schools, departments, and even multiple sections of the same classes, and these variations potentially influence professors' grading practices. Deciding whether and how to adjust scores raises many complex questions: What are the advantages and disadvantages to adjusting scores? What methods are available to do so? What are the pros and cons of each, and how do these relate to the grading goals and context at hand? Given the importance and complexity of the task, but with little or no explicit training, it is not surprising that many professors find it challenging, perhaps daunting, to grapple with these questions.

In our discussion of score adjustment, we focus our attention in two ways. First, we limit ourselves to the context of norm-referenced assessment. Criterion-referenced assessments do occur in higher education, but they are uncommon in the classroom (Imrie, Cox, & Miller, 2014). More often, they are constructed and administered by third parties who target specific knowledge, skills, and abilities required for certification or licensure in a profession. Our emphasis on assessments pertains largely or entirely to rank ordering students and assigning grades that reflect this relative standing. Second, we discuss making score adjustments on particular assignments rather than on assigning final course grades in order to streamline the overall discussion, with the understanding that much of what we review would generalize to score adjustments in the final grading process.

We begin by presenting the results of a self-constructed survey designed to examine professors' beliefs and practices about score adjustment. Informed by these data, we address some of the most salient concerns and suggest that in many—but by no means all—instances, a thoughtfully chosen score adjustment method can be worthwhile. In the remainder of the article we overview of a number of methods that professors can use to adjust scores. We review the characteristic features of each method that potentially make it more (or less) appropriate to achieve certain goals in the grading process. We illustrate the use of each method and offer a user-friendly Excel file with a series of worksheets that can be used to implement the score adjustment methods reviewed here (see http://ruscio.pages.tcnj.edu/quantitative-methods-program-code/). Our primary aim is to encourage professors to consider whether score adjustments may be helpful in certain circumstances and, if so, to help them make a strategic choice among the many available methods.

Survey of College Professors About Score Adjustment

To ensure that our own beliefs about the grading process and the role of score adjustment within that process were as broadly informed as possible, we interviewed 100 full-time tenured or tenure-track faculty teaching undergraduate classes at a mid-size state college. Participants represented all seven schools at the college (Arts and Communication; Business; Education; Engineering; Humanities and Social Sciences; Nursing, Health, and Exercise Science; and Science) as well as all academic departments within each school. We sampled widely to ensure that we would hear from individuals with different educational backgrounds, teaching experiences, disciplinary traditions, and so forth. Each professor who agreed to participate took part in a semi-structured interview that lasted approximately 10-15 minutes. The questions designed to guide each interview are in Appendix A. Participants' responses were documented in real time as fully as possible and later transcribed. The data were examined for trends or patterns to help understand the professors' familiarity with and use of various methods of score adjustment when assigning grades. Table 1 summarizes the findings.

Overall, a strong majority of participants stated that they had adjusted scores on at least one assessment. When asked about their decision to adjust scores, most of the participants who did so indicated that this had not been their original plan; rather, this decision was based on unexpected results. A majority of the participants adjusted scores on specific assessments rather than at the end of the semester; however, some participants considered adjustments under both circumstances. Scores were more often adjusted on multiple-choice or written exams than on labs, projects, performances, or papers. Most participants used the same approach for adjusting scores in all of their classes, regardless of whether the courses were more versus less advanced or offered primarily for majors versus the college core curriculum.

When asked about their familiarity with score adjustment methods,

Table 1
Summary of Survey Results

School	*HSS*	*SCI*	*BUS*	*EDU*	*ARTS*	*ENG*	*HES*	*Total*
Full sample	**N = 28**	**N = 28**	**N = 14**	**N = 14**	**N = 9**	**N = 5**	**N = 2**	**N = 100**
Adjust scores	89	86	93	79	67	60	50	83
Knowledge of method(s) to adjust	24	71	57	57	22	60	0	47
Why adjust scores?								
Motivation	57	29	64	43	22	40	0	43
Error/bad item	32	50	36	50	11	40	0	38
Borderline grades	46	11	0	0	22	0	50	29
Fairness	50	0	29	21	22	60	0	25
Average too low	18	29	21	21	22	40	0	23
Difficulty	0	36	36	7	22	40	0	20
Accuracy	11	7	0	29	0	20	0	10
Better grades?	0	0	21	0	22	0	50	6
Why not adjust scores?								
Maintain expectations	11	39	14	36	89	40	50	32
Arbitrary	0	14	14	57	56	0	0	19
Unfair	0	21	29	29	0	20	0	15
Inaccurate	0	18	21	14	33	0	0	13
Resources provided	0	7	0	43	33	0	0	11
Unnecessary	4	0	7	0	0	40	50	4
Average class size	18-25	13-48	20-30	20-24	18-25	20-30	20-25	

Table 1 (*continued*)
Summary of Survey Results

School	*HSS*	*SCI*	*BUS*	*EDU*	*ARTS*	*ENG*	*HES*	*Total*
Subsample	***n* = 25**	***n* = 24**	***n* = 13**	***n* = 11**	***n* = 6**	***n* = 3**	***n* = 1**	***n* = 83**
Assessments adjusted								
Exams	36	96	92	82	50	67	100	82
Papers	44	4	23	55	50	0	0	25
Projects	0	17	31	64	33	0	0	20
Quizzes	4	17	15	27	17	33	0	14
Homework	0	21	8	0	17	0	0	8
Adjust after assignments	56	54	77	73	50	67	0	60
Adjust at end of semester	24	25	8	9	17	33	100	21
Adjust at both times	20	21	15	18	33	0	0	19
Adjustments for unexpected results	72	83	92	100	100	100	100	86
Adjustments planned	44	17	8	0	0	0	0	19
Adjust for all classes	92	75	92	73	33	100	100	81
Adjust for some classes	8	25	8	27	67	0	0	19
Preferred score adjustment methods								
Remove poor item	36	13	31	82	0	0	0	30
Boost borderline grade*	52	21	15	18	17	0	100	29
Add points	20	29	54	18	17	33	0	28
Adjust curve/range	0	33	31	0	50	67	0	20
Rewrite/redo	36	8	23	18	0	0	0	19
Extra credit	16	0	15	0	17	0	100	10
Drop lowest grade	16	4	0	0	0	0	0	6
Weighted final	0	4	8	0	0	33	0	4

Notes. All cells but class size contain percentages, which may sum to more or less than 100% due to multiple or lack of participant responses in each category. School abbreviations: HSS = Humanities and Social Sciences; SCI = Science; BUS =Business; EDU = Education; ARTS = Arts and Communication; ENG = Engineering; HES = Health and Exercise Science. *Adjustment method does not apply to a specific assessment.

about one-half of all participants identified at least one method. The range of methods provided was rather limited, with the most common being the following four: adding the same number of points to all students' scores (the example that had been provided by the interviewer), adjusting the curve or range of scores in some way, allowing rewrites or re-dos of an assignment, and boosting borderline grades.

Participants often expressed fairly strong opinions about the desirability of adjusting scores. The most common reason offered in support involved providing motivation to students who might otherwise be discouraged by low grades. Other common justifications for making adjustments involved ensuring fairness in grading; taking into account the difficulty of the assessment; achieving a good class average or norm; avoiding problems associated with retaining a poorly functioning item; and giving students the benefit of the doubt and the encouragement to succeed, especially when dealing with borderline grades. On the other side of the argument, many participants provided reasons not to adjust scores. These included the notions that students should "get the grade they deserve" based on how well they met established expectations for the class; that score adjustments can be arbitrary, unfair, or confusing to students; and that multiple resources already had been made available to help students succeed (for example, office hours, online resources, tutoring center). Despite these strong opinions, there was no general agreement on whether, or when, scores should be adjusted.

Had participants' opinions about adjusting or not adjusting scores reflected context-specific concerns, it would have been neither surprising nor alarming to observe their substantial disagreements. However, much of the support for or objection to adjusting scores was expressed in general terms, suggesting that the lack of consensus does not stem from situational factors that arguably should inform important decisions about grading practices. Consistent with this interpretation, many participants revealed that they had never discussed this topic before the interview. Because this finding arose spontaneously rather than in response to a question asked of everyone, it is unclear how many participants had ever discussed score adjustment.

Concerns About Adjusting Scores

The survey results described above helped to identify four especially important concerns that many professors hold about adjusting scores.

Reasons for Adjusting Scores

First, the decision to adjust scores often results from a discrepancy between expected and observed scores. Because none of our survey participants mentioned or endorsed any adjustment methods that could lower any scores, we believe that this first issue nearly always involves the perception that students performed worse than expected. There are several possible explanations for such a gap, and they are not mutually exclusive (Livingston, 1988). Broadly speaking, the causes most often involve the students, the assessment, and / or the professor. The students may be of unusually low ability, or they may have prepared inadequately for or worked insufficiently hard on an assignment; the assessment may have been more difficult than intended, or it may have measured performance poorly (for example, badly written items, disconnect between intended and actual content coverage); or the professor may have taught the material ineffectively. Unless the professor feels confident that poor performance stemmed from low student ability or effort, and not from factors related to instruction and assessment, some adjustment of the scores when assigning grades may be warranted.

Fairness in Grading

A second concern involves fairness in grading. Like the discrepancy between expected and observed scores, fairness emerged as one of the most common themes when survey participants discussed reasons to—or not to—adjust scores. Considerations of fairness likely felt clear and reasonable to the individuals who advanced them, but there was very poor agreement on how this concept should be defined, let alone methods for best achieving it. Who is to judge what is fair to whom, and by what criteria? Does fairness imply equal treatment, consideration of individual effort, recognition of achievement that should be rewarded, or ensuring of accountability for mistakes? Is it fair for students in sections of a course taught by a more demanding professor to receive lower grades than those in other sections? Is it fair for a professor to reward some students' hard work and determination by boosting borderline grades, or should the professor solely respect actual student achievement? No simple, objective standard involving fairness exists to determine when to adjust scores, and therefore, it is reasonable that different professors will reach different conclusions.

Before proceeding, however, we revisit one fairness-related rationale for not adjusting scores that emerged frequently in our interviews: "Students should get the grades they deserve." The fairness and objectivity intended by this approach depends critically on the assumption that there are no plausible alternative explanations for student performance other than student ability and effort, including the dubious belief that professors can and do always construct assessment instruments that achieve precisely the appropriate and intended levels of difficulty. We do not wish to be overly critical on this point, but we encourage professors to be open-minded in considering whether their assessments and instruction are sufficiently close to flawless to ensure the fairness of unadjusted scores serving as final grades. Perhaps if professors reflected on their own experiences as students, they might recognize that instructors do not always teach as effectively as they imagine, and that assessments at times turn out to be more difficult than intended. Indeed, anyone who has performed item analyses to improve an assessment instrument knows that even subtle changes in the phrasing of items or response options potentially have significant influences on responses (Nunnally, 1978). In other words, scores on an assessment can be adversely affected when students honestly and innocently misunderstand what the professor had in mind when crafting the assignment. Equating grades with raw scores presumes a degree of objectivity that may seldom, if ever, be attainable in practice. As a result, students may not "get the grades they deserve" without some appropriate score adjustment.

Potential for Student Misunderstanding

A third concern includes the idea that adjusting scores may seem arbitrary or confusing to students. This possibility suggests the importance of providing students with an explanation for why and how scores are adjusted. If scores are adjusted upward, students are unlikely to be upset. But if adjustments will affect some students' grades more than others', a sound rationale for justification should be provided, because some students may perceive the procedure as unfair. Describing the reasons for making an adjustment as well as the method to do so helps prevent misunderstandings or misperceptions. It remains possible that some students will not find a professor's explanation satisfactory, but no grading policy is likely to meet the approval of all students. This resonates as true even when the grading policy is perceived by students as straightforward and objective, involving no score adjustments: An implicit or explicit refusal to adjust scores even when an assessment turns out to be more difficult than intended can seem quite arbitrary and harsh. As with all course policies, professors have the responsibility to explain their choices so that students know what to expect. We recommend taking the opportunity to present a rationale for grading policies (Frisbie, 2004), including any circumstances under which scores will be adjusted as well as why and how the adjustment will be done.

Possible Effects on Student Motivation

A fourth concern about adjusting scores involves the possible effect on students' motivation. Low grades despite high effort sometimes discourage students. To the extent that a score adjustment potentially alleviates this problem, it seems worth considering. However, easily attained high grades at times reduce future effort, and a score adjustment might contribute to this problem. Balancing these risks is not necessarily an easy task, and once again, we suggest only that adjusting scores may be a helpful approach under certain circumstances.

An Underappreciated Rationale for Adjusting Scores

Before turning to an overview of the score adjustment methods themselves, we address something that was conspicuously absent in our survey of professors: the psychometrics of grades. Those professors who perform empirical research keenly acknowledge the importance of avoiding artificial or unnecessary restrictions in the range of observed data, such as those due to floor or ceiling effects in measurement (Bordens & Abbott, 2011). A truncated range of scores constrains variability, with many undesirable consequences. For example, reduced variability makes it more difficult to detect or interpret correlations between variables or differences across experimental conditions. Likewise, reduced variability attenuates measurement reliability and validity. All else being equal, these psychometric properties will be enhanced to the extent that measurements span a broad range of values (Nunnally, 1978).

Whether they perform empirical research or not, few professors appear to appreciate that the same principles apply when assessing their students and assigning grades. Knowledge, skills, and abilities can be assessed using items or instruments of varying difficulty levels. When professors are unwilling to allow a large proportion of students to fail and do not intend to adjust scores, they tend to construct assessments such that scores are restricted to the upper end of the possible range. For example, a professor might write an exam expecting that most students will score at least 60% correct, with an average of 80% correct or higher. This amounts to a planned restriction of range, and artificially constraining the variability of scores in this way consequently compromises measurement reliability and validity.

A psychometrically superior alternative incorporates constructing an assessment instrument such that scores span a broader range, thereby preserving measurement reliability and validity. The assessment must be considerably more challenging than is typically the case, with expected scores much lower than 60% for many students and an average score well below 80%. The resulting broader range of scores may be unacceptably low to assign as grades, in which case scores can be adjusted upward. Rather than presuming that the assessment as constructed will yield an appropriate range of grades, planning to adjust scores encourages a thoughtful consideration of merited grades. Examining students' performance affords an opportunity to reconsider factors such as student ability and effort, content coverage, difficulty level of the assessment, and instructional effectiveness, all of which play into assigning grades.

To underscore the value of constructing more challenging assessments rather than striving for the restricted range of high scores that subsequently equate with student grades, we performed simulations to examine the relationship between the difficulty level of an assessment and the measurement reliability of the resulting scores. In each of 1,000 simulated classes, 200 students with widely varying and normally distributed ability levels responded to three versions of a 25-item test that varied in item difficulty levels (easy, moderate, and high). Naturally, mean test scores differed across the easy (85% correct), moderate (72% correct), and high (54% correct) item difficulty levels. Measurement reliability was estimated by calculating the internal consistency using Cronbach's alpha (α; Cronbach, 1951) of each version of the test within each simulated class of students. Across all simulated classes, the mean values of α included .67, .73, and .76 for the easy, moderate, and high item difficulty levels, respectively. The correlation between the 1,000 classes × 3 versions = 3,000 tests' mean scores and α levels was -.82. Thus, measurement reliability decreased substantially and systematically with less challenging assessments.

This simulation demonstrates that constructing an assessment instrument designed to yield scores sufficiently high to assign as grades needlessly sacrifices a nontrivial amount of measurement reliability. Given that reliability constrains validity (Nunnally, 1978), this is an important psychometric consequence. Though we simulated fairly large classes of students with normally distributed ability levels taking the tests, the basic trend that we observed—that more challenging assessments improve measurement reliability— by no means solely pertains to these conditions. Though it would vary in magnitude, the direction of this effect would be expected to hold with classes of any size, with student ability distributions of any shape, and with other kinds of assessments besides tests. This fundamental point is well known to those who perform empirical research, and it applies with full force to professors in their role as instructors. In addition to whatever other reasons one might consider when deciding whether to adjust scores when assigning grades, sound psychometric reasons exist that support beginning with more challenging assessments than are the norm and then adjusting scores in a thoughtful manner.

Methods for Adjusting Scores

If it is determined that scores need to be adjusted, the method for doing so needs to be selected carefully: Is the goal of adjusting scores to reach a particular class average or a desired grade distribution? Should all scores be adjusted by the same amount, or might it be preferable to increase lower and higher scores by different amounts? Is it acceptable or desirable for all students to receive passing grades? Is it alright if some adjusted grades exceed 100%?[1] Are there any "curve breaking" scores—statistical outliers at the upper end of the distribution—that might pose special challenges to making any adjustments? There are many ways to adjust scores when assigning grades, each with distinct advantages and disadvantages with respect to specific grading goals and contexts (Richeson, 2008).

In this section, we review a broad array of score adjustment methods drawn from a variety of sources (for instance, Kulick & Wright, 2008; Richeson, 2008; Svinicki & McKeachie, 2011), including personal experience. To highlight some broad conceptual differences, we break these down into linear methods (that is, transformations accomplished by a straight-line transformation of scores into grades), nonlinear methods (that is, transformations accomplished via functions other than straight lines), and qualitative methods (that is, techniques that do not involve quantitative transformation of scores into grades). A summary of the characteristic features of each method, along with its relationship to various grading goals, is provided in Table 2. Figure 1 and Figure 2 display the effects of applying different adjustment methods to the same data. The illustrative data consist of one realistic set of raw scores that a professor might want to consider adjusting. This sample contains N = 200 scores in a negatively skewed distribution, meaning that many scores bunch together at the upper end, with scores spread further apart toward the lower end. The scores range from 33 to 95 (M = 75.03; SD = 11.11; Mdn = 77; IQR = 69 to 83). Applying grading thresholds of 90% to earn an A, 80% to earn a B, and so forth to these illustrative data yields 6% A's, 33% B's, 34% C's, 18% D's, and 10% F's.

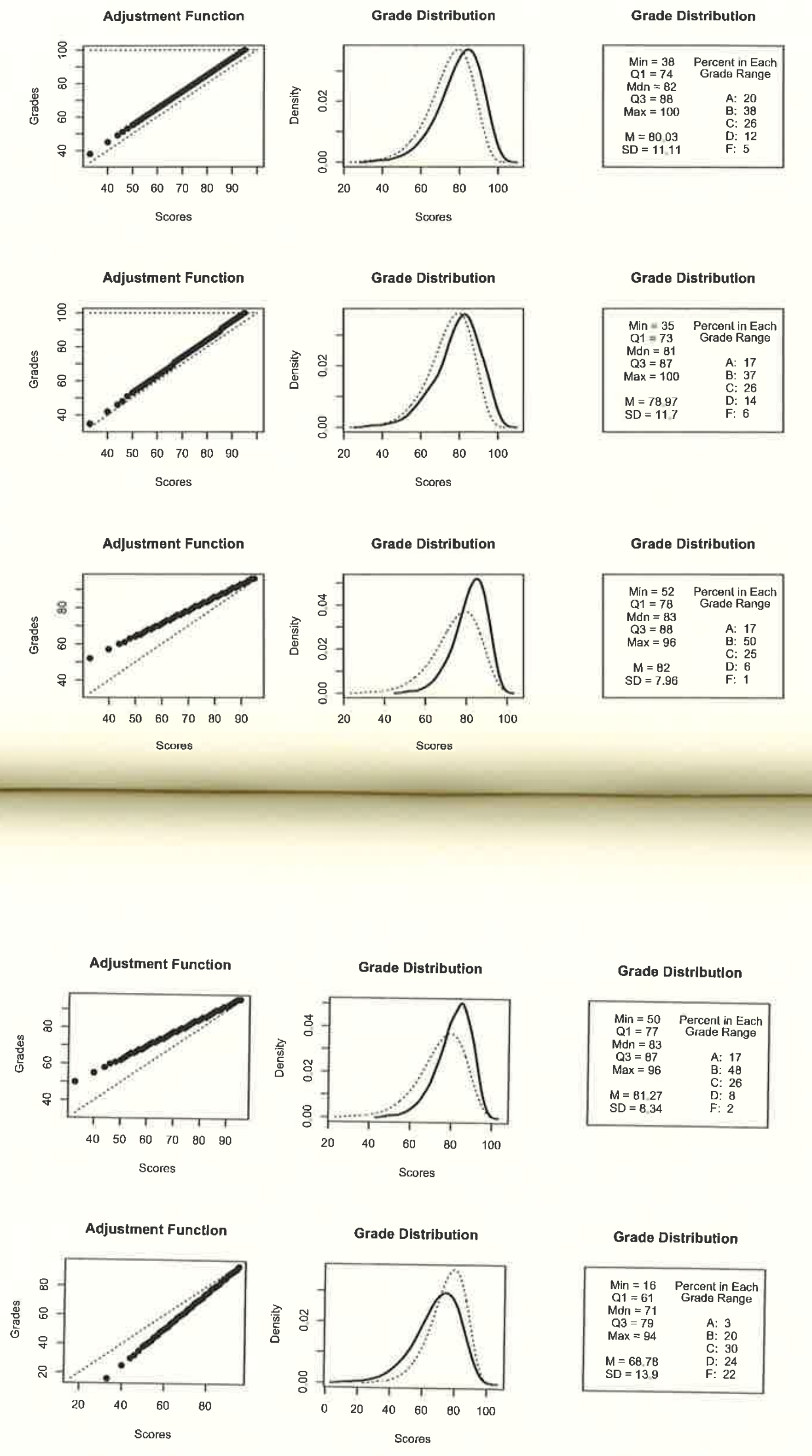
Table 2

Grading Goals and Characteristic Features of Score Adjustment Methods

	Score Adjustment Method											
Grading Goal or Characteristic Feature	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Simple to use and easy to explain	X	X	X		X			X		X	X	?
Compensate for difficult assessment and/or poor item		X	X	X	X		X	X	X	X	X	?
No grade can exceed 100%	X	?	?		X	X	X	X		X	?	X
Highest score does not constrain adjustment options	X		?		X	X	X	X		X	?	X
Will not alter shape of distribution	X	X	X	X	X	X						
Same increase for all scores		X										
Lower scores increase more than higher scores					X		X					
Higher scores increase more than lower scores			X									
Variable effect on scores (some may even decrease)				X		X		X	X	X	X	X
Attain desired average and spread of grades				X					X			?
Remove influence of guessing						X						
Additional effort to promote mastery of material										X	?	
Additional effort to redeem poor performance										X	X	
Allow fixed percentage of students to pass												X
May detract from initial effort		?	?		?		?				X	?

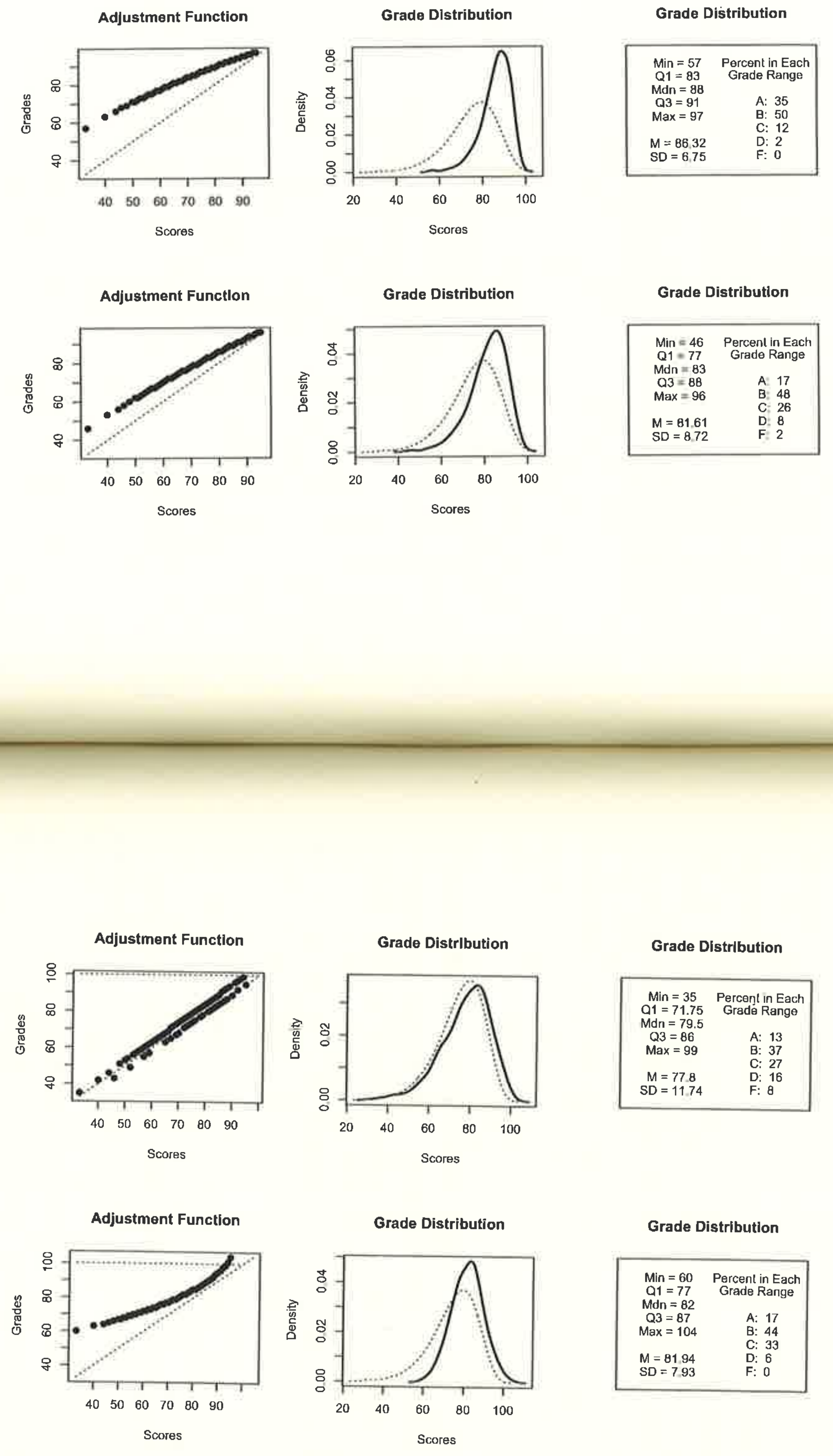
Notes. An "X" indicates that a method is certain or highly likely to meet the stated goal or feature the characteristic indicated. A "?" indicates that a method may meet the goal or feature the characteristic, depending on how it is implemented or other contextual factors. Score adjustment methods are numbered as follows: 1 = no adjustment; 2 = add points; 3 = percent of highest score; 4 = setting *M* and *SD*; 5 = give back points; 6 = control for guessing; 7 = root of score; 8 = remove poor item; 9 = percentile transformation; 10 = return, rewrite, regrade; 11 = extra credit; 12 = thresholds.

Figure 1
Linear Score Adjustment Methods



Note. Dotted lines show the grades if scores are unadjusted (and, in some cases, the reference level of a grade of 100). First panel = adding points (C = 5 points). Second panel = percent of highest score. Third panel = setting mean and standard deviation (desired M = 82, SD = 8). Fourth panel = giving back points (25%). Fifth panel = control for guessing (20% chance-level guessing parameter used).

Figure 2
Nonlinear Score Adjustment Methods



Note. Dotted lines show the grades if scores are unadjusted (and, in some cases, the reference level of a grade of 100). First panel = root of score (root = .5). Second panel = root of score (root = .7). Third panel = remove poor item (an item worth 5% of total points, equivalent to 1 out of 20 multiple-choice items; 22% had gotten the item correct [lower line], 78% had gotten the item incorrect [upper line]). Fourth panel = percentile transformation (desired *M* = 82; *SD* = 8).

Linear Methods

A linear method of adjusting scores will not affect the shape of the original distribution, meaning that the relative spacing between scores will remain the same when they are transformed into grades. In addition, the rank ordering of scores will remain the same. Instead, a linear adjustment method changes the center of the distribution, which will usually shifts upward, and the spread of the distribution, which either increases or decreases.

No Adjustment

In this method, there is no transformation, making this the easiest and simplest method for a professor to use and explain to students. Our survey results also suggest this method as the norm, or the default option. Experience suggests that ordinarily, no explanation for the decision not to adjust scores is provided to (or requested by) students, and we include this method merely as a point of reference. Formally, the raw score x becomes the grade: $f(x) = x$

Add Points

With this method, one adds the same number of points to all scores. It may be best used to compensate for the difficulty level of an assessment. This adjustment method is easy to use and explain to students: Everyone's score increases by the same amount to yield a grade. Formally, $f(x) = x + C$, where C is the number of points to add. For example, one might choose to add 5 points to raise the average grade in the illustrative data from a 75 to an 80 (see Figure 1, first panel).

Because some grades potentially exceed 100 if C is sufficiently large, this method presents a dilemma. Should grades above 100 be allowed? Some professors may be comfortable with this, while others may not. Should grades be capped at 100? This may create resentment among students who now earn the same or similar grades as others who performed more poorly. To avoid this dilemma, one can set the highest grade equal to 100 and increase all scores by the same amount (that is, if the highest score of 93 becomes a grade of 100, all other scores would also receive an additional $C = 7$ points). This approach provides the advantage of guaranteeing that no grades exceed 100, but the disadvantage exists that a single unusually high score—a statistical outlier at the upper end—constrains the ability to adjust other scores. Moreover, if the student with the highest score is identifiable, it may create some resentment among the rest of the class.

Percent of Highest Score

Another way to compensate for the difficulty level of an assessment involves calculating each grade as the percentage of the highest score. This guarantees that no grade will exceed 100. Unlike the previous method, this one increases higher scores by more than lower scores. For example, if the highest score is 90, that score would increase by 10 points to a grade of 100. A score of 60, in contrast, would increase by only 7 points, to a grade of 67. Formally, $f(x) = 100x / H$, where H is the highest score. Throughout most the range of scores in the illustrative data, the effect of this score adjustment is fairly similar to adding points (see Figure 1, second panel).

A potential drawback to this method is that it a single high score can "spoil the curve" for everyone else. This may be avoided by calculating grades as the percentage of a value other than the highest score. For example, if this highest score is an outlier of 98, with no other scores above 90, one might substitute a value of 90 into the formula for the true value of $H = 98$. Doing so might be a reasonable way to adjust scores throughout the class, although it would create the dilemma of whether to credit a grade of 109 for the highest score ($100 \times 98/90 = 108.89$) or to cap this single outlying grade at 100.

Setting *M* and *SD*

This method replaces the observed center (M) and spread (SD) of scores with a desired M and SD in the distribution of grades. This affords great versatility for professors to attain a certain desired class average or spread in grades, although the process of determining the desired M and SD potentially requires considerable trial and error to find values that yield acceptable results.[2] In the common case of an unintentionally difficult assessment, one chooses values such that the average increases (for example, to compensate for the difficulty level by increasing scores) and the variability decreases (for example, to prevent a large proportion of grades from exceeding 100). Formally, $f(x) = (x - M_0) \times (SD_1 / SD_0) + M_1$, where M_0 and SD_0 are the observed mean and standard deviation of the scores, and M_1 and SD_1 are the desired mean and standard deviation of the grades. For example, the illustrative data contains a low average ($M_0 = 75.03$) and a fairly high variability ($SD_0 = 11.11$). Using this method, values reset to a higher average (for example, $M_1 = 82$) with less variability (for example, $SD_1 = 8$), and no grades fall above 100 (see Figure 1, third panel).

It is important to understand that standardizing a distribution—setting its central tendency and variability to desired levels—does not alter the shape of a distribution. While this succession resonates with all linear score adjustment methods, a common misconception persists that simply redefining the M and SD of a distribution produces a normal or bell-shaped curve. This is not the case. If the score distribution skews and/or contains outliers, a standardized grade distribution follows the same pattern. Alternatively, using the percentile method, a nonlinear adjustment described below, better approximates a normal or bell-shaped distribution if that is desired.

Give Back Points

With this method, grades are assigned by giving back a certain percentage of points. This method is easy to use, simple to explain, likely to be well received by students, does not constrain the highest scores on the assessment, and ensures that no grades will exceed 100. At the same time, it achieves results highly similar to the more complex method of setting a desired M and SD. A demanding search for a pair of values that yields acceptable results—the desired M and SD—is replaced with a simpler search for a suitable percentage of points to give back. Formally, $f(x) = x + (100 - x) \times (P / 100)$, where P is the percentage of points to give back. For example, giving back 25% of points for a score of 80 yields a grade of 85, the original score of 80 plus an additional 5 points (25% of the remaining 20 points). Giving back 25% of points for all scores in the illustrative data yields results highly similar to those achieved by setting the desired $M = 82$ and $SD = 8$ (see Figure 1, fourth panel).

Control for Guessing

When students' random responding yields scores above 0, it may be desirable to adjust for the influence of chance-level guessing. Although this method can be justified on the grounds that it adjusts for the influence of guessing, which arguably should not affect grades, it results in the overall reduction of scores, which may be perceived as harsh or unfair by students and professors. In our survey, no respondent mentioned or endorsed any score adjustment method that would reduce any students' scores. Perhaps the only circumstance under which one desires to control for guessing would be in a low-level course used to "weed out" students from a highly competitive major or program, in which case one more likely justifies removing the influence of guessing and lowering scores.[3] Formally, $f(x) = 100 \times (x - G)/(100 - G)$, where G represents the expected score for chance-level guessing (that is, identifying the correct option on a multiple choice item by guessing at random). For example, if multiple-choice items have 5 response options, the guessing parameter G would equal 20. Readers familiar with the well-known formula for Cohen's Kappa statistic should recognize its similarity to this function. For the illustrative data, adjusting scores in this way decrease the average and increase the variability (see Figure 1, fifth panel).

Nonlinear Methods

Whereas linear score adjustment methods preserve the relative spacing of scores as they are transformed into grades, nonlinear methods do not. Scores within some region of the distribution may be pushed further apart or pulled closer together than scores in other regions of the distribution. One of the three nonlinear methods described below even at times alters the rank ordering of scores.

Root of Score

To compensate for an overly difficult assessment, one can increase scores in a way that does not depend on the highest score and ensures that no grades will exceed 100. As with the case in several of the linear methods, lower scores tend to increase more than higher scores. The simplest version of the root-of-score method involves the familiar square root. Formally, $f(x) = 10x^{1/2}$. Scores of 0 and 100 are not affected, but the closer a score is to 25, the more it increases. Because scores at or below 25 are uncommon in classroom assessment, practically speaking, lower scores will usually increase more than higher scores. This remains the case for the illustrative data (see Figure 2, first panel). A more general version of this method allows the root to vary: $f(x) = 100^{1-a} \times x^{a}$, where a is chosen such that $0 < a < 1$. The smaller the value of a, the larger the adjustment to scores becomes. Using $a = .7$ achieves a more modest adjustment in the illustrative data (see Figure 2, second panel) than previously, using the square root function (in which case $a = .5$). Unless one uses a rather extreme value for a, the results will be similar to those for giving back a percentage of points (see above; Figure 1, fourth panel). Both of these methods prohibit grades above 100 and increase lower scores more than higher scores, but giving back a percentage of points is simpler to implement and easier to explain.

Remove Poor Item

Whereas several of the methods described above indirectly compensate for a poor item by increasing everyone's score, the most direct method to accomplish this—and only this—goal involves recomputing all scores without the poor item. As a consequence, the remaining items each exert a slightly greater weight in determining grades. What makes this method nonlinear is that no single line captures the adjustment. For example, dropping an item worth 5% of each student's score differentially affects those who had and had not answered that item correctly (see Figure 2, third panel). The 22% of the students who answered correctly receive grades that fall along one line in the first graph, whereas the other 78% of grades fall along another, higher-scoring line.

Because some students' scores increase (those who answered the poor item incorrectly and who no longer lose credit for that response) and others' scores decrease (those who answered the poor item correctly and no longer receive credit for that response), this adjustment possibly creates some resentment among students who answered the poor item correctly and now feel penalized. To avoid this problem, one could either add points to all students' scores rather than removing the poor item (for example, adding 5 points would not penalize anyone who answered the poor item incorrectly, but would award extra credit to those who answered it correctly) or score the item as correct for everyone (for example, add 5 points only to the scores of students who had answered the poor item incorrectly). Reasonable people likely disagree as to which of these three options is the "fairest" solution.

Percentile Transformation

This method converts scores that may be distributed in a very skewed manner into a normal or bell-shaped grade distribution with a desired *M* and *SD*. This is accomplished through a three-step process. The first step is to convert scores into percentiles. For example, within the illustrative data, a score of 80 falls at the 61st percentile. Second, percentiles are converted into the corresponding *z* scores in the standard normal distribution. The 61st percentile corresponds to a *z* score of 0.28. Third, the *z* scores are converted to a distribution that achieves the desired *M* and *SD*. For example, if the desired grade distribution has $M = 82$ and $SD = 8$, a *z* score of 0.28 yields a grade of 84. Though the three-step process may seem off-putting, this method achieves a simple and intuitive result: Any score distribution is transformed into a normal distribution with the desired center and spread. The percentile transformation is nonlinear in that it tends to make larger adjustments to more extreme scores, thinning or thickening the tails of the distribution as needed to achieve a normal, bell-shaped curve. For the illustrative data, one could achieve a normal distribution with a desired $M = 82$ and $SD = 8$ (see Figure 2, fourth panel).

Two common misconceptions regarding normal curves, however, need to be addressed. First, as noted earlier, simply setting the desired *M* and *SD* for a distribution does not transform it into a normal curve. Whereas setting the *M* and *SD* (see above; Figure 1, third panel) left the grade distribution negatively skewed, the percentile transformation achieves the same desired *M* and *SD* in a normal distribution. Thus, those professors who want their grade distribution to fit a normal curve should use the percentile transformation. Second, adjusting scores to follow a normal curve does not constrain the average or spread in grades. In other words, the average grade can be high, moderate, or low depending on the desired *M*. Likewise, the desired *SD* affects the spread in grades. A common misconception persists that a bell-shaped curve must be centered at an average grade (for example, a C), with equal numbers of B and D grades and equal numbers of A and F grades. A normal distribution centers on any grade one prefers, not necessarily a C, and grades vary as much or as little as one wishes.

Even following the percentile transformation, the grade distribution approximates a normal curve only as well as the original scores allow. Percentile transformation yields a better approximation to a normal curve with a larger set of scores, few of which are tied with one another. The same applies for any other adjustment method one might design to normalize scores.

Qualitative Methods

The remaining three score adjustment methods distinguish themselves by the absence of one feature shared by the linear or nonlinear methods described in the preceding sections: The qualitative methods described below do not involve quantitative transformations of scores into grades.

Return, Rewrite, Regrade

Students may learn from their mistakes by redoing written assessments such as papers, projects, or problem-solving sets. This potentially helps students achieve mastery of the course material and motivates them to redeem a poor performance. After regrading work that has been returned and rewritten, a professor can either average the initial grade and the new grade, weight one grade more heavily than the other, or use only the new grade. Unlike the quantitative score adjustment methods described above, this method demands considerable extra effort from students and the professor, who need to complete and grade the new work, respectively. Because returning, rewriting, and regrading leaves everyone with less time for other activities, this method should be undertaken only when the expected benefits outweigh the costs.

Extra Credit

In this method, extra points are awarded to students based on completion of an optional, add-on assignment or participation in an activity pertinent to course goals. For present purposes, we consider only extra credit work that affects the grade on a particular assignment, not separate assignments that contribute only to the final course grade. Like the previous method, offering opportunities for extra credit may motivate students to redeem a poor performance. On the other hand, knowing the availability of extra credit could detract from their initial effort. In addition, it is important to treat extra credit strictly as a bonus, rather than simply raising expectations for all students in light of the opportunity to earn the additional credit, as the latter approach may create resentment among students who, perhaps justifiably, perceive the extra credit assignments as mandatory.

Thresholds

Applying a predetermined series of thresholds transforms a set of scores into letter grades. For example, thresholds can be located by (a) using standard deviation units to split scores into letter-grade categories (a practice often apparently based on the dubious assumption that score distributions tend to approximate normality), (b) using percentiles to assign fixed percentages to each letter-grade category, (c) using naturally occurring gaps in the score distribution to differentiate letter-grade categories, or (d) using fixed numerical ranges to define letter-grade categories (for example, 93-100 = A; 90-92 = A-; 87-89 = B+; and so on). The first two approaches yield predictable grade distributions, which can be useful to attain a desired class average, to comply with administrative policies that require particular grade distributions, or to administer an assessment that only a fixed percentage of students are allowed to pass. Such an adjustment, however, may be perceived as focusing on class-level rather than individual-level performance, which potentially deflates student motivation.

The other two approaches will not necessarily yield predictable results. The third approach, searching for gaps, contains the advantage of ensuring that differences in grades correspond to substantial—rather than minor—differences in scores, but the disadvantage exists that gaps may not emerge, especially in small classes, and may be due to little more than chance (that is, "sampling error" in the language of statistics). The fourth approach, defining a conversion between numerical scores and letter grades, contains the advantage of being simpler to implement and explain, but the disadvantage is that some students who score close together (for example, 89 vs. 90) will receive different grades (B+ vs. A-), even though other students who score further apart (for example, 93 vs. 100) will receive identical grades (A for both).

Closing Remarks

Grades are an extremely important tool for students, professors, and third parties such as employers and the admissions committees of graduate and professional schools. Provided that professors can be motivated to consider their options and to make informed choices, an increased understanding of score adjustment methods likely improves the accuracy and fairness of students' grades, which have many important connotations and consequences. Our survey results helped us to identify reasons that lead some professors to adjust scores under certain circumstances as well as a number of concerns about making adjustments. We also learned that many professors had never discussed this subject with anyone before the interview. Considerable interest consistently arose in learning more about score adjustments, and to that end, we hope the possibilities discussed in this article prove helpful. For those who wish to experiment with or implement methods described here, we prepared a Microsoft Excel file that contains separate worksheets for each of the linear and nonlinear adjustment methods reviewed (see http://ruscio.pages.tcnj.edu/quantitative-methods-program-code/). Each worksheet allows the user to copy and paste (or enter manually) a set of scores as input, specify relevant parameters to make the adjustment, and observe the effects on grades graphically and through summary statistics.

Footnotes

[1] Throughout this article, we presume that both scores and grades can range from 0 to 100. This is done purely for ease of exposition. Any of the score adjustment methods described here could be adapted for use with other ranges of values. Also, we work exclusively with scores and grades that are rounded to the nearest whole number. This, too, is done for simplicity; one can allow fractional values if desired in practice.

[2] Because each grade is rounded to the nearest whole number, there can be a slight discrepancy between the desired values and the observed values of the *M* and *SD*.

[3] We are neither endorsing nor critiquing the use of "weed-out" courses. Rather, we merely suggest that if a course is designed to serve this pur-

pose, one might find this score adjustment both defensible and useful. Readers interested in a more complex, though potentially instructive, way to control for guessing should consult Bickel (2010).

References

Bickel, J. E. (2010). Scoring rules and decision analysis education. *Decision Analysis, 7,* 346-357.

Bordens, K. S., & Abbott, B. B. (2011). *Research design and methods: A process approach* (8th ed.). New York, NY: McGraw-Hill.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297-334.

Davis, B. G. (1993). *Tools for teaching.* San Francisco, CA: Jossey-Bass.

Frisbie, D. A. (2004). Issues in formulating course grade policies. *North American Colleges and Teachers of Agriculture, 21,* 15-18.

Imrie, B. W., Cox, K., & Miller, A. H. (2014). *Student assessment in higher education: A handbook for assessing performance.* New York, NY: Routledge.

Kulick, G., & Wright, R. (2008). The impact of grading on a curve: A simulation analysis. *International Journal for the Scholarship of Teaching and Learning, 2,* 1-17.

Livingston, S. A. (1988). *Adjusting scores on examinations offering a choice of essay questions* (Research Report No. 88-64). Princeton, NJ: Educational Testing Service.

Mavis, B. (2014). Assessing student performance. In W. B. Jeffries & K. N. Hugget (Eds.), *An introduction to medical teaching* (pp. 209-241). Dordrecht, The Netherlands: Springer.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw-Hill.

Piontek, M.E. (2008). Best practices for designing and grading exams. *CRLT Occasional Papers, 24,* 1-12.

Richeson, D. (2008). How to curve an exam and assign grades. Retrieved from http://divisbyzero.com/2008/12/22/how-to-curve-an-exam-and-assign-grades/

Rojstaczer, S., & Healy, C. (2012). Where A is ordinary: The evolution of American college and university grading, 1940-2009. *Teacher's College Record, 114,* 1-23.

Strashny, A. (2003). A method for assigning letter grades: Multi-curve grading. *Econometrics 0305001,* Econ WPA. Retrieved from http://econwpa.wustl.edu:80/eps/em/papers/0305/0305001.pdf

Svinicki, M., & McKeachie, W. J. (2011). *Teaching tips: Strategies, research, and theories for college and university teachers* (13th ed.). Belmont, CA: Wadsworth.

Walvoord, B. E., & Anderson, V. J. (1998). *Effective grading: A tool for learning and assessment.* San Francisco, CA: Jossey-Bass.

Kaitlin Kuhlthau *currently works as a Research & Insights Manager to better understand consumer perceptions and behavior in the ever-changing digital and non-digital landscapes. With an M.Sc. in Environmental Psychology and a B.A. in Psychology and Spanish, Kaitlin's range of research interests include environmental stressors, behavioral economics, design and user experience, and cultural implications.* **John Ruscio** *is a Professor in the Department of Psychology at The College of New Jersey. His research and teaching interests include theory and application of behavioral economics, measures of scholarly impact, modern and robust statistical methods, and the taxometric method for distinguishing categories from dimensions.* **Christine Luce** *is a Senior Research Assistant in the Research and Development division at Educational Testing Service (ETS) in Princeton, NJ. Christine supports research work in the English Language Learning and Assessment group and the Cognitive Sciences group, focusing on projects related to the development of assessments for the Global English language learning population.* **Matthew Furey** *is currently a Senior Analyst in the Organizational Analytics Department at Johnson & Johnson. His focus is in Talent Management, specifically Talent Acquisition, Global Mobility, Leadership & Learning, High Potential Employee Development Programs, Diversity & Inclusion Programs, and Succession Planning. Matthew received a bachelor's degree in Psychology and Business Management from the College of New Jersey. He is currently pursuing an M.B.A. at Rutgers University with a concentration in Analytics and Information Management.*

Appendix A
Questions to Guide Interviews With College Faculty

1. Have you ever/do you transform scores when grading? [if "no," skip to #2]
   A. Under what circumstances do you transform scores?
   B. Do you transform scores after each assignment or at the end of the semester after final averages are computed (i.e., dropping lowest quiz grade)? [if only at end of the semester, skip to #1C]
      B1. For what type of assignments do you consider transforming scores (i.e., essay exams, multiple-choice exams, presentations, quizzes, etc.)?
   C. Do you plan transformations of scores in advance, or is this based on unexpected results (e.g., very low or very high scores)?
   D. Do you do this in all your classes? If not, which ones (i.e., lecture/seminar/lab, major/liberal learning/elective, intro/intermediate/advanced)?
   E. What is typically the average size of these classes (how many students)?
2. Why do/don't you transform scores?
3. What (other) methods can be used to transform scores? [if none, skip #3A]
   A. What do you think are the pros and cons of each method? Under what circumstances or for what purpose might someone use each method?

Bernauer, J. A., & Fuller, R. G. (2017). Beyond measurement-driven instruction: Achieving deep learning based on constructivist learning theory, integrated assessment, and a flipped classroom approach. *Journal on Excellence in College Teaching, 28*(2), 111-132.

Beyond Measurement-Driven Instruction: Achieving Deep Learning Based on Constructivist Learning Theory, Integrated Assessment, and a Flipped Classroom Approach

James A. Bernauer
Richard G. Fuller
Robert Morris University

The authors focus on the critical role of assessment within a flipped classroom environment where instruction is based on constructivist learning theory and where desired student outcomes are at the higher levels of Bloom's Taxonomy. While assessment is typically thought of in terms of providing summative measures of performance or achievement, it can also serve as a powerful tool for guiding both teaching and learning if it is artfully incorporated into the teaching/learning process; the authors refer to this as "integrated assessment." This integrated assessment/constructivist learning approach is described within the context of a flipped classroom environment and is consistent with student-centered teaching, less reliance on direct instruction, and the attainment of higher-level or deep learning.

Introduction

Almost 20 years ago, one of the authors wrote an article in which he described the rationale for using assessment as the starting point rather than the end point of instruction (Bernauer, 1998). Based on their past 20 years of teaching experience, both authors now believe that this concept, while still useful, actually does not go far enough. In this article, we offer a