

Setting standards on educational tests

John J Norcini

Objective This instalment in the series on professional assessment provides an introduction to methods of setting standards.

Method A standard is a special score that serves as a boundary between those who perform well enough and those who do not. The practical steps in selecting it include: deciding on the type of standard; deciding the method for setting it; selecting the judges; holding the meeting; calculating the cutpoint, and deciding what to do afterwards. Four of the more popular methods are illustrated for both written and clinical examinations.

Results The most important criteria for selecting a method for setting standards are whether it is consistent with the purpose of the test, based on expert judgement, informed by data, supported by research, transparent, and requires due diligence. The credibility of the standard will rely largely on the nature of the standard setters and the selection of a broadly repre-

sentative and knowledgeable group is essential. After the standard has been set, it is important to ensure that stakeholders view the results as credible and that the pass rates have sensible relationships with other markers of competence.

Conclusions A standard is an expression of professional values in the context of a test's purpose and content, the ability of the examinees, and the wider social or educational setting. Because standards are an expression of values, methods for setting them are systematic ways of gathering value judgements, reaching consensus and expressing that consensus as a single score on a test.

Keywords education, medical, undergraduate/* standards; *educational measurement; *standards; reproducibility of results; Great Britain.

Medical Education 2003;37:464-469

Introduction

On a test, a standard or cutpoint is a special score that serves as the boundary between those who perform well enough and those who do not. In the broadest sense, it is an expression of professional values in the context of a test's purpose and content, the ability of the examinees, and the wider social or educational setting. The fact that standards are an expression of values means that a 'method' for setting them is not a technique for divining a scientifically correct solution. Instead, it is a systematic way of gathering value judgements, reaching consensus, and expressing that consensus as a single score on a test. Because standards are judgmental, the methods used to set them will not differ in terms of their ability to discern the 'truth'. They will, however, vary in their credibility according to who sets the

standards, the characteristics of the method they use, and the outcome.^{1,2,3}

The purpose of this instalment in the professional assessment series is to provide an introduction to methods of setting standards. It is based largely on previous work and describes the six steps in setting standards using four of the more popular methods.^{1,2} It also includes a brief section on the application of the methods to clinical examinations.

Steps in setting standards on written examinations

Step 1: Deciding on the type of standard

There are two types of standards, relative and absolute, and the decision between them is related to the purpose of the test. Relative standards are expressed as a number or percentage of examinees, so the cutpoint is set, for example, at the score that passes the 50 best performers or separates the top 20% from the bottom 80%. Relative standards are most appropriate for examinations where the purpose is to identify a certain

Foundation for Advancement of International Medical Education and Research (FAIMER®), Philadelphia, Pennsylvania, USA

Correspondence: John J Norcini PhD, 3624 Market Street, 4th Floor, Philadelphia, Pennsylvania 19104, USA. Tel: 00 1 215 823 2170; Fax: 00 1 215 386 2321; E-mail: jnorcini@ecfmg.org

Key learning points

Methods for choosing a standard are systematic techniques for gathering and reaching consensus on professional values.

The steps in selecting a standard or the cutpoint are: deciding on the type of standard; deciding the method for setting it; selecting the judges; holding the meeting; calculating the cutpoint, and establishing what to do afterwards.

As standards are expressions of values, the most important contributors to their credibility are the number and nature of the judges.

number of examinees. This includes tests that are used to select the highest or lowest scorers for admissions or placement, where a limited number of students can be accommodated.

Absolute standards are expressed as a number or percentage of the test questions, so that the cutpoint is set, for example, at 70 correct responses of the 100 questions (70%) on the test. Absolute standards are most appropriate for tests of competence, where the purpose is to establish that the examinees know enough for a particular purpose. These include final or exit examinations and tests for certification and licensure.

Step 2: Deciding on the method for setting standards

The credibility of the results of any standard setting exercise will be enhanced if the method produces standards that are consistent with the purpose of the test and based on expert judgement informed by data about examinee performance. Further, the exercise should not be so brief as to indicate a lack of reasonable consideration nor so long as to be unduly burdensome to the standard setters. A method that is supported by a body of published research is preferable as a means of justifying the final result, as is a method that is transparent, and thus easy to implement and explain.

Over time, a variety of different methods have evolved and they all have advantages and disadvantages depending on the specific application. Thorough surveys of these methods can be found elsewhere,^{2,5-8} but, according to a classification scheme developed by Livingston and Zieky,⁴ they fall into four categories: relative methods, absolute methods based on judgements about test questions, absolute methods based on judgements about individual examinees, and compromise methods. For this paper, four methods corresponding to each of these categories will be

described: the Fixed Percentage method, Angoff's method, the Contrasting Groups method, and Hofstee's method.^{2,4-10}

Fixed Percentage method

In this method, each judge is asked what he/she believes is the percentage of the examinees qualified to pass. They then engage in a discussion during which judges are free to change their estimates. At the end of the discussion, the judges' pass rates are averaged to arrive at the cutpoint.

This method is easy to use; judges are comfortable in applying it, and it is equally applicable to different forms of the same test and different types of tests (i.e. clinical and written examinations). However, this method produces relative standards so they are independent of test content and how much of it the examinees know. Further, standards set in this fashion will vary from one administration to the next, depending on the ability of the group of examinees taking the test. Hence, the Fixed Percentage method is best suited to situations where there is a desire to identify a certain number of the best (or worst) examinees.

Angoff's method

In this method, judges are asked to first define the characteristics of a borderline group of examinees (a group with a 50% chance of passing). They then consider the difficulty and importance of the first item on the test. Each judge estimates what percentage of the hypothetical borderline examinees will respond correctly to the item. This judgement is often informed by data on the performance of the examinees. The judges discuss their estimates and are free to change them, and then proceed in the same manner through the remainder of the items on the test. The judges' estimates are averaged for each item and the cutpoint is set at the sum of these averages. Table 1 presents an example of the results of an application of Angoff's method by six judges to a 12-item test. The judges began by defining the characteristics of a borderline group of examinees. They then considered the difficulty and importance of the first item on the test. Each judge estimated what percentage of the hypothetical borderline examinees would respond correctly to the item and their estimates were recorded in view of the group. The judges discussed them, changed them as they wished, and then proceeded in the same manner through the remaining 11 items on the test. The judges' estimates were averaged for each item and the cutpoint is the sum of those averages.

Angoff's method is relatively easy to use, there is a sizeable body of research to support it, and it is frequently

Table 1 Example of the results of an application of Angoff's method by six judges to a 12-item test

Question	Judge 1	Judge 2	Judge 3	Judge 4	Judge 5	Judge 6	Average
1	0.85	0.80	0.80	0.95	0.85	0.90	0.86
2	0.60	0.70	0.50	0.55	0.65	0.70	0.62
3	0.45	0.55	0.50	0.60	0.90	0.35	0.56
4	0.90	0.95	0.90	0.95	0.85	0.90	0.91
5	0.80	0.75	0.65	0.70	0.85	0.55	0.72
6	0.70	0.65	0.60	0.70	0.75	0.60	0.67
7	0.40	0.50	0.35	0.50	0.55	0.50	0.47
8	0.75	0.65	0.60	0.70	0.75	0.60	0.68
9	0.65	0.55	0.70	0.65	0.65	0.60	0.63
10	0.55	0.50	0.45	0.60	0.65	0.55	0.55
11	0.50	0.45	0.40	0.50	0.55	0.50	0.48
12	0.95	0.95	0.95	0.90	0.95	0.95	0.94
Cutpoint							8.09

applied in licensing and certifying settings. It also has the virtue of focusing attention on each of the questions and thus can be very helpful from a test development perspective. This method produces absolute standards, so it is best suited to tests that seek to establish competence. However, judges sometimes feel as though there is no firm basis for their estimates and application of the method can be tiresome for longer tests.

Contrasting Groups method

The first step in the application of this method is to draw a random sample of examinees. The judges then consider the responses of the first examinee to all of the questions on the test. As a group they make a decision (consensus or majority) about whether the performance is of a pass or fail level. This process is then repeated for all of the examinees in the sample. After the judgements are made, the scores of the passers and failers are graphed and the cutting score calculated in any of a number of ways (e.g. the point of least overlap between the distributions). Figure 1 presents an example.

Educators are comfortable making the types of judgements that form the basis of the Contrasting Groups method and those judgements are directly informed by the actual test performances of the examinees. Moreover, the method allows direct manipulation of the false positive and false negative rates. For example, it is possible to approximately equalise the number of false positives and negatives by selecting as the standard the point of least overlap in the score distributions. This method produces absolute standards, so it is best for tests of competence. However, it can be time consuming as the judges must review the entire test performance of each selected examinee and the sample size must be

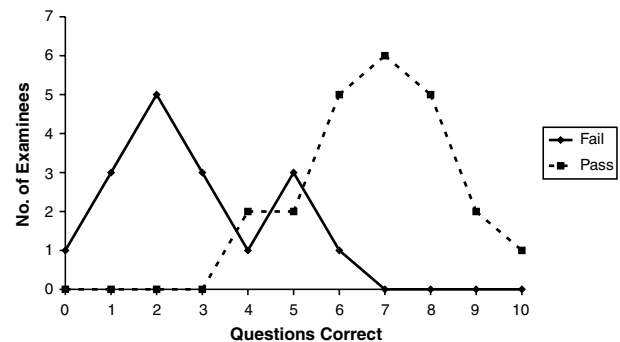


Figure 1 Example of an application of the Contrasting Groups method to a 10-item test. A random sample of 40 examinees is drawn and the judges consider the responses of the first examinee to all of the questions on the test. They decide as a group about whether the performance is passing or failing. This process is repeated for the other 39 examinees in the sample. After the judgements are made, the scores of the passers and failers are graphed separately and the cutting score calculated. In the example below, it is set at 7 if the intent is to minimise false positives, at 4 if the intent is to minimise false negatives and somewhere between to serve other purposes.

relatively large to ensure a reliable result. Furthermore, the exact selection of a strategy for calculating the cutpoint can be subjective.

Hofstee method

This method produces a compromise between relative and absolute standards. As is consistent with absolute standards, the judges are asked to specify the minimum and maximum acceptable cutpoints and, as is consistent with relative standards, they are also asked to indicate the minimum and maximum acceptable fail rates. Their responses serve as the focus for discussion,

with all being free to change their estimates. The final judgements are combined as in Fig. 2.

This method is easy to implement and judges are comfortable with the questions they are asked. Under some circumstances, however, the cutpoint may not be within the bounds they define and when this happens, the standard becomes the maximum or minimum acceptable pass rate. Consequently, this method would not be ideal for ongoing application in a high stakes test of competence, but is well suited to occasional use and lower stakes settings.

Step 3: Selecting the judges

Because standards are an expression of values, the most important contributors to their credibility are the number and nature of the judges. In a low stakes setting like a classroom, one judge, usually the teacher, is enough. In a high stakes examination, however, many more are needed to ensure the incorporation of a variety of perspectives and to produce reliable results. Six to eight judges are a minimum for these purposes.¹¹

More important than the number of judges are their characteristics. Specifically, a mix of professional roles

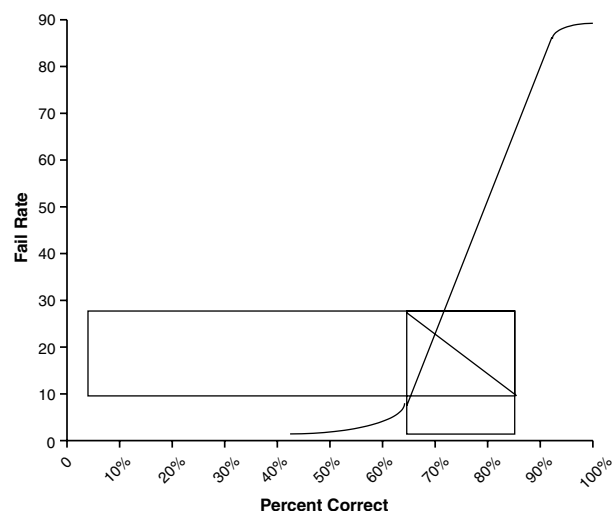


Figure 2 Example of an application of the Hofstee method. The judges are asked to specify the minimum and maximum acceptable cutpoints and fail rates. After discussion, the results are averaged and graphed to identify the rectangle bounded by fail rates of 10% to 28% and percent correct scores of 65% to 85%. A diagonal is drawn through the rectangle and the examinee performance curve is graphed (i.e. the percentage of examinees failing at every score on the test). The standard is the percent correct point where the diagonal intersects the examinee performance curve so it is 70% correct and the fail rate is 25%.

such as teacher, practitioner, generalist and specialist is advantageous, depending on the purpose of the test. A balance of personal attributes like gender, race and age is important, as is the avoidance of real or perceived conflicts of interest.

Step 4: Holding the standard setting meeting

Independent of method, there are a variety of important features that all standard setting meetings should have in common.

- All judges should attend throughout the meeting. Any absences will generate missing data and an absent judge's lack of participation in the dialogue will have an effect on others.
- The judges should discuss the purpose of the test, the characteristics of the examinees, and the nature of competence. This will serve to orient the group to the standard setting exercise and clarify some of the issues around how to make judgements.
- The judges should have detailed familiarity with the test content and format, although this can be accomplished before the meeting. This is most important for the Fixed Percentage and Hofstee methods because a systematic review of the content is not a part of the process.
- The method should be explained and the judges given practice before collecting data that will be used in the calculation of the standard.
- The judges should be given feedback throughout the process. Specifically, they need to know the consequences of their judgements at several points in the exercise.

Step 5: Calculating the standard

The exact calculation of the standard will vary by method but some issues are common to all. It is not unusual to have relatively small numbers of judges and this raises the possibility that one or two outliers could significantly influence the standard. In such instances, it may be reasonable to use the median instead of the mean or to remove the data of the judges with the highest and lowest standards.⁴ The removal of data should be done cautiously and only when its effect is sizeable, as this adversely affects the credibility of the standard.

As part of the calculation of the standard, it is important to determine whether the process leading to it is sufficiently reliable or reproducible for the purpose of the test. Reliability or reproducibility coefficients can be calculated in a number of ways and they address the issue of whether the cutpoint would be similar if the

standard setting exercise was repeated with a comparable group of judges.¹¹ If not, the exercise can be performed again or a second standard setting meeting with additional or different judges can be held.

Step 6: After the test

After the test has been given, it is important to ensure that the standard produces reasonable results. In an ongoing testing programme, this issue has three aspects. Firstly, stakeholders should be questioned to ensure that they view the results as credible. Secondly, the pass rates should be compared against contemporaneous markers of competence to ensure that they have the expected relationships. Finally, the results of applying a standard should be compared against future performance.

Setting standards on clinical examinations

All of the traditional standard setting methods were developed for multiple choice questions. The past three decades, however, have seen a significant increase in the use of clinical examinations and the need to set standards for them. Of the methods presented above, the Fixed Percentage and Hofstee methods can be directly applied to such tests without modification and without extra effort on the part of the judges.

Angoff's method and the Contrasting Groups method can also be applied directly to clinical examinations but they require considerably more work on the part of the judges. For example, using Angoff's method would be extremely time consuming if it required that judgements be gathered for every item on the checklists for every station in an objective structured clinical examination (OSCE). Likewise, using the Contrasting Group method would require that the judges evaluate all of the checklists for all of the OSCE stations for a sizeable number of examinees. To overcome these problems, both methods have been modified.

In one modified version of Angoff's method, judgements are captured at the level of the OSCE station rather than the individual items on the checklist. Consequently, each judge estimates the station score of a hypothetical borderline group of examinees. These estimates are averaged across the judges for each case and then summed across cases to arrive at the cutpoint, just as the proportions are in Table 1.

One very efficient variation on the traditional methods of setting standards takes advantage of the fact that physician-examiners are already scoring the performance of examinees on each station of an OSCE.¹² In support of the standard setting process, they are also asked to rate the examinees' responses to the needs/problems of the

patient on a 6-point scale where the options are outstanding, excellent, borderline pass, borderline fail, poor, and inadequate. For each case, the scores of the examinees rated as borderline are averaged and these averages are aggregated across all stations on the test.

A modification of this efficient variation, consistent with the Contrasting Groups method, considers as a 'pass' all scores of examinees rated as excellent or outstanding and as a 'fail' all scores of those rated as poor and inadequate. Station-by-station, these data would be treated just as they are in Fig. 1. Other modifications of the Contrasting Groups method have also been developed.^{13,14}

Conclusions

Methods for choosing a standard are systematic techniques for gathering and reaching consensus on professional values. This paper presented the practical steps in selecting the cutpoint including:

- 1 deciding on the type of standard;
- 2 deciding the method for setting it;
- 3 selecting the judges;
- 4 holding the meeting;
- 5 calculating the cutpoint, and
- 6 establishing what to do afterwards.

Four of the more popular methods were illustrated and references were provided to practical guides and more comprehensive reviews.^{2,4-8}

References

- 1 Norcini JJ, Shea JA. The credibility and comparability of standards. *Appl Measurement Educ* 1997;10:39-59.
- 2 Norcini JJ, Guille RA. Combining tests and setting standards. In: Norman G, van der Vleutin C, Newble D, eds. *International Handbook of Research in Medical Education* 2002;811-34.
- 3 Kane M. Validating the performance standards associated with passing scores. *Rev Educational Res* 1994;64:425-61.
- 4 Livingston SA, Zeiky MJ. Passing scores: a manual for setting standards of performance on educational and occupational tests. Princeton, New Jersey: Educational Testing Service 1982.
- 5 Jaeger RM. Certification of student competence. In: Linn RL, ed. *Educational Measurement*. New York: American Council on Education and Macmillan Publishing Co 1989.
- 6 Berk RA. A consumer's guide to setting performance standards on criterion-referenced tests. *Rev Educational Res* 1986;56:137-72.
- 7 Cusimano MD. Standard setting in medical education. *Acad Med* 1996;71:112-20.
- 8 Meskauskas JA. Evaluation models for criterion-referenced testing: views regarding mastery and standard setting. *Rev Educational Res* 1976;45:133-58.

- 9 Angoff WH. Scales, norms and equivalent scores. In: Thorndike RL, ed. *Educational Measurement*. Washington DC: American Council on Education 1971.
- 10 de Gruijter DNM. Compromise models for establishing examination standards. *J Educational Measurement* 1985;22:263–9.
- 11 Brennan RL, Lockwood RE. A comparison of the Nedelsky and Angoff cutting score procedures using generalisability theory. *Appl Psychol Measurement* 1980;4:219–40.
- 12 Dauphinee WD, Blackmore DE, Smee S, Rothman AI, Reznick R. Using the judgements of physician examiners in setting standards for a national multicentre high stakes OSCE. *Adv Health Sci Educ* 1997;2:201–11.
- 13 Clauser BE, Clyman SG. A contrasting groups approach to standard setting for performance assessments on clinical skills. *Acad Med* 1994;69:S42–S44.
- 14 Mills CN, Jaeger RM, Plake BS, Hambleton RK. An investigation of several new methods for establishing standards on complex performance assessments. Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, California, 1998.

Received 2 October 2002; editorial comments to author 10 October 2002; accepted for publication 13 January 2003