2008; 30: 836–845 MEDICAL TEACHER

AMEE GUIDE

Setting and maintaining standards in multiple choice examinations: AMEE Guide No. 37

RAJA C. BANDARANAYAKE

Abstract

The process of setting a standard when pass/fail decisions have to be made inevitably involves judgment about the point on the test score scale where performance is deemed to be adequate for the purpose for which the examination is set. As with any process which involves human judgment, setting this standard is likely to include a certain degree of error, which may result in some false positive and false negative decisions. The customary practice of maintaining a constant point on the test score scale at which pass/fail separations are made cannot be justified, as examinations vary in difficulty. The aim of standard setting procedures is to minimize such errors while accounting for the varying difficulty of examinations.

A standard may be norm-referenced, where it is dependent on the performance of the particular group of examinees, or criterion-referenced, where it is based on predetermined criteria, irrespective of examinee performance. Where certification of competence is the primary purpose of an examination, the latter is preferred as the decision to be made is whether an individual is competent to practise rather than competent compared to peers. Several methods of standard setting have been used, some of which are based solely on predetermined criteria, while others compromise between norm- and criterion-referenced standards.

This guide examines the more commonly used methods of standard setting, illustrates the procedure used in each with the help of an example, and discusses the advantages and disadvantages associated with the use of each. The common errors made by judges in the standard setting process are pointed out and the manner in which judges should be selected, trained and instructed emphasized. A method used for equating similar tests set at different times with the intention of maintaining standards from one examination to the next is illustrated with an example. Finally, the guide proposes a practical method for arriving at a pre-determined standard by the proportionate selection of test-items of known relative difficulties in relation to minimally competent examinees.

Introduction

Simply stated, standard setting is the process of determining *how much is good enough*. In medical education the standard is intimately associated with the notion of competence. Competence, like all attributes, is measured along a scale, and is hence a continuous variable. The *standard*, or criterion level of performance, is a point on this scale at which a separation of competence and incompetence occurs. This is an artificial but necessary dichotomy imposed on the continuous variable. The terms *cut-score*, *cut-off score* and *passing score* are synonymous terms which represent this standard or criterion level on a given test for making decisions pertaining to the purpose for which the test was conducted, such as to certify competence.

Measurement on the scale of competence is associated with error. This error may arise from several sources, including the measuring instrument (test), measurer (examiner) or subject of measurement (examinee). The *true score* of an individual in a particular aspect of competence, say in Anatomy, is a conceptual measure indicating the true extent of competence that the individual possesses. The *observed score*, which the individual is assigned as a result of taking a test in Anatomy purporting to measure competence in this subject, has an

Practice points

- Standards set for examinations which certify competence should be criterion-referenced rather than norm-referenced.
- All standard setting methods involve judgment, with the possibility of false positive and false negative errors around the cut-off point
- The degree of error can be substantially reduced by the proper selection, training and monitoring of judges.
- While several standard setting methods are available, the Angoff method is the most popular, though the flexibility afforded by the Hofstee method, is more acceptable.
- Studies directed towards validation of the method used should be undertaken in the initial stages of its use, so that the method can be defended on scientific grounds.
- Standards can be maintained by test equating methods using "marker questions' from previous examinations to determine the relative difficulty of each examination".
- A practical procedure would be to specify the performance standard and develop a test to fit that standard, rather than apply a standard setting procedure to an existing test (Kane, 1994).

Correspondence: Raja C. Bandaranayake, 20 Strickland Street, Rose Bay, NSW, Australia 2029. Tel: +61-2-93717739; fax: +61-2-93717739; email: rajabanda@yahoo.com.au

ISSN 0142–159X print/ISSN 1466–187X online/08/09–100836–10 © 2008 Informa UK Ltd.
DOI: 10.1080/01421590802402247

inherent error within it, which may be either positive or negative. Rarely do true and observed scores coincide, unless positive and negative errors are equivalent, as educational measurement is rarely devoid of measurement error. The reliability of a test, and the related statistic called the *standard error of measurement* (SE_{meas}), are estimates of the amount of error in such measurement.

"The passing (cut-off) score is a particular point on the observed score scale that is used to make decisions about examinees, whereas the standard is a conceptual boundary (on the true score scale) between acceptable and non-acceptable performance" (Cusimano, 1996). As each examinee's score is referenced to the cut-off score a competence/incompetence decision is made. As both examinee's score and the cut-off score are on the observed score scale there is likely to be measurement error associated with the score. Thus there is also likely to be a classification error in the decision made at the borderline between competence and incompetence. This error may be a *false positive* (passing and incompetent examinee) or a *false negative* (failing a competent examinee).

The standard is an artificial, but necessary, dichotomy imposed on the continuous variable of competence in order to make a decision pertaining to the competence of an individual in the given field in which a test is taken.

Justification

In medical education the significance of setting an appropriate standard is obvious as the decision based on the standard set has the potential, not only to alter the potential careers of examinees, but also, and more importantly, to affect the lives of those whom examinees certified as competent would serve. Certification examinations have as their main purpose the determination of an individual's competence so as to protect the public from unsafe practice. The negative consequences of certifying an incompetent examinee (false positive) may far outweigh those of not certifying a competent one (false negative) (Cusimano, 1996). Despite all the efforts that may be taken to maintain a valid and reliable examination, the examination process would be a failure if the cut-off score is not set properly. The high stakes nature of licensing and certifying examinations mandates careful scrutiny of the manner in which standards are set (Fielding et al., 1996).

No standard setting procedure exists, or will ever be found, where human judgment is not involved. This single fact remains the critical issue in the debate about standard setting. This is probably the reason why none of the methodological developments in standard setting has had universal acceptance. Gross (1985) points out that, in spite of the fact that in all walks of life important decisions are made based on judgment, considerable resistance remains to judgmental standard setting in testing. However, he cautions, "judgmental standard setting is not tantamount to capricious standard setting". The best one can hope for under the circumstances is to make as informed a judgment as possible, and to reduce, as far as possible, error in the measurement on which such judgment is based.

The decision based on the standard set has the potential, not only to alter careers of examinees, but, more importantly, to affect the lives of those whom examinees certified as competent would serve.

Types of standards

Fundamentally, there are two types of standards: *norm-referenced* and *criterion-referenced*.

A norm-referenced standard is a relative standard based on the performance of a group of examinees at the same examination for which the standard is set. Thus the standard varies with the performance of the group of examinees. A given examinee's performance is judged relative to the performance of the entire group, rather than on its own merits. Thus an examinee has a better chance of achieving the standard if placed in a relatively weak group of examinees than in a relatively strong group. While the process of arriving at a norm-referenced standard is much simpler than that of arriving at a criterion-referenced standard, there can be no assurance that the standard is equivalent from one examination to the next, as examinee group performance may vary between the two examinations and cut-off scores are determined by group score distributions. Cut-off points based on norm-referenced standards are not relevant to making judgments of competence or incompetence of an examinee. Therefore they are not appropriate for licensing examinations where the aim is to certify competence to ensure safe practice (Norcini, 1994). The important issue is to determine whether an individual is a safe practitioner rather than safe compared to others (Holmes, 1986). On the other hand, they are well suited when a desired number of passes is required, such as in selection examinations, or when ranking of examinees is a goal.

A criterion-referenced standard is an absolute standard which is referenced to a specified level of examinee performance on a given examination. Each examinee is judged in relation to this absolute standard irrespective of the performance of the examinee group in that examination. A criterion-referenced standard is prescribed prior to the administration of the examination. While both false positive and false negative errors may result from either norm-referenced or criterion-referenced standards, the latter are preferable in high-stakes licensure and certifying examinations, because of the risk of false positive errors with the former (Boulet et al., 2003). This is because norm-referencing does not tie the standard to a criterion of competence.

From the above discussion it is apparent that we should basically be concerned with criterion-referenced standards, especially when they are being set for certification purposes. The following section will review some methods by which such standards are set. However, compromise methods, which use both criterion-referencing and norm-referencing in a particular examination, are being increasingly used currently, and will also be reviewed.

Norm-referenced standards are not relevant to making judgments of competence or incompetence of an examinee. Criterion-referenced standards should be used for such examinations.

Methods of setting criterion-referenced standards

Several methods of setting criterion-referenced standards have been described in the literature. These methods have been classified in various ways.

(1) Test-centred and examinee-centred methods (Kane et al., 1994; Boulet et al., 2003)

- *Test-centred standards* are those derived from hypothetical decisions based on the test content. In these methods a group of expert judges set the standard by reviewing the items in the test and deciding on the level of examinee performance on these items that will be considered just adequate for demonstrating competence. Methods included in this category are the Nedelsky method, the Angoff method and its modifications and the Ebel method (see below).
- *Examinee-centred standards* are those derived from reviewing the performance of examinees or a similar group prior to making judgments about what constitutes borderline performance between competence and incompetence. Methods included in this category are the borderline group method (Livingstone & Zieky, 1982) and the contrasting group method (Livingstone & Zieky, 1982).

(2) Berk (1986) proposed another system of classification of methods for determining criterion-referenced standards, based on the extent to which performance data influence the judgment involved.

- *Judgmental methods*, which are based on the judgment of one or more persons independently, without prior review of performance data (e.g. Nedelsky, Angoff and Ebel methods).
- *Judgmental-empirical methods*, which are based on judgments of one or more persons with performance data made available before the judgments are made (e.g. Hofstee method, described below).
- *Empirical-judgmental methods*, which are based primarily on performance data from one or more groups of examinees (e.g. borderline & contrasting group methods, described below).

Each of the methods of standard setting mentioned above, irrespective of the system of classification used, will be described briefly below.

A. Test-centred methods

The Nedelsky (1954) method

Each of a panel of judges reviews each multiple-choice item in an examination and identifies those response options that a minimally competent examinee should be able to eliminate as incorrect. The minimum passing level (MPL) *for that item* is the reciprocal of the number of remaining options. *For each judge* the MPL is the sum of these reciprocals for all items in the examination. The average MPL so obtained for all judges is the cut-off score *for the examination* (see Example 1).

Example 1
In an examination paper containing n items each of which has 5 options, Judge A identifies, in item 1, 2 options as those which a minimally competent examinee should be able to eliminate as incorrect.
The MPL for that item for Judge A (MPLA1) $= 1/(5-2) = {}^{1}/_{3}$.
Similarly for item 2 he identifies 3 options, giving an MPLA2 of $1/(5-3) = ½$.
The MPL for the entire paper for Judge A (MPLA) is (MPLA1) + (MPLA2) + (MPLA3) + ⋯ (MPLAn).
Similar values are obtained for Judges B, C, . . . , N, where N = number of judges. The MPL for the examination = (MPLA + MPLB + MPLC + . . . MPLN)/N.

While the Nedelsky method was originally designed for the single-response type of multiple choice questions, each of which had 4 or 5 response options, it can also be used for multiple true-false questions. If each true/false option is considered a separate question, then the MPL for that question would be either 0.5 or 1.0. In the case of single response questions which have, say, 5 options each, application of the Nedelsky formula would yield an MPL for each question which ranges from 0.2 to 1.0.

The Angoff (1971) method

This method is also based on the judgments of expert judges in relation to minimal competence. A panel of judges first meets to discuss the characteristics of a "borderline" examinee, i.e. a minimally competent individual who is at the borderline of pass and fail. Each judge is asked to consider a number, say 100, of minimally competent individuals and estimate the proportion of this number who would answer the item correctly. Once all the judges have made their independent judgments with regard to all the items in the examination, group discussion may take place among them to explain gross differences in their judgments. Judges may now independently alter their previous judgments if they desire. Once they have completed this process each judge's estimates for all the items are summed to obtain that judge's MPL. The average MPL of all the judges would represent the cut-off point for making pass/fail decisions (see Example 2).

Example 2
Each of *N* judges considers 100 minimally competent individuals taking an examination of *n* items.
Judge A estimates that, of these individuals, 50 would answer item 1 correctly, 20 item 2, 70 item 3 and so on. The MPL for Judge A $(MPL_A) = (0.5 + 0.2 + 0.7 + \cdots x_n)/n \times 100 = A\%$.
Similarly, for Judges B, C, D & E . . . *N*, the MPLs are B%, C%, D%, E% . . . N%, respectively.
The MPL (cut-off score) for the examination = (A% + B% + C% + D% + E% + ⋯ N%)/*N*.

Several variations of the Angoff procedure have been described. One set of variations is based on the degree of freedom given to judges in making their judgments. For example, they may be given a three-choice option of 'yes',

'no' or 'don't know' to the question "Can a person with minimal competence answer the item correctly?" (Nassif, 1978 cited by Berk, 1986); or they may be given a multiple-choice format to select from with seven choices of percentages ranging from 5% to 95% with a 'don't know' option (Educational Testing Services, 1976, cited by Berk, 1986). In a further variation of the latter, a nine-choice option is used followed by adjustments for standard error of the mean and for random guessing (Bernknopf et al., 1976, cited by Berk, 1986). Other variations are based on the degree of freedom given to judges to interact with each other after they have made independent Angoff judgments to arrive at consensus. It is preferable, however, for judges to be allowed to alter their initial judgments independently after such discussion, as in the study by Norcini et al (1988). In yet another variation, previous performance data on the item, such as difficulty index, are provided to the judges before they make judgments. Many authors, however, refer to 'a modified Angoff' method without specifying the modification used, creating difficulties in evaluating it.

The Ebel (1972) method

This method too is based on the judgment of a group of experts. A panel of judges rates each test item along two dimensions: perceived difficulty (*easy, medium, hard*); and relevance (*essential, important, acceptable, questionable*), assigning it to one cell in a 3×4 matrix (see Example 3). The judge then estimates the percentage of items in each cell that a minimally acceptable individual should be able to answer correctly. The MPL for each judge is obtained by multiplying the number of items in each cell by its respective estimate (expressed as a percentage), summing the products of all cells and dividing by the total number of items. This process yields a weighted average. The cut-off score is the mean MPL for all the judges.

Comparison of test-centred methods

The three methods described above are based on the notion of a borderline candidate with minimal competence. This is difficult for many judges to conceptualize, leading to considerable subjectivity and variation among them. As will be described later, variation can be considerably reduced through training and discussion, but individual idiosyncracies are inevitable. Meskauskas and Webster (1975) found that MPLs determined by 6 judges using the Nedelsky method varied from 36% to 80%. While there is always some arbitrariness in setting standards, some system is better than no system.

The Nedelsky method yields a lower cut-off score than either the Angoff or Ebel methods. Andrew & Hecht (1976) found that the corresponding cut-off scores derived from the Nedelsky and Ebel methods on the same examination were 49% and 68%, respectively, while Harasym (1981), using medical school faculty as judges, found that 99% and 88% of students taking an examination would have passed using the Nedelsky and modified Angoff methods, respectively.

Proponents of standard setting methods argue that one cannot expect the three methods to yield identical cut-off scores as they are based on different philosophical conceptualizations, even though the underlying assumption of minimal competence is common. Different results would have been acceptable if the intention was to measure different things. However, when labels of mastery or incompetence are affixed to judgments based on these methods, congruence is essential for validity considerations. The lack of such evidence has probably resulted in a reluctance to depend solely on these methods on a wider scale. The compromise methods described below may have received wider acceptance for this reason.

It is difficult to find clear evidence to support one method over the others. The lack of external criteria of validity, such as the actual performance of those certified according to standards determined by each method, precludes valid comparisons. Berk (1986) has identified two sets of criteria for evaluating standard setting methods: technical adequacy (e.g. sensitivity to performance, statistical soundness, predictive validity) and practicability (e.g. ease of implementation, interpretation, public credibility). In an exhaustive review he concludes that, of the above three methods, the Angoff method offers the best balance between technical adequacy and practicability. This method is convenient to use, is flexible in that it allows improvements in specific procedures to overcome limitations or problems identified, and has a relatively small standard error for passing scores (Kane, 1994).

B. Examinee-centred methods

Borderline group method (Livingston & Zieky, 1982)

In this method the judges are requested to judge a group of individuals as borderline candidates, based on their previous experience or some procedure other than the test itself. The scores on the test of these candidates are arranged in rank order and the median score for the group is taken as the cut-off score.

Example 3

	Essential	Important	Acceptable	Questionable
Easy	$15^1 \times 100\%^2 = 1500$	$20 \times 80\% = 1600$	$10 \times 50\% = 500$	$10 \times 30\% = 300$
Medium	$25 \times 80\% = 2000$	$40 \times 60\% = 2400$	$25 \times 40\% = 1000$	$15 \times 20\% = 300$
Hard	$10 \times 60\% = 600$	$20 \times 50\% = 1000$	$5 \times 10\% = 50$	$5 \times 0\% = 0$

[1 = number of items; 2 = estimated% of these items which a minimally competent individual would answer correctly, according to Judge A]
MPL for Judge A (MPL_A) = (1500 + 1600 + 500 + 300 + 2000 + 2400 + 1000 + 300 + 600 + 1000 + 50 + 0)/200 = 56.25%.
MPL for examination (cut-off score) = average MPL for all judges.

If this method is used, it is important that the scores of the group should form a cluster rather than be spread out. If they do not form a cluster the method is not applicable.

Contrasting groups method (Livingston & Zieky, 1982)

The same authors described a second examinee-centred method in which the judges categorize a sample of examinees into two groups, competent ("qualified") and incompetent ("unqualified"), based on any knowledge they have from their previous performances, but not from the test itself. A score that best discriminates these two groups, with or without the use of statistical analysis, is chosen as the cut-off score. For example, the score distributions can be plotted, and the point of intersection of the two distributions taken as the cut-off score.

C. Compromise methods

As pointed out above, a reluctance to be solely dependent on test-centred or examinee-centred methods, due to validity considerations stemming from the subjectivity of judgments, has led to the use of compromise methods. These, while depending to a large extent on one of the judgmental methods described above, provides flexibility for adjusting the standard based on performance data in the examination for which the standard has been determined. Thus, in effect, these methods are a compromise between absolute and relative standards, to prevent gross deviations from average pass rates. Each method consists of two stages: (1) an estimation phase, in which judgmental data are obtained and an estimated cut-off score determined; (2) an establishment phase, in which the estimated cut-off score may be accepted or adjusted after considering the effect of using the estimated cut-off score on pass rates (Mills & Melican, 1988).

Hofstee (1983) method

In this method judges set about determining the cut-off score using one of the methods above, say the Angoff method. They are also called upon to state what the minimally acceptable ($c_{\min}$) and maximally acceptable ($c_{\max}$) cut-off scores are. They, or others, also agree on what would be the minimally acceptable ($f_{\min}$) and maximally acceptable ($f_{\max}$) failure rates are for the examination. Possible cut-off points are then plotted against resulting failure rates (see graph below), and point A, corresponding to the minimally acceptable cut-off with maximally acceptable failure rate ($c_{\min}$, $f_{\max}$), and point B, corresponding to the maximally acceptable cut-off with minimally acceptable failure rate ($c_{\max}$, $f_{\min}$) are found. The cut-off score corresponding to the point at which the line joining A and B intersects the distribution curve is taken as the operational cut-off score for the examination. This cut-off score would, obviously, give an acceptable failure rate for the examination (see Example 4 and Figure 1).

De Gruijter's (1980) model

This model is similar to the Hofstee method in that it can be used by individual judges or a group of judges. In addition to each judge stating the ideal cut-off score and the corresponding failure rate, he/she also estimates the degree of uncertainty in relation to this judgment. These values are then used to determine the cut-off score.

Beuk's (1984) model

In this method, each judge is asked to independently state the minimum level of knowledge expressed as a percentage of the total score that a candidate should possess to pass the given examination, and also the expected percentage pass rate. The mean and standard deviation of each of these values are determined and used to determine the cut-off score.

The mathematical procedures involved in the last two procedures are omitted from this guide for the sake of simplicity.

In a study comparing relative, absolute (modified Angoff) and compromise (Hofstee) methods of standard setting, Fielding et al. (1996) concluded that relative methods were not appropriate for high stakes examinations, while the other two methods were. The modified Angoff method worked well but was time-consuming to apply, as was the Hofstee method. The latter had the further disadvantage that the judges felt uncomfortable estimating maximum and minimum acceptable failure rates.

Example 4
A plot of cut-off scores for a given examination against resulting failure rates is given below:
$c_{min} = 40\%$
$c_{max} = 45\%$
$f_{min} = 10\%$
$f_{max} = 20\%$
A = point representing c_{min}, f_{max}
B = point representing c_{max}, f_{min}
Line AB intersects the curve at a cut-off point of 42.5%
Thus, operational cut-off score = **42.5%**.

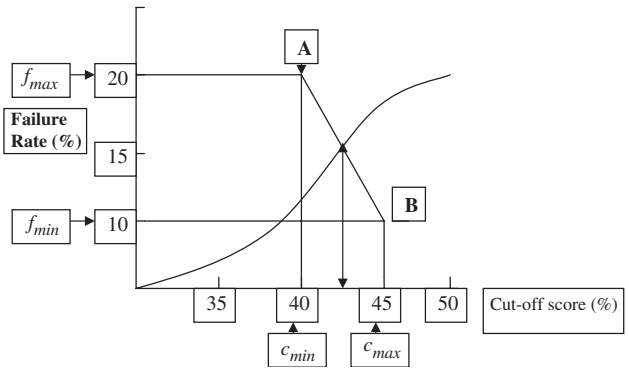

Figure 1. The Hofstee curve.

Norm-referenced standards should not be used in professional examinations unless the purpose is to rank candidates and select a given number for purposes such as admission from a pool of applicants. Criterion-referenced standards should be used for licensing examinations which attempt to separate ''safe'' from ''unsafe'' practitioners. Of the test-centred methods, the Angoff method is the most popular. However, in order to compensate, to some extent, the subjectivity of the judgments of experts, the Hofstee method is the most popular of the compromise models, as it resulted in more ''reasonable'' failure rates (Case et al, cited by Cusimano, 1996).

The validity and reliability of standards

The basic definition of validity of any set of scores obtained from a test is the extent to which those scores accurately reflect a measurement of what the test purports to measure. In the case of a test standard, validity is a reflection of the extent to which the proposed cut-off score, based on which pass-fail decisions are made, represents a performance standard which separates the incompetent from the competent.

Kane (1994) points out that, when a passing score is identified for making a decision on competence, two assumptions are made: (1) the passing score is an accurate reflection of the performance standard specified as separating competence from incompetence; (2) the specified performance standard is appropriate for the decision that has to be made. If a relationship is established between the passing score and the performance standard the first assumption is validated. However carefully a passing score is established, by paying particular attention to the procedure used to do so, establishing its relationship with the performance standard requires evidence external to the standard setting process. Kane refers to this as external validity. If, however, adequate procedural precautions are not taken in establishing the standard, it can be invalidated. The difficulty in establishing the external validity of a standard stems from the problems associated with obtaining criterion measures of performance, in professional practice, without contamination by external variables, such as post-test learning and contextual factors.

Evidence of internal validity of the standard is obtained through the stability of the passing score, if several samples of judges are used to arrive at the standard for the same test. As judgment is a subjective process, some degree of variability is to be expected. However, as with determining the reliability of a test, the variance among the samples of judges can be used to estimate the standard error of the passing score as a measure of the reliability of their judgments. While reliability is prerequisite for validity, it does not establish internal validity. For example, if all judges were to do Angoff ratings within a narrow range, the standard error of the ratings would be low (and reliability correspondingly high), but the standard may yet be inappropriate.

Procedural evidence relates to both the appropriateness of the procedures used, and the manner in which those procedures are implemented. Thus standard setting bodies must ensure that an appropriate procedure is set in place for setting the standard, and that the implementation of this procedure is monitored to ensure that it is carried out as planned.

Validating a standard involves two steps: determining the extent to which the passing score is an accurate reflection of the performance standard; and obtaining evidence that the specified performance standard is appropriate to the decision that has to be made from the test results.

The reliability of a performance standard is reflected by the degree of agreement among the judges (Ben-David, 2000). It is thus obvious that reliability would be substantially increased (1) by training judges in the standard setting procedure; and (2) if the standard is revised after discussion among judges (Fielding et al., 1996).

The role of judges in setting standards

The judges are the key elements in the standard setting process. Their appropriate selection, training, interaction and monitoring are critical steps in increasing the chances of arriving at a defensible standard. These steps will be addressed in this section.

The process of judgment is necessarily a subjective one with inherent variability. The aim should thus be to reduce this subjectivity as much as possible, while recognizing the fact that there is no known strategy for producing objective judgments (Berk, 1986). It is well known that, in general, a high degree of consensus is rarely found in the standards set by different judges on the same examination. As pointed out earlier, there is little consistency in the standards arrived at by different methods on the same examination. Colton & Hecht (1981), cited by Gross (1985), reported that rater consistency was lower for the Nedelsky method than for the Angoff method, while different results have been reported when the Nedelsky and Ebel methods are compared (Skakun & Kling, 1980). Even when the same technique is applied to the same examination on different occasions, different results have been reported (Holmes, 1986). The latter must stem from both inter-judge and intra-judge variability.

The standard set by a given judge is influenced by factors within the judge, in the social milieu in which the process is carried out and by interaction between these two sets of factors. For example, a ''weak'' judge could be influenced by the opinions of peers when comparisons are made between their respective standards (Meskauskas & Norcini, 1980). Many judges who have not clearly understood the process may feel that ''they are pulling probabilities from thin air'' (Berk, 1986). Judges who are not well versed in the standard setting process may have difficulty eliminating the pass mark which they are accustomed to from their minds, when they are making judgments. They tend to either forget or misunderstand the concept of ''borderline candidate'' and unwittingly set an inordinately high standard by basing their judgments on pass rates across the entire group of candidates.

The concept of ''expert judge'' is subject to many interpretations. What is the degree of expertise in a given discipline that is necessary to be a judge in, say, an undergraduate examination? Does level of content expertise have a bearing on the judgments made? Teachers may be able to rank test items according to difficulty level, but may not accurately estimate levels of examinee performance (Impara & Plake, 1998). Verhoven et al. (2002) found that, while teachers

who wrote items and recent graduates were equally capable of ranking items according to difficulty level, the former tended to overestimate expected student performance. It is often thought that judges may, because of their expertise, tend to look at an item from their own viewpoint rather than that of the candidate for whom the examination is set, thereby setting an inappropriately high standard. However, in a medical school adopting an integrated, problem-centred curriculum, where Angoff judges for objective structured practical examinations were multidisciplinary, Bandaranayake (2000) found judges within a discipline were significantly more lenient than those outside it, resulting in lower mean Angoff scores for the former.

The first step in increasing objectivity is the appropriate selection of judges. Criteria for selection include expertise in subject matter, familiarity with similar candidate groups, skill in conceptualizing and skill in self-monitoring (Ben-David 2000; Kane, 1994). While expertise in the subject matter is a *sine qua non*, a heterogeneous group, such as one including teaching faculty, practitioners and recent graduates is beneficial, as they bring different perspectives into the standard setting process (Kane, 1994).

The second step is adequate training in the process. As mentioned above, one major difficulty judges have is to understand and keep in mind the construct of "borderline candidate" or "minimally competent candidate". Judges often start off with different interpretations of this construct. The terms must be thoroughly discussed by the group of judges with the help of a facilitator. A common understanding must be arrived at, before independent judgments are made. Failure to do so would likely result in a divergent set of cut-off scores. In addition, judges, particularly novice ones, must be given explicit instructions on the purpose of the exercise and the process to be followed. The gravity of the decision which would be based on the product of their endeavours must be stressed. Periodic re-training of judges who have participated previously in such exercises would help to reinforce their learning. Kane (1994) recommends exposing judges to the consequences of setting different cut-off scores to help them realize the gravity of their judgments.

In spite of training, outliers in the form of "hawks" or "doves" inevitably occur among most groups of judges. One course of action is to eliminate the standards set by such outliers. A better procedure to follow is to provide an opportunity for judges to discuss their individual judgments before they contribute to a mean cut-off score. If the latter procedure is followed, the judges are given time to discuss their findings, bearing in mind the ultimate purpose of the exercise and the level of performance expected of the candidates for the role they would be called upon to play if they pass. Such discussion also serves the purpose of providing feedback to individual judges as part of their training for similar exercises in the future. Berk (1986), however, points out that group interaction produces a norming effect. Thus instead of setting the final cut-off score through a process of consensus, it is preferable for judges to undertake a second rating independently after the group discussion, with opportunity to change their earlier judgments if necessary (Norcini et al., 1988).

Finally, the examination coordinator must monitor the results obtained from different judges, paying particular attention to outliers in spite of their interaction. Reasons for such deviations must be identified and appropriate action taken. A data bank of the judgments of individual judges should be maintained by the examination coordinator, so that the history of each judge with regard to the judgmental process can be determined. Future selection of judges would be facilitated by such a data bank.

> Judges are the key elements in the standard setting process. They must be selected appropriately, trained adequately and monitored closely if the standard arrived at is to be valid and reliable.

Maintenance of examination standards

Previous sections of this guide have focused on setting standards for a given examination. Some examinations are set several times a year or are set annually. This is particularly true of postgraduate examinations where licences are granted to practise a given specialty. An example of this is the Part 1 examination of the Royal Australasian College of Surgeons (RACS), which is set three times a year across several centres in Australia, New Zealand and South East Asia. While the examination at all these centres at a given point of time is the same, examinations set at different times may vary in difficulty even though the same test specifications are used. The importance of maintaining a constant standard across examinations over a period of time is obvious, to ensure safe practice in this specialty. However, as Meskauskas & Norcini (1980) have pointed out, knowledge in various branches of Medicine changes rapidly, and the difficulty of an examination has to be considered in the context of the practice of the specialty at a given time. In other words, a test which may have been considered difficult a decade ago may not be considered so at the present time, because of the rapid advances in knowledge and procedures that are likely to have occurred in the intervening period. Nevertheless, attempts to maintain standards over a given period of time require procedures for obtaining data concerning the approximate equivalence of tests with regard to their difficulty level.

Test equating is "often viewed as the process of making statistical adjustments to the scores obtained on different forms of the same test to compensate for differences in relative difficulty" (Holmes, 1986). Through this process an attempt is made to increase the chances that consistent standards are maintained across several administrations of a licensing examination.

Holmes categorizes methods of test equating as follows:

(1) Single group methods, where two tests are administered to the same group at different times for comparison of standards. This is neither a practical method, because of the costs involved, nor is it likely to be legal. The RACS sets three MCQ papers, each of which contains 120 test items, on three consecutive days with the same test specifications. This practice offers an opportunity to compare difficulty levels of the three papers at a given examination. However, it does

not provide a solution to the comparison of difficulty levels of, say, the three examinations conducted in a given calendar year, as the candidate group varies from one examination to the next.

(2) Equivalent group methods, where each test is administered to approximately equivalent groups at the same time. The legal implications of such a practice are questionable, as different candidates are subjected to different tests.

(3) Anchor test methods, where each test is given to a different group, but a common anchor test, which is either a different test, or a subset of both tests, is given to both groups. The former (i.e. a different test) requires a candidate to sit two tests, and may give rise to complaints from candidates. A procedure such as this was used to determine intern placement positions for candidates who were successful at the final licensing examinations of the different medical schools in Sri Lanka, This, however, was for a different purpose from test equating – it was for deciding on an order of merit amongst all successful candidates. The latter, where a common subset of items is used in two tests to two different groups, offers the most promising possibilities for test equating. Slinde and Linn (1977) call this procedure "horizontal equating", and it allows statistical adjustments to be made to the scores on two tests to compensate for unintended differences in difficulty levels. A similar procedure is used by the RACS to equate tests and is described in the box below.

As Holmes is careful to point out, it is important for testing bodies to construct tests carefully in the first instance, rather than try to compensate for poor construction through such procedures as test equating. Careful construction includes ensuring content validity and content sampling through the appropriate use of test grids, as well as weeding out from test item banks those items which have repeatedly been found to give poor item analysis data in spite of attempts to improve their quality.

The procedure adopted by the RACS to equate tests is described below:

The probability of guessing in single-response MCQ with 5 options per item is 20%. Thus the "total ignorance score" is assumed to be 20%. The maximum possible score is 100%. The effective range of scores is, therefore, 20% to 100%. The mid-point of this range is 60%. A factor of 5% is added to the mid-point score to derive a nominal cut-off score of 65%. (RACS does not use test-centred standard setting procedures). The College wishes to maintain this standard in all its MCQ examinations by adjusting for the varying difficulty level of each examination. It does this with the use of a group of previously set questions, labeled "marker questions", in each examination as the anchor set for adjusting the cut-off score according to how the group of candidates who sat a given examination answered this set compared to previous groups of candidates.

The procedure involves four steps, as in Example 5 below:

(1) Comparison of examination scores
(2) Comparison of "marker question" scores
(3) Estimating relative examination difficulty
(4) Determining the cut-off score.

Example 5

1. **Comparison of examination scores**	
Mean score in this examination:	56.7%
Average examination mean score over last 4 years:	59.4%
Thus mean score in this examination is:	2.7% **lower**
Assuming this candidate group is of same standard as in the last 4 years, this examination is:	2.7% **harder**
2. **Comparison of "marker" scores**	
Mean score in this examination on previously used questions (N = 162):	62.5%
Mean score on same questions when they were each last used:	60.5%
Thus, compared with previous candidates, this group of candidates, on these items, scored (62.5-60.5)%=	2.0% **higher**
Thus this group of candidates is: than previous groups	2.0% **better**
3. **Estimating examination difficulty**	
Thus it is expected that their mean score in this examination would be:	2.0% **higher**
But their mean score in this examination is:	2.7% **lower**
Thus this examination is really:	4.7% **harder**
4. **Determining cut-off score**	
The cut-off level for an average examination is:	65.0%
Thus the cut-off level for this examination should be	(65–4.7)% = **60.3%**

De Gruijter (1985) suggested that once a cut-off score has been set for the first examination, rather than cut-off scores being determined through test-centred methods for subsequent administrations, a system of test equating (such as the above) should be practised. This requires, however, ensuring that the anchor test items are not revealed, if the same items are repeatedly used. Furthermore, if the number of candidates is small, equating cannot be achieved as statistical procedures may not be meaningful. A practical suggestion that follows has the potential to avoid repeated, expensive standard setting methods for successive tests.

Attempts to maintain standards over a given period of time require procedures for obtaining data concerning the approximate equivalence of tests with regard to their difficulty level. However, such procedures should not be used as compensation for poor construction of tests in the first instance.

A practical suggestion

Kane (1994) suggested that, rather than apply a standard setting procedure to an existing test, it would be better to specify the performance standard and develop the test to fit that standard. The test can be so constructed to yield "high precision around the passing score". Surprisingly, this advice seems to have been unheeded by testing bodies. The suggestion which follows has, to the author's knowledge, not been attempted before, but seems to have the potential to achieve a defensible standard, without actually carrying out the standard setting process each time a test is administered. It is partly based on the same principle that applies to the use of item analysis data on items stored in a MCQ bank for future use, namely that difficulty index of a given item is reasonably constant on repeated administrations to approximately equivalent groups of candidates. The method also requires the development of an adequately large bank of items for which Nedelsky or Angoff values have been determined from

previous administrations of each. The bank should be large enough to enable the examination to be set according to a predetermined table of specifications.

Each item in the bank is allocated by expert judges into one of the nine cages in the table below. When an examination is due, the specified number of items in each cage is selected from the bank. If required, further specification of the items in each cage, for example, according to disciplines, body systems, topics or themes, can be achieved to ensure a fair spread of items across those components which are examined. Assuming that difficulty levels remain fairly constant, the prespecified cut-off mark (say 60% in Example 6) would result.

Example 6.

Developing a test to fit a specified performance standard

	Easy	Medium difficulty	Hard
Essential	$6 \times 100\% = 600$	$12 \times 80\% = 960$	$7 \times 50\% = 350$
Important	$12 \times 80\% = 960$	$24 \times 60\% = 1440$	$19 \times 40\% = 760$
Acceptable	$5 \times 60\% = 300$	$12 \times 50\% = 600$	$3 \times 10\% = 30$

MPL $= (600 + 960 + 350 + 960 + 1440 + 760 + 300 + 600 + 30)/100 =$ **60%**

Rather than apply a standard setting procedure to an existing test, it would be better to specify the performance standard and develop the test to fit that standard.'' (Kane, 1994)

Conclusions

In high-stakes examinations important decisions are made in regard to competence and incompetence, which may affect, on the one hand, the careers of professionals, and on the other, the safety of the professional's clients. Thus the standard setting process cannot be taken lightly, nor can arbitrary standard be used as is customary in many examinations. While judgment is inevitable in any standard setting process, and errors are likely to be made at the boundary between competence and incompetence, the aim should be to make the process used as objective as possible. The proper selection, training and instruction of judges cannot be overestimated in this regard. While pure criterion-referenced standards are determined prior to measuring the performances of a given group of examinees, the consequences of applying such standards may not always be acceptable because of gross departures from customary pass rates (or failure rates). Thus compromise methods are being increasingly used to heighten acceptability of standard-setting.

The practical issues involved in setting standards, as well as the widely different results that are produced when different standard setting procedures are used on the same examination, may be the reasons for a general reluctance to use such procedures. Test equating methods hold promise for maintaining examination standards once they have been set. A practical procedure has been suggested above which, though requiring a considerable amount of preliminary work for examining bodies as with any standard setting process, is likely to lighten the burden in the long term in future examinations of a similar nature.

As examining bodies increasingly realize the significant effects of the decisions they make based on the examinations conducted by them, and as the public increasingly realize their legal rights pertaining to such decisions, it is inevitable that defensible standard setting methods would be used more commonly and carefully. The choice of method would, however, be the prerogative of each examining body, but the effects of the choice must be meticulously investigated in the process of establishing it as the method of choice.

Declaration of interest: The author reports no conflicts of interest. The author alone is responsible for the content and writing of the article.

Notes on contributor

RAJA C. BANDARANAYAKE, MBBS, PhD, MSEd, FRACS is a retired professor of Anatomy and an international consultant in medical education. His most recent appointments were as Director and Associate Professor in the School of Medical Education, University of New South Wales, Sydney, Australia, and as Professor & Chairman, Department of Anatomy, College of Medicine & Medical Sciences, Arabian Gulf University, Bahrain. He was Deputy Chairman of the Board of Examiners, Chairman of the Examinations Committee, and Chairman of the Anatomy Sub-committee, of the Royal Australasian College of Surgeons, as well as Technical Adviser and Member of the Board of Examiners of the Australian Medical Council. As a member of a working party of the World Federation for Medical Education, and of a core committee of the Institute for International Medical Education in New York, he has contributed to the development and testing of international standards in medical education. As a consultant of international repute he has facilitated the development of curricula, including student assessment and programme evaluation procedures, in many health professional schools around the world.

References

Andrew BJ, Hecht JT. 1976. A preliminary investigation of two procedures for setting examination standards. Educ Psycholog Measur 36:45–50.

Angoff WH. 1971. Scales, norms and equivalent scores. In: Thorndike RL, editor. Educational Measurement. 2nd ed. Washington DC: American Council on Education. pp 508–600.

Bandaranayake R. 2000. Content expertise of Angoff judges. In: Proceedings of the Ninth International Ottawa Conference on Medical Education. Cape Town, 1-3 March 2000.

Ben-David MF. 2000. Standard setting in student assessment. Med Teach 22:120–130, (AMEE Guide No. 18).

Berk RA. 1986. A consumer's guide to setting performance standards on criterion-referenced tests. Rev Educ Res 56:137–172.

Beuk CH. 1984. A method for reaching a compromise between absolute and relative standards in examinations. J Educ Measure 21:147–152.

Boulet JR, de Champlain AF, McKinley DW. 2003. Setting defensible performance standards on OSCEs and standardized patient examinations. Med Teach 25:245–249.

Colton DA, Hecht JT. 1981. A preliminary report on a study of three techniques for setting minimum passing scores. Presented at the Annual Meeting of the National Council on Measurement in Education. New York.

Cusimano MD. 1996. Standard setting in medical education. Acad Med 71:S112–S120.

De Gruijter DNM. 1985. Compromise models for establishing examination standards. J Educ Measure 22:263–269.

Ebel RL. 1972. Essentials of Educational Measurement. Englewood Cliffs, NJ: Prentice-Hall.

Fielding DW, Page GG, Rogers WT, O'Byrne CC, Schulzer M, Moody KG. 1996. Standard setting for a test of Pharmacy practice knowledge: Application in high-stakes testing. Am J Pharmaceut Educ 60:20–29.

Gross LJ. 1985. Setting cutoff scores on credentialing examinations. Eval Health Profess 8:469–493.

Harasym PH. 1981. A comparison of the Nedelsky and modified Angoff standard-setting procedure on evaluation outcome. Educ Psycholog Measure 36:45–50.

Hofstee WKB. 1983. The case for compromise in educational selection and grading. In: Anderson SB, Helmick JS, editors. On Educational Testing. San Francisco: Jossey-Bass. pp 109–127.

Holmes SE. 1986. Test equating and credentialing examinations. Eval Health Profess 9:230–249.

Impara JC, Plake BS. 1998. Teachers' ability to estimate item difficulty: A test of assumptions in the Angoff standard setting method. J Educ Measure 35:69–81.

Kane M. 1994. Validating the performance standards associated with passing scores. Rev Educ Res 64:425–461.

Livingston SA, Zieky MJ. 1982. Passing scores: A manual for setting standards of performance on educational and occupational tests. Princeton, NJ: Educational Testing Service.

Meskauskas JA, Norcini JJ. 1980. Standard setting in written and interactive (oral) specialty certification examinations: Issues, models, methods, challenges. Eval Health Profess 3:321–360.

Meskauskas JA, Webster GW. 1975. The American Board of Internal Medicine recertification examination process and results. Ann Inter Med 82:577–581.

Mills CN, Melican GJ. 1988. Estimating and adjusting cutoff scores: Features of selected methods. Appl Measure Educ 1:261–275.

Nedelsky L. 1954. Absolute grading standards for objective tests 14:3–19.

Norcini JJ. 1994. Research on standards for professional licensure and certification examinations. Eval Health Profess 17:160–177.

Norcini JJ, Shea JA, Hancock EW, Webster GD, Baranowski RA. 1988. A criterion-referenced examination in cardiovascular disease. Med Educ 22:32–39.

Skakun EN, Kling S. 1980. Comparability of methods for setting standards. J Educ Measure 17:229–235.

Verhoeven BH, Verwinjen GM, Muijtjens AMM, Scherpbier AJJA, van der Vleuten CPM. 2002. Panel experstise for an Angoff standard setting procedure in progress testing: Item writers compared to recently graduated students. Med Educ 36:860–867.