

Standard setting in student assessment

An extended summary of AMEE Medical Education Guide No 18

M Friedman Ben-David

Published in Medical Teacher (2000) 22, 2, pp 120-130

The full text of this guide comprises 18 pages and 25 references and is available from:

Association for Medical Education in Europe (AMEE), Tay Park House, 484 Perth Road, Dundee, DD2 1LR.

Tel: +44 (0) 1382 631953; Fax +44 (0) 1382 645748;

Email: amee@dundee.ac.uk www.amee.org

Guide Overview:

Setting standards for performance assessment is a relatively new area of study. The goal of this guide is to familiarise the reader with the framework, principles, key concepts and practical considerations of standard setting approaches and to enable the reader to make "educated" choices in selecting the most appropriate standard setting approach for their testing needs.

1 Why should we use standard setting procedures?

Licensure, credentialing and academic institutions are seeking new innovative approaches to the assessment of professional competence. Central to these recent initiatives is the need to determine standards of performance, which separate the competent from the non-competent candidate. Test developers need an educational tool by which they determine the cut-off point on the scoring scale which separates the non-competent from the competent.

2 Key Concepts

Norm-referenced vs Criterion-referenced standards

In a norm-referenced orientation, the standard is based on performance of an external large representative sample (norm group) equivalent to the candidates taking the test. The norm-referenced approach employing a group referenced standard, may result in reasonable standards providing the group is representative of the candidates' population, heterogeneous and large.

At the school level, a *relative* standard can be set at the mean performances of the candidates, or by defining the units of standard deviation from the mean. These standards may vary from year to year due to shifts in the ability of the group and may result in a fixed annual percentage of failing students. The criterion reference orientation links the standard to the content of the competence level under consideration. A standard is defined as absolute if it can be stated in terms of the knowledge and skills a student must possess in order to pass the course. An absolute (criterion) standard stays the same over multiple

administrations relative to the content specifications of the test. The failure rate may vary due to changes in the group's ability, from one administration to the other.

Compensatory vs Conjunctive Standards

In a *compensatory standard*, the standard is set on the total test score. Thus candidates can compensate for relatively poor performance in some parts of the examination with good performance in others. In a *conjunctive standard*, standards are set for individual components of the examination and candidates cannot compensate for relatively poor performance in one part. Reliability of test components may be a problem in conjunctive standards. In an OSCE, for example, while a single station may be considered unreliable, a group of stations assessing similar competences may be acceptable. Conjunctive standard allows diagnostic feedback to candidates, since each skill component is considered separately. The higher the correlation among the test components the greater the inclinations towards a compensatory standard. Highly correlated test components may imply a common construct or dimension of performance, in which performance on the one component impacts on performance on the other. The aggregation of related component scores into one dimension score is the basis for a compensatory pass/fail standard.

3 Standard setting methods

Test-centred models

In test-centred models the judges set standards by reviewing the test items and provide judgements as to the "just adequate" level of performance on these items.

The Angoff model employs a test-centred approach, in which a group of expert judges make estimates about how candidates would perform on items in the examination. This is described further in a later section.

Ebel's approach requests judges to categorise the items in a test employing a number of categories according to levels of difficulty and levels of relevance to the decision to be made. After classifying the items into each category, judges then decide on the proportion of items in each category that a hypothetical group of examinees could respond to correctly.

The Nedelsky approach was originally designed for multiple choice items. For each item, the judges decide on how many of the distractors (response options), a minimally competent examinee would recognise as being incorrect.

Jaeger's method emphasises the importance of recognising the need to sample all populations that have a legitimate interest in the outcomes of competency testing. The focus in Jaeger's method is on the passing examinees rather than on the borderline or the minimally competent.

Examinee-centred models

In the Borderline-Group method the judges identify an actual (not hypothetical) borderline group. The median score for this group is used as the passing score. In the Contrasts by Group approach, the panellists sort the examinees into two

groups: competent and not competent. The judgement is based on characteristics of the examinees relative to the task other than the test scores (i.e., the test scores are not known to the panellist during the sorting process). After the sorting is completed, the score distributions for the competent and not competent groups are plotted. Commonly, the point of intersection of the two distributions could be considered as the passing score.

The Hofstee method is a standard setting approach that incorporates the advantages of both relative and absolute standard setting procedures.

Modified Angoff

The Angoff standard setting approach is a judgemental approach in which a group of expert judges makes estimates about how borderline candidates would perform on items in the examination, i.e. the proportion of borderline examinees who will answer an item correctly. Estimates are averaged over judges and summed over items to create a standard (cut off score).

The panellists are asked to make judgements about that borderline candidate's likelihood to respond correctly to each of the test items. In general, judges have a tendency to produce high standards. An example of one is described in the guide.

Selection of panellists

The selection of panellists in standard setting is of the greatest importance. In summary, panellists should be:

- Experts in the related field of examination
- Familiar with the examination methods
- Good problem solvers
- Familiar with level of candidates
- Interested in education (teachers).

Written vs performance standards

Most of the standard setting procedures employed for performance assessment were first applied to written tests. The difference between written MCQ tests and performance tests such as the OSCE need to be taken into consideration.

4 Educational benefits of standard setting

Faculty Development

Standard setting procedures can be employed as a form of faculty development. Faculty experience first hand information of candidates' performance on the task and are able to compare this with their own expectations relating to the competence. The performance of poor and excellent candidates can be compared to their expectations.

Quality control of test materials

The process of exposing faculty to test materials, scoring policy, and profiles of scored performance, constitutes a scrutinised quality control procedure.

Panellists in the process of reviewing test materials identify in appropriate items, which are either ambiguous, or irrelevant.

5 Practical steps of the Modified Angoff Approach

Four steps can be identified:

Step 1 – General Orientation: The facilitator presents to panellists a general overview of the OSCE test, the purpose of the examination, the duration of each station, test components, scoring methods and any other information from which the panellists may benefit. A mini OSCE may be run as part of this orientation process.

Step 2 – Orientation to a "practice" station: Test developers present "practice" stations to panellists. The "practice" orientation materials may include:

1. A full descriptions of the stations including history and physical examination checklists;
2. Videotapes of one low performer and one high performer for the practice stations;
3. A blank checklist for the panellist for the two component skills while viewing the video. The actual skill score is presented to the panellists following the completion of each video performance.

Step 3 – Characteristics of borderline candidates: Examiners are asked to indicate their expectations for the performance of a hypothetical borderline group. Following a discussion a consensus is reached on the appropriate borderline characteristics per skill component.

Step 4 – Panellists provide ratings: Rating forms are distributed to panellists for each of the skills being assessed, eg, history taking, physical examination. On each form the stations are listed which contribute to the assessment of that competence. For each station the maximum number of points are noted. In the next column, for each station the panellists enter their individual judgements as to how many points will be answered correctly by a borderline examinee in order to pass the station. The panellists discuss their ratings. For the practice station the performance of a similar cohort of students in the past is presented to the panellists. This indicates the percentage of the students who might fail if the panellists' average ratings are applied to the distribution as a cut-off score. Panellists are then asked to make a second rating on the rating form, adjusting their rating in view of their peers' ratings and the actual performance data. A final cut-off score is calculated by averaging all the ratings. It is possible to divide, after the orientation, a large group of panellists (i.e. 18) into 3 groups of 6 each. The groups will set standards on different stations but one or two stations will be rated by all.

6 Evaluation

The standard setting process should be evaluated. Evaluation materials should include data on the first and second ratings of the panellists for each of the test components rated, which should demonstrate increased consensus of raters. It should also include a questionnaire administered to panellists at the end of the standard setting process.

7 Conclusion

Much work is still needed to establish effective standard setting procedures. The length of procedures should also be considered and ways to shorten the process are needed. Further consideration must be given to fully compensatory models in which test items or components are averaged to produce a test standard. Obtained standards should be checked against other information available on the test taker to ensure validity. Effective methods of training panellists to recognise borderline characteristics are essential if the Angoff approach is widely used. The more standard setting procedures are applied to a variety of tests, the more the practice of high quality testing will be enhanced and the higher will be the confidence in the testing of professional competencies.

© 2004 AMEE

The AMEE Guides series comprises 29 guides on key topics in medical education and is available from:

Association for Medical Education in Europe (AMEE), Tay Park House, 484 Perth Road, Dundee DD2 1LR, UK

*Tel: +44 (0)1382 631953; Fax: +44 (0)1382 645748; Email: amee@dundee.ac.uk
www.amee.org*