Educational and Psychological Measurement

http://epm.sagepub.com

A Modification to Angoff and Bookmarking Cut Scores to Account for the Imperfect Reliability of Test Scores

Robert G. MacCann

Educational and Psychological Measurement 2008; 68; 197 originally published online Sep 6, 2007;
DOI: 10.1177/0013164407305584

Educational and Psychological Measurement
Volume 68 Number 2
April 2008 197-214
© 2008 Sage Publications
10.1177/0013164407305584
http://epm.sagepub.com
hosted at
http://online.sagepub.com

A Modification to Angoff and Bookmarking Cut Scores to Account for the Imperfect Reliability of Test Scores

Robert G. MacCann
NSW Board of Studies, Sydney, Australia

It is shown that the Angoff and bookmarking cut scores are examples of true score equating that in the real world must be applied to observed scores. In the context of defining minimal competency, the percentage "failed" by such methods is a function of the length of the measuring instrument. It is argued that this length is largely arbitrary, being heavily influenced by practical educational constraints. Hence, there is an ambiguity or nonuniqueness about the percentage failed. An argument is advanced that the failure rate should reflect the percentage of *true* scores below the cut score. A modification to the cut score is derived that achieves this outcome and simultaneously removes the nonuniqueness in the percentage failed.

Keywords: *minimum competency; standard setting; cut score; Angoff method; bookmarking; score reliability; true score*

Standard-setting methods are being widely used throughout the world to specify levels of student achievement in educational programs. One of the most popular methods is the Angoff (1971) procedure. This method has been well studied and has been extensively compared to other standard-setting methods (e.g., Berk, 1996; Brennan & Lockwood, 1980; Busch & Jaeger, 1990; Chang, 1999; Cross, Impara, Frary, & Jaeger, 1984; Giraud, Impara, & Buckendahl, 2000; Goodwin, 1999; Hambleton, 2001; Harasym, 1981; Jaeger, 1993; Livingston & Zieky, 1989; MacCann & Stanley, 2004). Another newer method that is gaining in popularity is the bookmark method (Beretvas, 2004; Mitzel, Lewis, Patz, & Green, 2001; Wang, 2003), which does not require the item by item judgment characteristic of Angoff.

Both methods provide a simple and convenient set of procedures for identifying levels of minimal competency in various courses. However, in spite of this apparent simplicity, it will be shown in this article that there is an ambiguity about the

Author's Note: Please address correspondence to Dr. Robert G. MacCann, Head of Measurement & Research Services, NSW Board of Studies, GPO Box 5300, Sydney, NSW 2001, Australia; e-mail: maccann.1@optusnet.com.au.

percentage of examinees identified as being below the level of minimal competence. For both methods, it will be shown that this percentage varies as a function of the length of a test, and given that the length of a test often depends on arbitrary factors, then the associated percentage in the "failure" band becomes ambiguous. This characteristic is not a feature of all standard-setting methods.

This article will show that both the Angoff and bookmark cut scores are effectively examples of *true score equating* (Levine, 1955), which explains the dependence of the percentage failed on the reliability of the scores. An interpretation of standard setting in terms of true scores is given, providing the rationale for removing the nonuniqueness of the failure rate by applying a simple modification to the cut score.

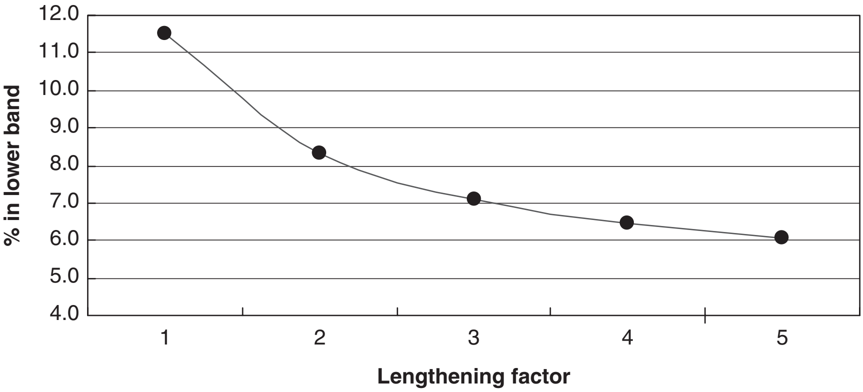
A Paradox

Although both the Angoff and bookmark methods have this same characteristic, the problem will be illustrated using the Angoff method. Consider a hypothetically perfect Angoff standard setting for establishing the cut score defining minimal competency. Although the Angoff judging is hypothetically perfect, the test contains errors of measurement and is of relatively low reliability. The judges, however, have been completely consistent and the cut score, C, cuts off 11.5% of the candidates in the bottom achievement band—those students who have failed to reach an acceptable level of performance. The judges are then shown examples of scripts just below the borderline and confirm that these candidates are appropriately placed in the bottom achievement band. This would normally conclude proceedings.

However, suppose that the students sat for a different test that was doubled in length by adding another section exactly parallel to the first test, the increase in length making the new test more reliable. (For a discussion on parallel measures, see MacCann, 2004.) As the new section is exactly parallel to the old, the judges (being completely consistent) obtain cut scores of C for the first part and C for the second part, obtaining an overall cut score of $2C$. The judges, however, are somewhat surprised when they are given the results of their decisions: The percentage of students in the bottom band has decreased. Only 8.3% have now been selected.

Continuing this scenario, suppose the test length is tripled, resulting in a cut score of $3C$. The judges now become alarmed that the percentage in the bottom band has dropped to only 7.1%. They are puzzled—surely increasing the test length is a desirable procedure, as it increases the reliability of measurement. Yet the borderline candidates now being identified are of lower ability than the judges intended. They wonder what would happen to the percentages if the process were continued. This question may be partially answered (for this particular example) by the graph in Figure 1, which shows the test being lengthened up to five times its original length. After five lengthenings, the bottom-band percentage is now down to about 6%.

Figure 1
Percentage in the Lower Band as a Result of Lengthening a Test



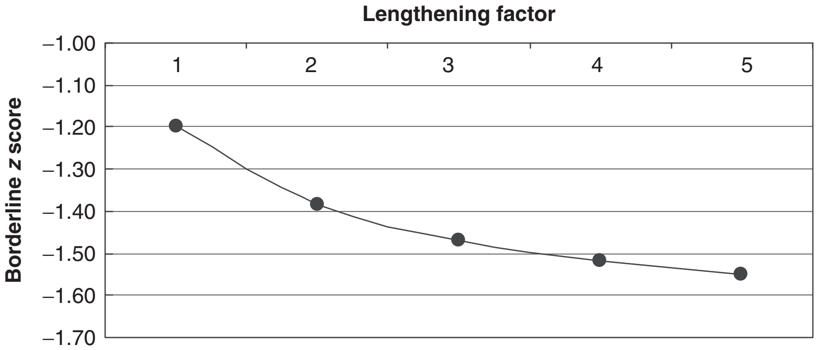
In addition to examining the fall in percentages, the corresponding standard scores (z scores) of the borderline candidates may be observed. These are plotted in Figure 2, showing how the initial borderline z score of -1.2 has steadily fallen to -1.55 by the fivefold lengthening. The equations and parameters for producing the above results will be given in a later section.

An important implication of this scenario results from the fact that the length of a test is largely arbitrary. Whether a test is 1 hour, 2 hours, or 3 hours depends heavily on practical educational constraints such as the time available (given the scheduling of other events), the costs of examining, the capacity of the students to sustain concentration in a single sitting, and so on. For example, if costs and scheduling factors were not relevant, one could test students in 2-hour sessions (one per day) for 5 consecutive days to obtain a test that is five times as long as the original 2-hour test.

Thus, the test length could in principle be any one of the lengths shown in Figure 1. It then follows that the percentage in the bottom band could also be any one of the corresponding percentages in Figure 1; that is, the percentage in the bottom band is arbitrary. This feature is not a consequence of any flaws or breakdown in the application of the Angoff method. It occurs under a theoretically perfect application of the Angoff method, where the judges are completely consistent across all test questions. Thus, if one is asked to estimate the percentage of students who should fail in Mathematics, one should logically reply, “Please specify the test length (reliability) you require, before I can answer.”

This state of affairs leads to a paradox that may be expressed as follows. If the judges apply the Angoff method to their complete satisfaction and (through

Figure 2
Decrease in Borderline *z* Scores as a Result of Lengthening a Test



inspection of sample answers) confirm that the appropriate percentage has been obtained, why does increasing the precision of measurement move the percentage selected further and further away from their original result, in the direction of reducing the failure rate?

Resolution of the Paradox

An important notion in resolving this paradox is that of an individual's *true score*. In the situation described here, this may be conceived as the expected value or long-run average score that a student would obtain over a series of testings (Lord & Novick, 1968). If a student's observed score on a particular occasion differs from the true score, then the difference is termed the error of measurement. Note that in the original test given in the examples above, there is a considerable degree of error of measurement in the test scores.

Now consider the bottom-band students with observed scores below *C*. Some will have true scores below this value and others will have true scores on or above this value. The latter students would have negative errors of measurement (on this occasion) that reduced their observed scores sufficiently to place them in the bottom band. As the test is lengthened and the score reliability increases, the magnitude of the errors of measurement decreases relative to that of the true scores, and a greater proportion of bottom-band students actually have true scores within this band. In the limit, when the test length effectively becomes infinite (and the reliability is perfect), the only students remaining in the bottom band are those whose true scores are in this band.

Hence, a sensible interpretation of the purpose of the standard setting would be that it seeks to determine the percentage of students who have true scores that are below *C*. From this point of view, with relatively unreliable test scores, the Angoff estimate gives too many students in the bottom band (students there because of negative errors of measurement). As the reliability increases, this percentage decreases, dropping students who should not be in the band. Eventually, in the limit, the unlucky students are weeded out entirely, leaving only those whose true scores are below the cut score. Thus, from this viewpoint, the Angoff method moves closer to achieving its goal as the reliability of the test scores increases.

This interpretation of the purpose of the Angoff method may be in conflict with the judges' intentions in practice. The judges are actually interested in observed scores, not true scores. They are interested in performance on the day, not how such students would score on a long-run average (which is what the true score measures). Students may have scored below the cut score on this particular occasion due to an unfortunate set of circumstances, but in the long run, over different samplings of items, they would tend to score above the cut score. For example, on this occasion, the topics they most intensively studied may not have appeared on the examination paper, or the teaching they experienced on those particular items may have been unusually poor. Whatever the reason, the performance they exhibited on this occasion was typical of the poor level of performance associated with the bottom band. The judges, being interested in the observed score performance on the day, collect samples of the various levels of performance for exhibition in Standards Packages, distributed on compact disks to schools. They do not (and cannot) differentiate between a performance that was a one-off poor result and a performance from a student who would repeatedly score in the lowest band. If the judges maintain the viewpoint that it is the observed scores on the day that matter, then the paradox is not resolved. They have defined and confirmed a standard for the particular occasion of testing, but then they are puzzled that testing on more reliable measures changes the standard they have set.

The resolution of the paradox depends on adopting the viewpoint that it is the percentage of true scores that matters. This has implications for removing the ambiguity about what should be the appropriate percentage of students in the bottom band, as noted earlier when discussing the various possibilities in Figure 1. Under this viewpoint, the appropriate percentage is the percentage of students with true scores below *C*.

A Key Relationship

It can be shown that the Angoff and bookmark cut scores are examples of true score equating, the two equated scores being the original test cut score and the hypothetical lengthened test cut score, where the lengthening is accomplished through the addition of parallel parts. The key relationship for this to occur is that if the test is lengthened by a factor, k, then the new cut score is k times the original cut score. That is,

$$C_Y = kC_X. \tag{1}$$

If this holds, then it can be shown that the relationship between C_X and C_Y is equivalent to that obtained from a true score equating.

Angoff Method

Under the Angoff method, the judgments are made by test item. Therefore, if the sum of the item cut scores on the original part were C_X, then it follows that the sum of the item cut scores for each parallel part (in an ideal judging) would also be C_X. Thus, the total cut score on Y would be given by Equation 1.

Bookmarking

In the bookmark method, the judges work through a booklet of items in ascending bookmark difficulty location (BDL) order (Beretvas, 2004). They consider whether borderline students would have a given probability, or higher, of correctly answering each item. This probability is called the response probability (often set at two thirds but sometimes at .5). When the judges find the item that the borderline students are likely to get wrong, they place a "bookmark" just before the item. The BDL of the previous item thus sets a cut score θ_c on the item difficulty/person ability scale. In a lengthened test (and an ideal judging), the judges would select the equivalent set of BDLs as being the appropriate cut score, and the average of this cluster of parallel BDLs would also estimate θ_c.

Let the original test comprising n items be denoted as Part 1. Now the expected score on an item (for ability θ_c) is given by the product of each score category value (h) and the probability of obtaining that score (P), summed over all score categories. For polychotomously scored items, where the maximum possible score is m, the expected score on item i on Part 1 may be written as

$$\hat{V}_{1i} = \sum_{h=0}^{m} h\hat{P}(X = h|\theta_c). \tag{2}$$

The algebraic expression for the probability (P) will vary according to the item response theory model used. For dichotomously scored items, $m = 1$. In this case, as the product will be zero for a score of zero, the expected item score is simply the probability of getting the item correct.

The expected cut score on the total test (Part 1) is then simply the sum of the expected item scores over all items. This is given by

$$C_X = \sum_{i=1}^{n} \hat{V}_{1i}(\theta_c). \tag{3}$$

On the lengthened test comprising k parallel parts, the expected cut score corresponding to θ_c is given by

$$C_Y = \sum_{i=1}^{n} \hat{V}_{1i}(\theta_c) + \sum_{i=1}^{n} \hat{V}_{2i}(\theta_c) + \sum_{i=1}^{n} \hat{V}_{3i}(\theta_c) + \ldots (k \textit{ terms}). \tag{4}$$

As the hypothetical parts are parallel, the expected cut score on the lengthened test would be given by $C_Y = kC_X$, which is Equation 1.

Angoff and Bookmarking Cut Scores as Examples of True Score Equating

Let Test X be the original test that is lengthened by a factor k to obtain Test Y. Let μ_X and σ_X denote the mean and standard deviation population parameters for Test X. Then the cut score for this test may be written in terms of a z score as

$$C_X = \mu_X + \sigma_X Z_X. \tag{5}$$

The mean for Test Y is given by summing the means over the parallel parts:

$$\mu_Y = \sum_{p=1}^{k} \mu_X = k\mu_X. \tag{6}$$

Now using the key Equation 1, the cut score on Test Y may be expressed as a z score as follows:

$$Z_Y = \frac{kC_X - \mu_Y}{\sigma_Y}.$$

Substituting from Equation 6, this z score may be written as

$$Z_Y = \frac{k(C_X - \mu_X)}{\sigma_Y}. \tag{7}$$

Substituting Equation 5 into Equation 7 gives

$$Z_Y = \frac{\sigma_X}{\sigma_Y} kZ_X. \tag{8}$$

Expressing the standard deviation of Y in terms of the variance of each parallel part gives

$$Z_Y = \sqrt{\frac{\sigma_X^2}{k\sigma_X^2 + k(k-1)\rho_{XX}\sigma_X^2}}\, kZ_X, \tag{9}$$

where ρ_{XX} is the reliability of X and is the correlation between parallel parts. This simplifies to

$$Z_Y = \sqrt{\frac{k}{1 + (k-1)\rho_{XX}}} Z_X. \tag{10}$$

Equation 10 shows how the z score borderline in the lengthened test (Y) relates to the z score borderline in the original test (X). This equation was used to derive the graphs in Figures 1 and 2, using the parameters $Z_X = -1.2$ and $\rho_{XX} = .5$. The associated percentages in the bottom band were obtained by using normal distributions. This example was chosen to vividly illustrate the theory, the reliability being far lower than one would typically encounter in reputable large-scale tests.

The Spearman-Brown formula (Feldt & Brennan, 1993) gives the reliability of the total test scores in terms of the reliability of the part scores by the equation

$$\rho_{YY} = \frac{k\rho_{XX}}{1 + (k-1)\rho_{XX}}. \tag{11}$$

From Equations 10 and 11 we obtain

$$Z_Y = \sqrt{\frac{\rho_{YY}}{\rho_{XX}}} Z_X. \tag{12}$$

Equation 12 is recognizable as one that derives from the definition of linear true score equating. From Levine (1955), C_X is linearly equated to C_Y by the equation

$$C_Y = \mu_Y + \sqrt{\frac{\rho_{YY}}{\rho_{XX}}} \frac{\sigma_Y}{\sigma_X}(C_X - \mu_X). \tag{13}$$

By algebraically converting to z scores, it can easily be seen that Equations 12 and 13 are equivalent. This implies that retaining a constant cut score for each component, as the test is lengthened, is equivalent to performing a true score equating.

Standard Setting Where the Percentage Failed Does Not Depend on Reliability

There are standard-setting methods where the percentage failed does not depend on the reliability of the test scores. These are methods that use *observed score equating*. Typically, they are not based on judging the difficulties of items but, instead, directly compare the quality of student work samples. These work samples are usually found by selecting them from the same percentiles in the two tests. The judges are required to judge whether the latest set of work samples are of the same standard, better, or worse than the earlier samples. They may then be given samples at a higher or lower percentile, depending on their initial judgments, and so on.

Consider the scenario where the same population sits for two tests, the original test and a lengthened, parallel version that is more reliable. Work samples at the 10th percentile in each test may be selected. This starting point represents equipercentile observed score equating. Given that the population of examinees is the same, and no improvement in performance over time has taken place in our hypothetical example, then it is likely that the judges would conclude that there had been no change in standards and retain the cut score at the 10th percentile.

The linear analog of equating percentiles is given by

$$C_Y = \mu_Y + \frac{\sigma_Y}{\sigma_X}(C_X - \mu_X), \tag{14}$$

which is the linear definition of observed score equating (Angoff, 1971).

In mentioning that there are methods for which the percentage failed does not depend on the reliability of the scores, it is not intended to imply that these are superior methods. Such a judgment should be based on a range of factors, in particular, the consistency of the decisions over different occasions, judging panels, and so on.

A Formula to Modify the Cut Scores

The dependence of the percentage failed on the reliability of the test scores yields two related problems. The first is that there is a *nonuniqueness* in the percentage failed. Logically, to give meaning to the percentage failed, one should specify the reliability of the particular test scores and give estimated values of the percentage who would have failed at other reliability values. It has been a long-standing view in the measurement and statistical literature that one usually wishes to generalize beyond the specific items that appear in a particular test. This generalization would be to tests that are similar to the current one but that would comprise different items that could lead to different test score reliabilities. The arbitrariness of having the percentage failed depend on the reliability can be dispelled by the data illustrated in Table 1. This, however, is cumbersome and awkward in reporting.

A second problem is the ethical one of failing a higher percentage of students than would have been failed had true scores been available. The above equations can easily be used to estimate the percentage failing under a perfectly reliable test (as will be shown below). If the standard-setting method fails 16%, but had true scores been used, only 12% would have failed, students may be entitled to query the examining body. This problem may be alleviated by using the above equations to modify the cut score so that the failure rate is equal to that had true scores been used (as in the last column of Table 1). At the same time, this overcomes the nonuniqueness problem.

Table 1
Example of Percentages Failing at Different Reliabilities

	Current Test	Parallel Tests of Differing Reliabilities				
Reliability	.75	.80	.85	.90	.95	1.00
% Failed	15.87	15.09	14.35	13.67	13.02	12.41

From Equation 12, in the limit as Test Y is increased in length, the reliability approaches 1. Thus, the z score cut score for true scores is equal to

$$Z_Y = \frac{1}{\sqrt{\rho_{XX}}} Z_X. \tag{15}$$

This is the z score that estimates the percentage of students whose true score is below C_Y. Thus, the modified cut score, C'_X, when expressed as a z score on the Test X scale, should equal that in Equation 15, giving

$$\frac{(C'_X - \mu_X)}{\sigma_X} = \frac{1}{\sqrt{\rho_{XX}}} Z_X.$$

Rearranging gives

$$C'_X = \mu_X + \frac{1}{\sqrt{\rho_{XX}}} Z_X \sigma_X. \tag{16}$$

Making μ_X the subject of Equation 5 and substituting into Equation 16 gives

$$C'_X = C_X + Z_X \sigma_X \left(\frac{1}{\sqrt{\rho_{XX}}} - 1 \right). \tag{17}$$

An alternative version of Equation 17 is obtained by replacing Z_X, using Equation 5 to get

$$C'_X = C_X + (C_X - \mu_X) \left(\frac{1}{\sqrt{\rho_{XX}}} - 1 \right). \tag{18}$$

Thus, the original cut score, C_X, is modified by a term that is a function of the reliability of the Test X scores. If this reliability is very high, then the correction is small. For minimal competency standard setting, the cut score is invariably below the mean so that the modification lowers the cut score slightly. In estimating C'_X, sample estimates of the population parameters would be used in Equation 18. Although the focus in this article has been on minimal competency, and hence the bottom achievement band, Equations 17 and 18 are obviously also applicable to high-level achievement bands above the mean.

Table 2
The Effect of Adjusting the Cut Scores

Course	α	z Cut	Cut	Adjusted Cut	Difference	Fail %	Adjusted Fail %
ESL	.884	−1.58	22.0	20.2	−1.8	7.47	5.95
English Standard	.877	−1.92	16.0	14.2	−1.8	3.35	2.36
Math	.945	−1.62	15.0	14.0	−1.0	6.01	4.76
Math Extension	.918	−1.63	21.4	19.9	−1.5	6.53	5.13
History Extension	.675	−1.14	52.0	48.7	−3.3	10.08	6.43
IPT	.902	−1.77	24.0	22.2	−1.8	5.56	4.66

Note: ESL = English as a Second Language; IPT = Information Processing and Technology.

Examples of the Adjustment

The adjustment in Equation 18 was applied to courses from the Year 12 public examination system in New South Wales, Australia, a program primarily for students leaving secondary school at age 18. The courses selected were ESL (English as a Second Language), English Standard (the main English course), Mathematics (the main Mathematics course), Mathematics Extension (a high-level Mathematics course), History Extension (a high-level History course), and IPT (Information Processing and Technology).

The bottom achievement band cut scores had been obtained through a modified Angoff judging process with six judges per course, the final cut score being the average of the six judgments. All marks reported are out of 100. The reliabilities were determined from Cronbach's α.

Table 2 shows Cronbach's α, the cut score as a z score (z Cut), the raw cut score, the adjusted cut score (using Equation 18), the difference between the raw and adjusted, the percentage failed under the raw cut score, and the percentage failed under the adjusted cut score. As expected, the α reliability coefficients for the Mathematics courses were higher than for the English, and much higher than for History Extension. The History examination comprised two essay questions in 2 hours (1 hour for each essay). Each question supplied source material that candidates had to read and use in responding to the question that followed. Each essay response was independently double marked, with a third marking used if the two markings diverged significantly.

As can be seen from Table 2, the reduction in the cut score using Equation 18 varied from 1 mark in Mathematics to 3.3 marks in History Extension, with corresponding reductions in the percentages in the failing band. Apart from History Extension, which had a relatively low reliability, the reduction in cut scores was between 1 to 2 marks on a scale out of 100.

The Effect on Examinees Misclassified

It is important to emphasize that lowering the cut score to reflect the failure rate had true scores been available is not a panacea for the problems of unreliability of the test scores. The errors of measurement in the individual test scores remain, so the candidates in the bottom band will still comprise a mixture of those whose true scores lie in this band and those who are only in the band due to negative errors of measurement. One cost of the modification is that the percentage of cases that deserve to fail (on the basis of their true scores) but are actually passing will rise slightly. No complaints should arise from these candidates. However, modifying the cut score will reduce the incidence of those students who do not deserve to fail yet are currently failing due to errors of measurement.

A discussion of the expected proportions of misclassified examinees using three-parameter item response theory, and assuming normal distributions, is given by Rudner (2001). The simulations below also use normal distributions. To gain a general idea of what would happen when the adjustment of Equation 18 is applied, the following simple model was simulated.

The observed score (X), with mean μ_X and standard deviation σ_X, may be written as the sum of a true score and error score, generated as follows:

$$True = \mu_X + \sigma_X \frac{A}{\sqrt{A^2+1}} R_1, \tag{19}$$

$$Error = \sigma_X \frac{1}{\sqrt{A^2+1}} R_2, \tag{20}$$

where R_1 and R_2 are uncorrelated, randomly generated $N(0,1)$ variables and A is a parameter that controls the reliability. The reliability (ρ_{XX}) is the ratio of true to observed variance, which from Equations 19 and 20 is given by

$$\rho_{XX} = \frac{A^2}{A^2+1}. \tag{21}$$

Thus,

$$A = \sqrt{\frac{\rho_{XX}}{1-\rho_{XX}}}. \tag{22}$$

Using these equations, sets of distributions were generated with normally distributed true scores and normally distributed error scores that were uncorrelated with the true scores. For each set, the relationship between true and error score was controlled by the scaling constant (A) to produce score sets with four reliability values: .75, .80, .85, and .90. These distributions were established with $\mu_X = 50$ and $\sigma_X = 16$, giving them an effective mark range from 0 to 100. The cut score for these distributions was set at 27 (/100), which gave a failure rate similar to many courses in the New South Wales program. Although these simulations may not exactly mirror the properties of many real-life distributions, which may vary from

Table 3
Pass/Fail Percentages for Varying Reliabilities

			Original Cut Score			Modified Cut Score		
Reliability			Pass	Fail	Total	Pass	Fail	Shift
.75	True	Pass	91.13	4.00	95.13	93.10	2.03	−1.97
	scores	Fail	1.34	3.53	4.87	2.03	2.84	−0.69
		Total	92.47	7.53	100.00	95.13	4.87	−2.66
.80	True	Pass	91.11	3.50	94.61	92.63	1.98	−1.52
	scores	Fail	1.37	4.03	5.39	1.98	3.42	−0.61
		Total	92.47	7.53	100.00	94.60	5.40	−2.13
.85	True	Pass	91.12	2.93	94.05	92.21	1.84	−1.09
	scores	Fail	1.33	4.62	5.95	1.84	4.11	−0.51
		Total	92.46	7.54	100.00	94.05	5.95	−1.59
.90	True	Pass	91.25	2.28	93.53	91.92	1.61	−0.67
	scores	Fail	1.24	5.23	6.47	1.61	4.86	−0.37
		Total	92.49	7.51	100.00	93.53	6.47	−1.04

the normal in skew, kurtosis, and so on, they should show the general effects of modifying the cut score.

Table 3 shows this data. For each reliability value, a set of scores (true and observed) comprising 10,000 examinees was generated. The percentages for each classification in the table were then calculated. This entire process was then repeated 100 times and the percentages averaged over 100 replications. The data in Table 3 comprises these mean statistics.

It is assumed here that the original cut score would have been appropriate had true scores been available, so it has been used to determine pass/fail on the true score scale. On the left side of the table, the original cut score is also used to determine pass/fail for the observed scores. On the right side on the table, the modified cut score from Equation 18 is used to determine pass/fail for the observed scores. Of particular interest are the two categories of misclassification: passing on true scores and failing on observed, and failing on true scores and passing on observed. As shorthand, denote the former as Category A and the latter as Category B. Then the lowering of the cut score via Equation 18 will reduce the percentage of examinees in A and increase the percentage in B.

From Table 3 it can be seen that Equation 18 has been successful in achieving an observed score pass rate that is the same as the percentage passing under true scores (95.13%, 94.60%, 94.05%, and 93.53% for reliabilities of .75, .80, .85, and .90, respectively). The effect of using Equation 18 is to shift examinees to the left along each row from the fail to the pass category. This shift is recorded in the far right-hand column. For a reliability of .75, 1.97% of examinees shifted out of Category A (decreasing from 4.00% to 2.03%). This is an important result as these are the students who deserved to pass on true scores but have failed because of errors of measurement. The price to be paid for this result is that 0.69% of examinees have shifted into Category B (increasing from 1.34% to 2.03%). These are the examinees who deserved to fail on true scores but passed due to errors of measurement. On balance, it seems a good result as the favorable shift is much larger than the unfavorable one at this relatively low level of reliability.

Similar results occur at the other reliability values. As one approaches the higher reliability levels, the results are attenuated, and it would appear that the ratio of the percentage shifting into Category A to the percentage shifting into Category B decreases. For a reliability of .90, overall an extra 1.04% were passed, with 0.67% shifting into Category A and 0.37% shifting into Category B. This attenuation of the effect at high reliabilities is expected from Equations 17 and 18.

Modifying the Cut Score Using the Standard Error

It is sometimes the case that an examining body will lower the cut score by subtracting a multiple of the standard error of measurement (Cizek, 1996). Although persuasive arguments can be made for this, it does not solve the nonuniqueness problem of the percentage failed depending on the reliability. For example, suppose the cut score was lowered by 1 standard error. Then the new *z* score for Test *X* will be given by

$$Z'_X = \frac{(C_X - \sigma_X\sqrt{1-\rho_{XX}} - \mu_X)}{\sigma_X}. \tag{23}$$

For the lengthened test *Y*, the new *z* score would be

$$Z'_Y = \frac{(kC_X - \sigma_Y\sqrt{1-\rho_{YY}} - \mu_Y)}{\sigma_Y}. \tag{24}$$

Simplifying Equations 23 and 24 gives

$$Z'_X = Z_X - \sqrt{1-\rho_{XX}}, \tag{25}$$

$$Z'_Y = Z_Y - \sqrt{1-\rho_{YY}}. \tag{26}$$

Combining Equations 12, 25, and 26, it can be shown that

$$Z'_Y = \sqrt{\frac{\rho_{YY}}{\rho_{XX}}}\left(Z'_X + \sqrt{1-\rho_{XX}}\right) - \sqrt{1-\rho_{YY}}. \tag{27}$$

Hence, the cutoff *z* score on the lengthened test is still a function of the reliability of the scores. In the limit as *Y* is lengthened, Equation 27 reduces to Equation 12, as expected.

This type of adjustment is one of several options that systems have adopted to gain a conservative estimate of the cut score. Other possibilities are to estimate the conditional standard error of measurement at the cut score point itself or to estimate the standard error of the mean (or the median) of the judges' independent decisions (MacCann & Stanley, 2004). If any of these is subtracted from the cut score to obtain a conservative estimate, then the value is being subtracted from a point estimate that moves with the reliability of the scores. Although these cases are not amenable to a simple analysis, it seems unlikely that the resulting value will be independent of the reliability of the scores. This is certainly not intended to discourage the practice of publishing the standard error of measurement or the standard error of the mean (or median) of the judges' decisions, which should be essential features of standards reporting, but is merely intended to confirm that the ambiguous nature of the percentage failed is likely to remain when these measures are used to reduce the cut score.

Discussion

This article has shown that two very prominent standard-setting methods are effectively examples of true score equating. However, when forced to work with observed scores in the real world, this leads to two problems. The first is the dependence of the percentage failed on the reliability of the test scores (which, in the context of parallel tests, means the test length). Given that the test length is usually influenced by practical constraints, there is a disconcerting nonuniqueness about the failure rate.

In addition, there is the ethical issue of failing students on the basis of errors of measurement. This obviously cannot be solved with our imperfect tests. However, it can be ameliorated to some extent by failing no more than would have been failed had true scores been available. Imagine a scenario where students and the public were fully knowledgeable about true scores and errors of measurement. Would they be satisfied with an examining body failing 16%, knowing that had true scores been used, only 12% would have failed? If only 12% truly deserved to fail, why is the examining body failing 16%? This problem may be overcome by using Equation 18 to modify the cut score so that the percentage failing is equal to that had true scores been used. By doing this, the system will simultaneously overcome the nonuniqueness problem by fixing the failure rate at that corresponding to perfect reliability.

A further issue concerns the relative costs of inevitably making some false decisions on the basis of a fallible measuring instrument. The context of this article is that of measuring achievement in high school students across a range of courses that are designed to provide a broad education. In such circumstances, one is usually inclined to give a student the benefit of the doubt in adjusting a cut score. Lowering the cut score via Equation 18 will reduce the number of students who failed due to errors of measurement but deserved to pass on true scores.

A reduction in these false negatives is usually considered desirable in this situation. The cost of using Equation 18, of increasing the numbers of students who passed but who deserved to fail, is considered to be small. The consequences of passing such students on tests designed to give a broad general education are not likely to have serious or life-threatening consequences. After all, these students would have marks close to the cut score, and the cut score itself may have been set at a different value had another panel of judges been appointed.

A good example of a system that strives to avoid false negatives comes from the legal profession through the well-known saying "Better that 10 guilty persons escape than that one innocent suffer." However, there are circumstances where the opposite emphasis is appropriate—on avoiding false positives. These circumstances are usually concerned with certifying competence in areas where incompetence could be life threatening—for example, aircraft pilots or medical practitioners. Here the cost of certifying incompetence could far outweigh concerns about failing students who really deserved to pass. Administrators of such systems would be unlikely to use Equation 18 to lower the cut score. Ultimately, it is a policy decision by the system administrators, weighing the costs of false negatives and false positives, that would determine whether lowering the cut score is appropriate.

The concept of making a final adjustment to the cut score is not new, as mentioned in the discussion on standard errors. Good arguments can be made for a final adjustment on ethical grounds as argued above or even on legal grounds (for example, Biddle, 1993). If a system were to use Equation 18 for a final adjustment, then the issue of how best to estimate the reliability of the scores may arise. Many systems would probably use Cronbach's α as a quick and convenient measure. However (assuming uncorrelated errors), unless the parts are essentially tau equivalent, α tends to underestimate the reliability (Novick & Lewis, 1967). If this is the case, then Equation 18 will overadjust, providing a conservative estimate of the cut score. This conservative estimate may be exactly what some systems are looking for. However, if the parts are not tau equivalent and also have correlated errors, then this creates two opposing biases and α may either underestimate or overestimate the reliability, depending on the relative strengths of the biases (Komaroff, 1997). If the test comprises a mixture of item types, systems requiring a more exact estimate of reliability could apply a structural equation modeling technique (Raykov, 2001).

Modifying the cut score via Equation 18 will have the most impact on tests with a relatively low level of reliability. Although it will not solve the problems of unreliable tests, it will reduce the incidence of an important class of error in the context of measuring school student achievement—those students who deserve to pass yet are currently failing due to errors of measurement. The percentage of these cases will drop with the modification. The correction also has the merit of giving a unique and defensible interpretation of the failure rate, something that at present is clouded in ambiguity.

References

Angoff, W. H. (1971). Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508-600). Washington, DC: American Council on Education.

Beretvas, S. N. (2004). Comparison of bookmark difficulty locations under different item response models. *Applied Psychological Measurement*, *28*, 25-47.

Berk, R. A. (1996). Standard setting: The next generations (where few psychometricians have gone before!). *Applied Measurement in Education*, *9*, 215-235.

Biddle, R. (1993). How to set cutoff scores for knowledge tests used in promotion, training, certification and licensing. *Public Personnel Management*, *22*, 63-70.

Brennan, R. L., & Lockwood, R. E. (1980). A comparison of the Nedelsky and Angoff cutting score procedures using generalizability theory. *Applied Psychological Measurement*, *4*, 219-240.

Busch, J. C., & Jaeger, R. M. (1990). Influence of type of judge, normative information, and discussion on standards recommended for the National Teacher Examinations. *Journal of Educational Measurement*, *27*, 145-163.

Chang, L. (1999). Judgmental item analysis of the Nedelsky and Angoff standard-setting methods. *Applied Measurement in Education*, *12*, 151-165.

Cizek, G. J. (1996). Standard setting guidelines. *Educational Measurement: Issues and Practice*, *15*, 13-21.

Cross, L. H., Impara, J. C., Frary, R. B., & Jaeger, R. M. (1984). A comparison of three methods for establishing minimum standards on the National Teacher Examinations. *Journal of Educational Measurement*, *21*, 113-130.

Feldt, L. S., & Brennan, R. L. (1993). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105-146). Phoenix, AZ: Oryx Press.

Giraud, G., Impara, J. C., & Buckendahl, C. (2000). Making the cut in public schools: Alternative methods for standard setting. *Educational Assessment*, *6*, 291-304.

Goodwin, L. D. (1999). Relations between observed item difficulty levels and Angoff minimum passing levels for a group of borderline examinees. *Applied Measurement in Education*, *12*, 13-28.

Hambleton, R. K. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods and perspectives* (pp. 89-116). Mahwah, NJ: Lawrence Erlbaum.

Harasym, P. H. (1981). A comparison of the Nedelsky and modified Angoff standard setting procedure on evaluation outcomes. *Educational and Psychological Measurement*, *41*, 725-734.

Jaeger, R. M. (1993). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 485-514). Phoenix, AZ: Oryx Press.

Komaroff, E. (1997). Effect of simultaneous violations of essential tau-equivalence and uncorrelated error on coefficient alpha. *Applied Psychological Measurement*, *21*, 337-348.

Levine, R. S. (1955). *Equating the score scales of alternate forms administered to samples of different ability* (ETS Research Bulletin No. 23). Princeton, NJ: Educational Testing Service.

Livingston, S. A., & Zieky, M. J. (1989). A comparative study of standard-setting methods. *Applied Measurement in Education*, *2*, 121-141.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

MacCann, R. G. (2004). Reliability as a function of the number of item options derived from the "knowledge or random guessing" model. *Psychometrika*, *69*, 147-157.

MacCann, R. G., & Stanley, G. (2004). Estimating the standard error of the judging in a modified-Angoff standards setting procedure. *Practical Assessment, Research & Evaluation*, *9*(5). Retrieved September 5, 2006, from http://www.pareonline.net/getvn.asp?v=9&n=5

Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed.), *Setting performance standards* (pp. 249-281). Mahwah, NJ: Lawrence Erlbaum.

Novick, M. R., & Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements. *Psychometrika*, *32*, 1-13.

Raykov, T. (2001). Estimation of congeneric scale reliability using covariance structure analysis with nonlinear constraints. *British Journal of Mathematical and Statistical Psychology*, *54*, 315-323.

Rudner, L. M. (2001). Computing the expected proportions of misclassified examinees. *Practical Assessment, Research and Evaluation*, *7*(14). Retrieved July 25, 2006, from http://pareonline.net/getvn.asp?v=7&n=14

Wang, N. (2003). Use of the Rasch IRT model in standard setting: An item-mapping method. *Journal of Educational Measurement*, *40*, 231-253.