An introduction to standard setting methods in dentistry

J. Puryer*[1] and D. O'Sullivan[1]

IN BRIEF

- Outlines contemporary standard setting methods used in dental assessments.
- Will benefit all practitioners involved in undergraduate and postgraduate education.
- Increases readers' confidence to participate in standard setting of assessments.

EDUCATION

The aim of this paper is to give readers an overview of contemporary standard setting methods used within dental education, and to provide a better understanding of the subject. We hope that it will be of benefit not just to those in academic dentistry, but all practitioners involved with both undergraduate and postgraduate assessment.

INTRODUCTION

In 2012, the General Dental Council (GDC) produced their *Standards for education* document[1] which applies to all programmes leading to registration with the GDC. The standards are the regulatory tool used by the GDC to ensure that a programme is fit for purpose. These standards cover four areas that the GDC expects providers to meet, and one of these areas is that of student assessment. Requirement 23 of these standards states that: 'Assessment must be fair and undertaken against clear criteria. Standard setting must be employed for summative assessments'.

Part of the GDC's Quality Assurance process includes inspecting course providers and awarding bodies, and of the 11 programmes inspected by the GDC in 2012/2013, requirement 23 was 'fully met' by four, 'partly met' by five, and 'not met' by two of the programmes.[2] The GDC observed that 'Standard setting of assessments and examinations appears to happen at a very basic level in some circumstances and this is a key area which could be improved'.[2]

As UK dental schools widen their use of standard setting in order to comply with GDC standards, it is inevitable that more individuals will become involved with the process because of the number of assessments that will need to be standard set across the various programmes. In addition to full-time academic staff, it is foreseeable that part-time staff from a variety of backgrounds may become involved. In order for standard setting to be a reliable and valid process, staff involved in the process must be aware of the various methods that can be used, so that they can be chosen appropriately and applied correctly. It is intended that this paper will give an overview of standard setting methods so that participants will feel more confident to become involved in the process, and be able to implement the methods appropriately.

[1]School of Oral and Dental Sciences, Bristol Dental Hospital, Lower Maudlin Street, Bristol, BS1 2LY
*Correspondence to: Dr James Puryer
Tel: +44 (0)117 342 4184; Fax: +44 (0)117 342 4443
Email: james.puryer@bristol.ac.uk

Refereed Paper
Accepted 13 August 2015
DOI: 10.1038/sj.bdj.2015.755
©British Dental Journal 2015; 219: 355-358

WHAT IS STANDARD SETTING?

A standard, also known as the 'minimum pass level' separates the competent students from those who are not. It is a statement about whether an examination performance is good enough for a particular purpose. Determining the numerical answer to the question 'How much is enough?' is the process of standard setting.[3] It is the process of determining the minimal level of skill and knowledge required, and identifying an examination score that corresponds to it.[4] This standard should not be set in an arbitrary way, but it should be established through a specific methodology that considers the test's objectives and content areas, the examinees' performance, and the wider social or educational setting.[5]

Many methods have been developed and used to set standards for either written or clinical examinations.[6] Standards may be classified as either 'relative' (norm-referenced) or 'absolute' (criterion-referenced).[7,8]

Relative standards

Relative standards are expressed in terms of the performance of the cohort taking the assessment. Students will pass or fail depending upon how well they perform relative to other students taking the assessment. The following are examples of relative standards:

- The 10 students with the highest score will be awarded a distinction
- The bottom 25% of examinees will fail.

This type of standard is appropriate for assessments intended to select a certain number or percentage of students, such as tests for admission to establishments where only a certain number of places are available.

Absolute standards

Absolute standards are expressed in terms of the performance of students against the test material, and do NOT compare the performance of one student with others taking the test. Students will pass or fail depending on how well they perform, regardless of the performance of other candidates. Thus all candidates potentially could pass or fail. The following are examples of absolute standards:

- Students must identify 80% of dental instruments correctly
- Students must complete 75% of clinical techniques safely.

This type of standard is appropriate for assessments intended to determine whether or not students have the necessary knowledge or clinical skills for a particular purpose such as graduation from Dental School. Unless there are strong reasons to fail a certain number of students (for example, limited training posts available), absolute standards should be used rather than relative standards. This is particularly important when accrediting dental students as qualified 'safe practitioners' as it shows that they have reached either a certain level of skills competency or have acquired an agreed level of knowledge. In order to have an absolute standard, one or more standard setting techniques should be used.

STANDARD SETTING TECHNIQUES

Any standard stetting technique should be:
- Defensible (against legal challenges)
- Credible (the method is easy to explain and implement)
- Supported by evidence in literature
- Feasible (depending upon staff resources)
- Acceptable to all stakeholders.

Techniques for absolute standard setting fall into two categories. They can be 'test-centred' (where judgements are made about individual test items), or they can be 'person-centred' (where judgements are made about individual students). In test-centred methods (such as Angoff, Ebel), a group of expert judges make estimates about how they perceive candidates would perform on items in the examination. They look at deciding which mark would be a suitable cut-off for a minimally passing or just competent student. In person-centred methods (such as Borderline regression), the judges identify an actual (not hypothetical) borderline group, and it is the median numerical score achieved by these students that is used as the passing score. The Hofstee method is an example of a standard setting approach that incorporates aspects of both relative and absolute standard setting methods, and such methods are sometimes called compromised techniques.

Angoff

This standard setting method involves a group of expert judges making estimates about how borderline candidates would perform on items in the examination. The panel members are asked to make judgements about a borderline candidate's likelihood to respond correctly to each of the test items. Estimates are then averaged and summed over items to create a standard (cut off score). Table 1 shows an example of results obtained from a panel of five judges for a 10-station OSCE which derives a final cut score of 53% for the assessment.

Table 1 An example of the results of panel judges for a 10-station OSCE

	Judge 1	Judge 2	Judge 3	Judge 4	Judge 5	Average
Station 1	35%	40%	40%	45%	40%	40%
Station 2	60%	60%	55%	50%	45%	54%
Station 3	60%	40%	50%	55%	45%	50%
Station 4	55%	60%	60%	45%	40%	52%
Station 5	70%	75%	80%	70%	80%	75%
Station 6	40%	40%	45%	30%	40%	39%
Station 7	55%	60%	50%	45%	45%	51%
Station 8	45%	45%	50%	40%	50%	46%
Station 9	60%	60%	70%	50%	60%	60%
Station 10	70%	70%	50%	60%	65%	63%
					Cut score	53%

Ebel

This standard setting method looks at both the difficulty and also the relevance/importance of each question. The process shares some similarities with the modified Angoff method, although with this method the judges read each question item and assign it two variables. Firstly, whether the item is judged to be 'easy, medium or difficult', and secondly whether the knowledge is deemed to be 'essential or non-essential'. Judgements are made about the percentages of items in each category that borderline candidates would have answered correctly. These judgements are recorded in a table, an example of which is shown in Table 2. These percentages are then combined with the number of items in the assessment that have been assigned that particular variable. Table 3 shows an example of a combined application of Ebel's method to a 100-item test which derives a standard (cut score) of 75 out of 100.

Table 2 An example of an Ebel table

	Easy	Medium	Difficult
Essential	95%	80%	70%
Non-essential	70%	50%	30%

Table 3 Application of Ebel's method to a 100-item test

Category	Average proportion correct	Number of questions	Expected score
Essential			
Easy	95%	20	19
Medium	80%	40	32
Difficult	70%	10	7
Non-essential			
Easy	70%	15	10.5
Medium	50%	10	5
Difficult	40%	5	1.5
Standard (cut score)			75

Borderline regression

This standard setting method has gained favour in recent years[9] in both medical and dental education due to its advantages in OSCE assessments. In this method, examiners are asked to complete the mark sheet for a candidate sitting an individual station (which may have previously been standard set using an alternative technique, such as Angoff). They are then asked to award a global score based upon their subjective opinion as to how well that candidate performed at that station overall. The global score should not be based on the numerical marks accrued for that station. The candidate is given a global score usually out of a choice of 3–5 grade descriptors, such as:
- Good pass
- Pass
- Borderline
- Fail.

The borderline grade reflects those candidates whom the examiner feels have not performed well enough to have passed the station, but equally not performed so poorly that they deserve to fail that station. Following the assessment, candidate's mark sheet scores are collated. The global scores are also collated and are then statistically

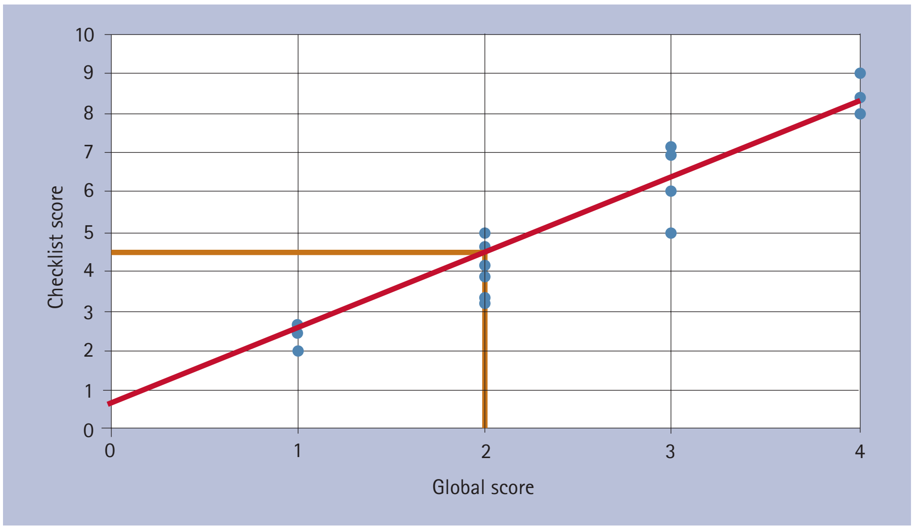

Fig. 1 An example of a Borderline Regression graph

regressed against the checklist scores. This process will then derive the cut or passing score for that particular OSCE station. Figure 1 shows an example of a completed borderline regression graph for a single OSCE station that has a maximum marksheet score of 10. The global score of '2' corresponds to those candidates that were deemed 'borderline' by the OSCE assessor. Borderline regression gives a cut score of '4.5 out of 10' for this particular station. Examiners often tend to favour this method of standard setting as it is less time consuming and is based on actual observation, rather than on a hypothetical borderline candidate's performance. This method of standard setting can be applied to other forms of assessments, but it is most useful for OSCEs and can be used as a check for the standard assigned to a station *vs.* the observed performance of candidates at that station.

Areas of concern

Whilst each standard setting method has its own merits, they share some common challenging issues. Ignoring these concerns during the standard setting process may result in dispute regarding the credibility and defensibility of the method used.[5,10,11] Potential issues include:

1) The subjective nature of the standard setting

All of the absolute standard setting methods require 'judgement'.[12,13] In test-centred methods, standard setters are asked to estimate the probability that a borderline candidate would correctly answer test items, whilst with person-centred methods, examiners are required to observe and rate a student's performance. In both cases, subjective judgement is used[6] and this may be criticised. However, it is important to remember that no purely objective method for determining the cut score exists.[12] In addition, human judgement plays a fundamental role in every level of dental student assessment and not merely in the standard setting.[13]

2) The definition of a borderline student

Crucial to each standard setting method is the definition of the 'borderline' student, although this concept is more obvious in some methods (for example, Angoff, in which standard setters are asked to envisage a borderline candidate and estimate their performance). It has been considered that the task of defining a borderline candidate is difficult, even for experts familiar with the students being assessed, to a degree that may impair their judgement. It has also been noted that judges may be tempted to envisage an 'average' student instead of the 'borderline' student, leading to the substitution of a criterion-based concept with a norm-referenced one.[12]

3) The choice of standard setting method

There appears to be little current consensus as to the best method of standard setting to use.[7,14] Whichever method is chosen, it should:

- Be closely aligned with the goal of assessment
- Require thoughtful effort of those participating in the process
- Based on research
- Easy to explain to participants
- Easy to implement.

The rationale for choosing a specific method is supported when evidence of defensible process is followed. Thus careful documentation of the whole process, including the number and background of judges, as well as collecting comments from judges and stakeholders should be considered. An assessment that is appropriately standard set may make the pass/fail decision defensible, but there is no conclusive way to ensure the validity of any standard setting method, and so relying on procedural evidence alone, provides weak justification for the credibility of decisions.[15]

HOW TO IMPROVE THE STANDARD SETTING PROCESS

The above issues are potentially overcome by use of the following good practices. Some of these should be followed prior to the standard setting process, and some take place afterwards. Most of these can be applied to all methods of standard setting.

Select appropriate judges

The number of judges used and their characteristics are key to the validity of the standard setting process. Their different educational backgrounds, professional role, familiarity with the students and the curriculum, experience and opinions will all have an impact on their cut score selected for each question item.[10,11,14] For the Angoff and Ebel methods, the involvement of an appropriate number and mixture of judges to allow a variety of viewpoints and to generate acceptable results is vital to the process.[11,12] There is still no consensus as to the exact number of judges needed, and although previous studies have suggested a range of 5–20, most authors suggest that a group of 10 is an appropriate number.[16–18] Judges should also be knowledgeable of the curriculum that is being assessed, the abilities of the student cohort and should be ideally selected with a balanced mix of age, gender, educational experience and subject experience.

Define the characteristics of a borderline student

Whichever method of standard setting is selected, stakeholders should decide upon student's performance levels that would be consistent with a 'just passing' student.[12,13,19] Descriptors of criteria relating to the minimally accepted level of competency should be available to judges. These descriptors will help to eliminate the issue of judges with 'extreme' views as to the acceptable standard, and which could otherwise influence the results. Methods for dealing with 'outliers' have been suggested[7] such as removing these outlying judgements or using the median instead of the mean. The removal of judgements should be a last resort since it undermines the credibility of the process and the selection of standard setters.

Train judges

It is essential that judges receive appropriate training on the method of standard setting selected (as well as descriptors of performance levels of the 'just passing' student) and they are then given practice at standard setting. This training will also help to fulfil requirement 21 of *Standards for education*[1] which states that 'Examiners/assessors must have appropriate skills, experience and training to undertake the task of assessment'; and also requirement 26 which states that 'The standard expected of students in each area to be assessed must be clear and students and staff involved in assessment must be aware of this standard'.

Determine reliability of assessment

It is important to determine if the results obtained would be the same if the standard setting method was repeated with either different or more judges. This reliability can be calculated using either Classical Test theory or Generalisability theory. It should be noted, however, that the reliability of an assessment does not guarantee the appropriateness of the assessment for a given purpose.

CONCLUSION

Standard setting is an important aspect of both undergraduate and postgraduate dental education and assessment, and this will become even more important as teaching establishments seek to ensure that their assessments are valid against possible legal challenges from students, and that they are also meeting the requirements of the GDC *Standards for education.* We hope that this introduction to standard setting will give readers a basic overview some of the methods used, and enable more clinicians to feel confident to become involved with this essential aspect of assessment.

1. The General Dental Council. *Standards for education.* London: GDC, 2012.
2. The General Dental Council. *Annual review of education.* London: GDC 2013.
3. Cusimano M. Standard setting in medical education. *Acad Med* 1996; **71:** 112–120.
4. Kane M. Validating interpretative arguments for licensure and certification examinations. *Eval Health Prof* 1994; **17:** 133–159.
5. Ricker K. Setting cut-scores: *a critical review of the Angoff and modified-Angoff methods. Alberta J Educ Res* 2006; **52:** 53–64.
6. Cizek G, Bunch M. *Standard setting: a guide to establishing and evaluating performance standards for tests.* Thousand Oaks, CA: Sage, 2007.
7. Livingston S, Zieky M. *Passing scores: a manual for setting standards of performance on educational and occupational tests.* Princeton, NJ: Educational Testing Service, 1982.
8. Ben-David M. AMEE guide No.18: Standard setting in student assessment. *Med Teacher* 2000; **22:** 120–130.
9. Schoonheim-Klein M, Muijtjens A, Habets L, Manogue M, van der Vleuten C, van der Velden U. Who will pass the dental OSCE? Comparison of the Angoff and Borderline Regression standard setting methods. *Eur J Dent Educ* 2009; **13:** 162–171.
10. Barman A. Standard setting in student assessment: is a defensible method yet to come? *Ann Acad Med Singapore* 2008; **37:** 957–963.
11. Norcini J. Setting standards on educational tests. *Med Educ* 2003; **37:** 464–469.
12. Zeiky M, Perie M. *A primer on setting cut scores on tests of educational achievement.* Princeton, NJ: Educational Testing Service, 2006.
13. Nichols P, Twing, J, Mueller C, O'Malley K. Standard-setting methods as measurement processes. *Educational Measurement: Issues and Practice* 2010; **29:** 14–24.
14. McKinley D, Norcini J. AMEE Guide No. 85: How to set standards on performance-based examinations. *Med Teacher* 2014; **36**: 97–110.
15. Hejri S, Jalili M. Standard setting in medical education; fundamental concepts and emerging challenges. *Med J Islam Repub Iran* 2014; **28:** 34.
16. Brennan R, Lockwood R. A comparison of the Nedelsky and Angoff cutting score procedures using Generalisability Theory. *Applied Psychological Measurement* 1980; **4:** 219–240.
17. Hurtz G, Hertz N. How many raters should be used for establishing cut-off scores with the Angoff method? A Generalizability theory study. *Educational and Psychological Measurement* 1999; **59:** 885–897.
18. Fowell S, Fewtrell R, McLaughlin P. Estimating the minimum number of judges required for test-centred standard setting on written assessments. Do discussion and iteration have an influence? *Adv Health Sci Educ Theory Pract* 2008; **13:** 11–24.
19. Chinn R. *Considerations in Setting Cut Scores.* Lexington, Kentucky: Council on Licensure, Enforcement, and Regulation, Resource Brief 2006.