Standard Setting in Student Assessment: Is a Defensible Method Yet to Come?

A Barman,[1] *MBBS (Bangladesh), DHE (India), MMed (UK)*

Abstract

Introduction: Setting, maintaining and re-evaluation of assessment standard periodically are important issues in medical education. The cut-off scores are often "pulled from the air" or set to an arbitrary percentage. A large number of methods/procedures used to set standard or cut score are described in literature. There is a high degree of uncertainty in performance standard set by using these methods. Standards set using the existing methods reflect the subjective judgment of the standard setters. This review is not to describe the existing standard setting methods/procedures but to narrate the validity, reliability, feasibility and legal issues relating to standard setting. Materials and Methods: This review is on some of the issues in standard setting based on the published articles of educational assessment researchers. Results: Standard or cut-off score should be to determine whether the examinee attained the requirement to be certified competent. There is no perfect method to determine cut score on a test and none is agreed upon as the best method. Setting standard is not an exact science. Legitimacy of the standard is supported when performance standard is linked to the requirement of practice. Test-curriculum alignment and content validity are important for most educational test validity arguments. Conclusion: Representative percentage of must-know learning objectives in the curriculum may be the basis of test items and pass/fail marks. Practice analysis may help in identifying the must-know areas of curriculum. Cut score set by this procedure may give the credibility, validity, defensibility and comparability of the standard. Constructing the test items by subject experts and vetted by multi-disciplinary faculty members may ensure the reliability of the test as well as the standard.

Ann Acad Med Singapore 2008;37:957-63

Key words: Difficulty and discriminating indices, Judges and judgment, Legal issues, Practicability, Reliability, Validity

Introduction

To validate any "adjective", be it for living or non-living, a criteria or standard is needed. Globalisation, mobility of doctors and the rising number of medical institutions make it imperative to have comparable standards in medical teaching learning and assessment.[1] Setting, maintaining and re-evaluation of assessment standards periodically is an important issue in medical education.[2]

Standard is the conceptual version, while the passing score is the operational version of the desired level of competence. Performance standard is a construct, while passing score is a number.[3] Standard setting is a policy-making activity and cut-score setting is operationalisation of the policy.[4] These are interrelated. The standard of performance can be transformed to cut score which is the score or number in the distribution of score obtained by the examinees and divides the series of score into 2 or more mutually exclusive categories.[5]

Standard for licensure and certification tests are to ensure that the candidates are prepared to handle the problems they will encounter in practice.[6] Examination standard setters are facing the problem of identifying the level of performance that reflects acceptable proficiency.[7-9] It would be appropriate if the cut score could be set at 100%, but there is likely to have few licensures and that may cause serious harm to the public services.[3] At the same time it cannot be compromised with the quality of the licensures (Fig. 1).

Assessment standards are relative/norm-referenced and absolute/criterion-referenced. George et al[10] found a significant difference in the outcome when the students were assessed by norm-referenced and criterion-referenced

[1] Department of Medical Education,School of Medical Sciences, Universiti Sains Malaysia, Malaysia

Address for Correspondence: Dr Arunodaya Barman, Associate Professor, Department of Medical Education, School of Medical Sciences, Universiti Sains Malaysia, 16150 Kubang Kerian, Kota Bharu, Kelantan, Malaysia.

Email: barman@kb.usm.my

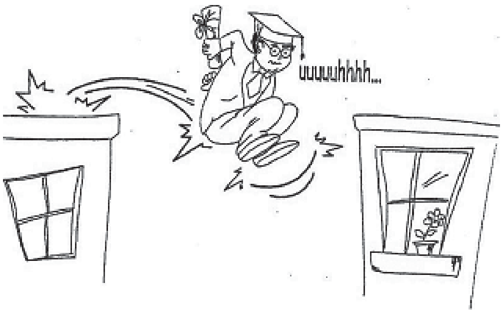

Fig. 1. Fate of a 75% achiever

tests. Relative standard is used to identify a certain number of best examinees from the group.[11] Norm-referenced system is easy to administer but in many respects, it fails to fulfil the authenticity of assessment.[12] Norm-referenced standards do not indicate the competency relating to the job.[13] However, Hoover[14] claimed that it tells much more about the achievement of students than the criterion-referenced test. Most examining bodies use norm-referenced system in reporting students score.[15] When the ultimate objective of medical education is to produce competent physicians, the assessment should be to differentiate the competent from non-competent and not rank order the individual prospective physicians in the group.[16,17] Criterion-referenced assessment assess students against specified standards of achievement[18] and differentiate the competent from the non-competent.[19] The criterion standard tells about the degree of competence of a particular examinee independent of the performance of other members in the group.[11,20] The test is in line with the goals of the taught subject and against the pre-specified criteria and standard.[21] Most high-stakes competency assessments seek criterion-referenced/absolute standard for pass/fail decision.[22] The criterion-referenced test is constructed in such a way that it gives measurement directly interpretable in terms of a specified performance standard.[23] Criterion-referenced test carries performance standard within it. Criterion is synonymous with standard or cut-off.[24]

In reality, examinees show a variety of performances ranging from non-competent to reasonably competent. To be fully competent, scoring 100% of marks is unrealistic in most examinations. Hence, most institutions accept the traditional cut-off scores of 50% to 70% as mastery in the subject.[25] Hambleton[26] stated, “It is well-known that cut-off scores are often “pulled from the air” or set to (say) 80% because that is the value other school districts are using”.

Setting standards is important but it is tiresome.[27] The method should yield cut-off score or decision validity evidence that permits dichotomous categorisation, is sensitive to the examinee’s performance, instruction or training, statistically sound and practicable, credible, easy to implement and interpret for the layman.[5] There were 38 standard setting methods/procedures until 1986[5] and nearly 50 until 1996.[28] It is not certain, for a particular situation, which one of the existing standard setting methods is appropriate.[5] This review is not to describe any of the existing standard setting methods/procedures but to narrate some of the issues relating to standard setting.

Methods

The literature search was performed using PubMed and EbscoHost databases. Key words used in searching include cut-off score, standard setting and pass/fail marks. No time limit was specified in searching. Abstracts of the searched articles were read through for relevance. Reliability, validity, feasibility, judges and judgment in setting standard and legal aspects of assessment were the key issues of inclusion criteria for in-depth study of the full articles. Some of the articles searched were from the reference lists of the articles of primary search.

Results

Standard/Cut-score Setting Methods

The absolute standard/cut score can be used to determine whether the examinee attained the requirement to be certified competent.[22] A number of methods for standard setting are described in literatures. All of them are judgmental with subjectivity and imprecision.[28] There is no perfect method to determine cut score on a test[3] and none is agreed upon as the best method.[29,30] All these methods have their advantages and disadvantages depending on the specific application.[11] There is no scientific way of choosing a standard setting method from the group.[31]

Though different standard setting methods are recommended for different nature and format of tests, they often produce different results.[3,5,7,8,26,31-41] Even when the same method is used on a test by different groups of panelists, resulting cut score may have great variations.[30,42,43] Thorough implementation of the best available procedure does not guarantee the appropriateness of the passing score.[3] Several standard setting procedures have been developed to face the challenge of making it defensible and controllable.[44] Significant advancement in standard setting technology has allowed reasonable confidence in the resultant standard.[45] Nevertheless, the cut score is not consistent among the methods. Cut scores obtained by using Angoff methods with and without reality check are not comparable to those obtained by using borderline regression method.[36] Moreover, passing scores ascertained by using these methods tended to be high for the group of students when they are more able.[38]

Angoff and Nedelsky methods of standard setting are commonly used methods for performance assessment and certification.[7,8,15,46-49] The Angoff method is most preferred as it provides reasonable standard.[50] Hobma et al[51] in their study obtained a credible standard by using Angoff and borderline regression methods.

The standard set by any method should be comparable to different groups of students undergoing similar educational programmes. Failing to assure comparable pass/fail decision makes the certificate indistinguishable. Though there is no national licensing examination for the graduates of different medical schools in the United Kingdom (UK), all the medical graduates are granted the same practising license by the General Medical Council (GMC). It means that all the graduates achieved the minimum standard. Using borderline group method or borderline regression method and the Angoff method, it was found that the undergraduate medical students' pass marks were inconsistent within 3 medical schools in the UK.[52]

The current 60% cut score of the Korean National Medical Licensing Examination is arbitrary and there is a possibility of the unfortunate failure of the competent and the fortunate pass of the non-competent depending on test difficulty. They suggested Bookmark and the modified Angoff methods with test equating as alternatives to the current system. Application of the test equating method may eliminate this problem, but it is not practically possible due to the drainage of test items. This causes a lack of comparability of cut score across different years.[53]

Judges and Judgment in Standard Setting

Setting standard is not an exact science.[54] Estimate of the cut score (pass/fail marks) is an arbitrary decision that comes from a subjective judgment of experts.[3,22,24,39,44,55,56] One of the sources of arbitrariness in standard setting is intrajudge inconsistency.[56] Some of the researchers[26,32,57,58] claim that it is arbitrary but not capricious, that is, it is not selected randomly without reasons. Glass[24] suggested that we should not, in most cases, set a standard because of the arbitrariness of most of the methods. It is unrealistic and unreasonable for a physician to take much time for setting the standard that is arbitrary anyway.[59]

Judgment is the most important factor in setting the standard and the judgment depends on the judges' knowledge, experiences and competency relating to the subject matter and standard-setting procedures. The judges vary widely in their knowledge of competency and proposing a reasonable and defensible test standard.[60] Stern et al[61] in their study found that the panelists within their group set similar standards. Though the primary act of standard setting is the creation of an absolute standard not bearing in mind examinees' normative test performance, studies recommended the provision of test performance data to the judges for modification of initial cut score and reduction of variation.[26,62] Judges tended to adjust their initial cut score during iterative standard setting exercise. In such a situation, there are ample possibilities of group-induced polarisation in standard setting.[63]

To generate an acceptable standard/cut score, the use of more than 1 type of judge was recommended[26,62] but no significant relationship was observed between the types of judges and test standard.[60]

Despite training, panelists were confused about their tasks and perceived differently what constitutes minimal competence,[64] which is the key issue in most of the cut-score setting methods. The judges often found it difficult to apply the concept of minimal competence.[49,65-67] There is no empirical definition of the borderline or minimally competent examinee.[68] Glass[69] concluded, "The idea of minimal competence is bad logic and even worse psychology". Berk[28] stated that item-judgment methods of standard setting are fundamentally flawed as judges require the nearly impossible task of estimating the probability of correctly answering test items by hypothetical borderline examinees. Panelists' familiarity with examinees' expected level of abilities are questionable.[70] For that, NAE[71] recommended discontinuation of any item judgment methods.

Cut score may be influenced by the cognitive, social, political and emotional status of the panelists. Panelists often find themselves in a tug of war. On one side, they want to set a high standard and cut score; on the other side, they were concerned about the high failure rate which points fingers towards the performance of teachers.[64] Panelists during the process of standard setting felt pressured to recommend a passing score that is acceptable to the school.[72] In such a situation, mathematical analyses of a good panelist are:

Difficulty level of question determined by judges – Difficulty level calculated on examinee's performance = 0 (Judges are efficient)

Difficulty level of question determined by judges – Difficulty level calculated on examinee's performance ≠ 0 (Judges are inefficient)

Validity of Standard/Cut score

High-stake tests results must be reasonably precise, reliable and valid.[73] Once a cut score is set, evidence must be produced on its quality, soundness or defensibility, which is difficult. Ben-David[25] on describing a number of existing standard setting methods in medical education concluded that much research is still needed to establish effective standard setting procedures. Validity evidences of a standard should direct towards input, output, consequences and process itself to show that the cut score distinguishes the competent from the non-competent. One way is to compare the performance of the competent and the non-competent in their workplace.[74] Such evidences are usually impossible to obtain as only the declared competent based on the set cut score are engaged for their performance, not the non-competent. Though standard setters very often rely on procedural evidence, it provides weak support for the appropriateness of the standard.[3]

To support the appropriateness of standard and the consequential decisions, a defensible standard setting procedures is essential.[25] Legitimacy of the standard is supported when performance standard is linked to the requirement of practice[3,75,76] and the requirement can be identified by practice and content analysis.[6,13,77-80] Chesser et al[81] stated that medical students should not be declared to have passed if they cannot show competency in all areas of test.

Validity of the cut score depends upon its accuracy in separating examinees into mastery and non-mastery.[26] Cut score should be high to minimise false positive when the licensures will be in a task failing which will cause a serious effect on the individual or society taking the services.[35] Hence, generally the judges produce a very high standard.[82] The more we focus on raising test scores, the more instruction is distorted, and the less credible are the scores themselves.[83]

Certifying a practitioner or student as competent needs assessment and that should be documented, accounted for and defended.[44] Documentation includes independent evidence that the passing score is reasonable.[46] An appropriately set standard may make this certification defensible,[55] but there is no absolute criteria that can validate the set score.[77] There is no universally agreed upon decisive factor that can define the effectiveness of any standard-setting method.[84]

Reliability of Standard/Cut score

A frequently used technique to evaluate quality of standard is estimation of its reliability, but it does not guarantee the appropriateness of the cut score for the given purpose as reliability does not tell the meaning of the score. Moreover, the reliability co-efficient has a chance of being influenced by dominating judges exercising authority on the panelists or if panelists have the same misconception about the process. Sometimes during iterative exercises, panelists felt they were under pressure to reset their ratings for cut score to be consistent with others in the group.[64]

Koffler[33] concluded that no one standard setting procedure is relied upon to determine cut score and recommended using a number of procedures. Though there are a lot of refinements in standard-setting methods, unreliability still persists.[45]

Practicability of Using Standard Setting Procedures

The practicability of using standard-setting methods, especially those requiring complex statistical procedures and repeated review of test data, is low.[5] Resource-intensive procedures also have low feasibility of their usage. Standard failing to assure comparable pass/fail decision makes the certificate indistinguishable. It requires the use of robust equating designs and procedures in the adjustment of the standard over examinations.[85] In order to estimate a reproducible cut score using the Angoff method, substantial resources are required.[70] Contrasting group method is time-consuming and it is difficult to make unbiased judgments.[86]

It is difficult to sort the items into categories according to perceived difficulty and relevance, as required by the Ebel method. Panelists felt greater confidence in their knowledge estimates and standard set by them using the Angoff method than using Nedelsky and Jaeger procedures.[48] However, Verhoeven et al[70] using the Angoff method found it difficult to set the standard for a comprehensive test that is used in undergraduate medical education, while Schindler et al[22] successfully utilised the Hofstee method in multiple performance measures for clinical clerkship.

Difficulty and Discriminating Indices and Standard Setting

One of the modifications of the commonly used Angoff method is the addition of the difficulty level of items obtained from the actual performance of the examinees.[25] Cut score needs to be changed relating to the difficulty level of tests.[87] Setting pass/fail marks using the difficulty level estimated by panelists may be susceptible to judgmental biases.[88] Questions are difficult when the examinees do not know the answers or the questions are vague or highly difficult in structure, such as those containing uncommon words and advanced form of sentences. The difficulty index is not solely determined by the content of the item as it also reflects the ability of the examinees[89] and the instruction they have had.[90] For a well-prepared group of examinees, item difficulty indices may range from 70% to 100%.[89] Hence, the passing score is usually higher when the examinee group is more able.[38] When the difficulty index moves towards high or low from 50%, the discriminating index becomes low. A rigid content specification should be maintained in generating the items[89] and for that purpose, items with high difficulty indices may need to be accepted. In criterion-referenced measurement, many good items may have discrimination indices of zero.[89] For validity, a well-constructed test accepts items with low discriminating indices.[90] The so-called "assessment by ambush" is one aspect of unfair examination, where for high discrimination, potentially important areas are not tested.[91] Hence, the use of difficulty and discriminating indices in standard setting may not be appropriate when the objective of assessment is to differentiate the competent from the non-competent.

Legal Issues and Standard Setting

When tests are challenged in courts, the test setters used the standard to argue but the court does not consider standard as authoritative.[77] It is evident that, in case of legal argument, courts evaluate the curricular validity and the reliability of the test.[92,93] It is shown in studies that for judgment, courts identified the rational relationship between essential elements in the curriculum and the content tested,[92] and relationship of the material taught in the classroom and the test items.[93] Test-curriculum alignment and content validity are important for most educational test validity argument.[92,94] Content validity refers to the degree to which test content is congruent with the testing purposes.[26,95] The test items should have a good fit to the curriculum[96-98] especially the core of knowledge and understanding specific to curriculum.[34] Curriculum policy and assessment practices should be well balanced.[99] Assessment methodology should reflect the outcome as well as the way of learning.[3,100] Face and content validity measure how well the test items represent the domain of learning objectives. There is no statistics to establish the content validity.[101] The test is judged to have content validity as it is designed and evaluated by expert faculty.[93,102,103] Assessment must measure the competency deemed important and stated in policy documents.[104,105] Performance on the sample test items should provide a basis for estimating achievement in alignment to the learning objectives.[73] Sequential development, expert review and document analysis are the approaches to determine whether expectations and assessment are in alignment.[106]

Conclusion

Setting the standard is tiresome but it is important. It is imperative to identify the cut score to differentiate the competent from the non-competent. Percentage of must-know survival knowledge and skills in the curriculum may be the basis of test items and hence the standard/cut score for pass/fail decision. Test items on need-to-know and nice-to-know areas of curriculum may be added to make up the full marks of 100%, which will not be decisive of their basic competence of pass and fail; these may improve their grades if included. Practice analysis may help to identify the curricular content specially the must-know areas. Test items should reflect the representative samples of the learning objectives. When 70% of the learning objectives in the curriculum are of must-know category then it may be appropriate to set the pass/fail marks at 70%. This may give the credibility, validity and defensibility of the standard. This procedure may ensure the comparability of the graduates produced by medical schools and test administrators year after year. Constructing the test items by subject experts and vetted by multi-disciplinary faculty may ensure the reliability of the test as well as the standard.

Acknowledgements

The authors would like to thank Mohd Nazif Saadon, Graphic designer, USM for the cartoon.

REFERENCES

1. Lilley PM, Harden RM. Standards and medical education. Med Teach 2003;25:349-51.
2. Senanayake MP, Mettananda DS. Standards medical students set for themselves when preparing for the final MBBS examination. Ann Acad Med Singapore 2005;34:483-5.
3. Kane M. Validating the performance standards associated with passing score. Rev Educ Res 1994;64:425-61.
4. Ricker KL. Setting cut-scores: A critical review of the Angoff and modified Angoff methods. Alberta J Educ Res 2006;52:53-64.
5. Berk RA. A consumer’s guide to setting performance standard on criterion-referenced tests. Rev Educ Res 1986;56:137-72.
6. Kane M. Model-based practice analysis and test specifications. Appl Meas Educ 1997;10:5-18.
7. Andrew BJ, Hecht JT. A preliminary investigation of two procedures for setting examination standards. Educ Psychol Meas 1976;36:45-50.
8. Skakun EN, Kling S. Comparability of methods for setting standards. J Educ Meas1980;17:229-35.
9. Ebel RI. The case for minimum competency testing. Phi Delta Kappan 1978;59:546-9.
10. George S, Haque MS, Oyebode F. Standard setting: comparison of two methods. BMC Med Educ 2006;6:46.
11. Norcini JJ. Setting standards on educational tests. Med Educ 2003;37: 464-9.
12. Carlson T, Macdonald D, Gorely T, Hanrahan S, Burgess-Limerick R. Implementing criterion-referenced assessment within a multi-disciplinary university department. High Educ Res Dev 2000;19: 103-16.
13. Truxillo DM, Donahue LM, Sulzer JL. Setting cutoff scores for personnel selection tests: issues, illustrations, and recommendations. Hum Perform 1996;9:275-85.
14. Hoover HD. Some common misconceptions about tests and testing. Educ Meas 2003;22:5-14.
15. Supernaw RB, Mehvar R. Methodology for the assessment of competence and the definition of deficiencies of students in all levels of the curriculum. Am J Pharmaceut Educ 2002;66:1-4.
16. Turnbull JM. What is … normative versus criterion-referenced assessment. Med Teach 1989;11:145-50.
17. Black PJ. Formative and summative assessment by teachers. Stud Sci Educ 1993; 21:49-97.
18. Dunn L, Parry S, Morgan C. Seeking quality in criterion referenced assessment. Paper presented at the Learning Communities and assessment Cultures Conference organized by the EARLI Special Interest Group on Assessment and Evaluation, University of Northumbria, 28-30 August 2002. Available at: http://www.leeds.ac.uk/educol/documents/00002257.htm. Accessed 15 April 2007.
19. Harden RM. Ten questions to ask when planning a course or curriculum. Med Educ 1986;20:356-65.
20. Glaser R. Instructional technology and the measurement of learning outcomes: Some questions. Am Psychol 1963;18:519-21.
21. Gipps C. What do we mean by equity in relation to assessment? Assess Educ Princ Pol Pract 1995;2:271-81.
22. Schindler N, Corcoran J, DaRosa D. Description and impact of using a standard-setting method for determining pass/fail scores in a surgery

clerkship. Am J Surg. 2007;193:252-7.

23. Glaser R, Nitko AJ. Measurement in learning and instruction. In: Thorndike RL, editor. Educational Measurement. 2nd ed. Washington DC: Americal Council on Education, 1971:625-70.
24. Glass GV. Standard and criteria. J Educ Meas 1978;15:237-261.
25. Ben-David MF. Standard setting in student assessment. Med Teach 2000;22:120-30.
26. Hambleton RK. On the use of cut-off score with criterion-referenced tests in instructional settings. J Educ Meas 1978;15: 277-90.
27. Jolly B. Setting standard for tomorrow's doctors. Med Educ 1999;33: 792-3.
28. Berk RA. Standard setting: The next generation (Where few psychometricians have gone before!). Appl Meas Educ 1996;9:215-235.
29. Linn RL. Performance standards: Utility for different uses of assessments. Educ Pol Anal Arch 2003;11:1-20.
30. Boursicot KA, Roberts TE, Pell G. Standard setting for clinical competence at graduation from medical school: a comparison of passing scores across five medical schools. Adv Health Sci Educ Theory Pract 2006;11:173-83.
31. Halpin G, Halpin G. An analysis of the reliability and validity of procedures for setting minimum competency standards. Educ Psychol Meas 1987;47:977-83.
32. Popham W. As always provocative. J Educ Meas 1978;15:297-300.
33. Koffler SL. A comparison of approaches for setting proficiency standards. J Educ Meas 1980;17:167-78.
34. Angoff WH. Proposals for theoretical and applied development in measurement. Appl Meas Educ 1988;1:215-22.
35. Walter RA, Kapes JT. Development of a procedure for establishing occupational examination cut scores: A NOCTI example. J Ind Teach Educ 2003;40:25-45.
36. Kramer A, Muijtjens A, Jansen K, Düsman H, Tan L, van der Vleuten C. Comparison of a rational and an empirical standard setting procedure for an Objective Structured Clinical Examinations. Med Educ 2003;37:132-9.
37. Mills CN. A comparison of three methods of establishing cut-off scores on criterion-referenced tests. J Educ Meas 1983;20:283-92.
38. Livingston SA, Zieky MJ. A comparative study of standard-setting methods. Appl Meas Educ 1989;2:121-41.
39. Downing SM, Tekian A, Yudkowsky R. Procedures for establishing defensible absolute passing scores on performance examinations in health professions education. Teach Learn Med 2006;18: 50-7.
40. Humphrey-Murto S, MacFadyen JC. Standard setting: a comparison of case-author and modified borderline-group methods in a small-scale OSCE. Acad Med 2002;77:729-32.
41. Jaeger RM. Certification of student competence. In: Linn RL, editor. Educational Measurement. 3rd ed. New York: ACE Macmillan, 1989:485-514.
42. Cizek GJ. Adapting testing technology to serve accountability aims: The case of vertically moderated standard setting. Appl Meas Educ 2005;18:1-9.
43. Reckase MD. Rejoinder: Evaluating standard setting methods using error models proposed by Schulz. Educ Meas 2006;25:14-17.
44. Cusimano MD. Standard setting in medical education. Acad Med 1996;71(10 suppl):S112-20.
45. Cizek GJ. Reconsidering standards and criteria. J Educ Meas 1993;30: 93-106.
46. Cizek GJ. Standard setting guidelines. Educ Meas 1996;15:12-21.
47. Plake BS. Setting performance standards for professional licensure and certification. Appl Meas Educ 1998;11:650-80.
48. Cross LH, Impara JC, frary RB, Jaeger RM. A comparison of three methods for establishing minimum standards on the national teacher examinations. J Educ Meas 1984;21:113-29.
49. Chinn RN, Hertz NR. Alternative approaches to standard setting for licensing and certification examinations. Appl Meas Educ 2002;15: 1-14.
50. Jaeger RM. Establishing standards for teacher certification tests. Educ Meas 1990;9:15-20.
51. Hobma SO, Ram PM, Muijtjens AM, Grol RP, van der Vleuten CP. Setting a standard for performance assessment of doctor-patient communication in general practice. Med Educ 2004;38:1244-52.
52. Boursicot KA, Roberts TE Pell G. Using borderline methods to compare passing standards for OSCEs at graduation across three medical schools. Med Educ 2007;41:1024-31.
53. Ahn DS, Ahn S. Reconsidering the cut score of Korean National Medical Licensing examination. J Educ Eval Health Prof 2007;4:1.
54. Stern DT, Friedman Ben-David M, Norcini J, Wojtczak A, Schwarz MR. Setting school-level outcome standards. Med Educ 2006;40: 166-72.
55. Searle J. Defining competency – the role of standard setting. Med Educ 2000;34:363-6.
56. van der Linden WJ. A latent trait method for determining intrajudge inconsistency in the Angoff and Nedelsky techniques of standard setting. J Educ Meas 1982;19: 295-308.
57. Hambleton RK, Bernnan RL, Brown W, Dodd B, Forsyth RA, Mehrens WA, et al. A response to "Setting reasonable and useful standards" in the National Academy of Sciences' "Grading the Nation's Report Card". Educ Meas 2000;19:5-14.
58. Linn R. Demands, cautions and suggestions for setting standards. J Educ Meas 1978;15:301-8.
59. Norcini J, Guille R. Combining tests and setting standards. In: Norman GR, van der Vleuten CPM, Newble DI, editors. International Handbook of Research in Medical Education. Boston: Kluwer Academic Publishers, 2002:811-34.
60. Busch JC, Jaeger RM. Influence of type of judge, normative information, and discussion on standard recommended for the national teacher examinations. J Educ Meas 1990;27:145-63.
61. Stern DT, Ben-David MF, De Champlain A, Hodges B, Wojtczak A, Schwarz MR. Ensuring global standards for medical graduates: a pilot study of international standard-setting. Med Teach 2005;27:207-13.
62. Morrison H, McNally H, Wylie C, McFaul P, Thompson W. The passing score in the objective structured clinical examination. Med Educ 1996;30:345-8.
63. Fitzpatrick AR. Social influences in standard setting: The effects of social interaction on group judgments. Rev Educ Res 1989;59:315-28.
64. McGinty D. Illuminating the "Black Box" of standard setting: An exploratory qualitative study. Appl Meas Educ 2005;18:269-87.
65. Boursicot K, Roberts T. Setting standards in a professional higher education course: defining the concept of the minimally competent student in performance-based assessment at the level of graduation from medical school. High Educ Q 2006;60:74-90.
66. Boulet JR, De Champlain AF, McKinley DW. Setting defensible performance standards on OSCEs and standardized patient examinations. Med Teach 2003;25:245-9.
67. Wayne DB, Fudala MJ, Butter J, Siddall VJ, Feinglass J, Wade LD, et al. Comparison of two standard-setting methods for advanced cardiac life support training. Acad Med 2005;80(10 Suppl):S63-6.
68. Goodwin LD. Relations between observed item difficulty levels and Angoff minimum passing levels for a group of borderline examinees. Appl Meas Educ 1999;12:13-28.
69. Glass GV. 2003. Standards and criteria redux. Available at: http://glass.ed.asu.edu/gene/papers/standards/. Accessed 5 June 2007.
70. Verhoeven BH, van der Steeg AF, Scherpbier AJ, Muijtjens AM, Verwijnen GM, van der Vleuten CP. Reliability and credibility of an Angoff standard setting procedure in progress testing using recent graduates as judges. Med Educ 1999;33:832-7.
71. Shepard L, Glaser R., Linn R, Bohrnstedt G. Setting Performance Standards for Student Achievement. Stanford, CA: National Academy of Education, 1993.
72. Giraud G, Impara JC. Making the cut: The cut score setting process in a public school district. Appl Meas Educ 2005;18:289-312.

73. Case SM, Swanson DV. Constructing written test questions in the basic sciences. 3rd ed. Philadelphia, PA: National Board of Medical Examiners, 3750 Market Street, PA 19104, 2002:8-20. Available at: http://www.nbme.org/PDF/ItemWriting_2003/2003IWGwhole.pdf. Accessed 14 August 2007.
74. Kane MT. The validity of licensure examinations. Am Psychol 1982;37:911-8.
75. Raymond MR. A practical guide to practice analysis for credentialing examinations. Educ Meas 2002;21:25-36.
76. Levin HM. Educational performance standards: Image or substance? J Educ Meas 1978;15: 309-19.
77. Sireci SG, Green PC. Legal and psychometric criteria for evaluating teacher certification tests. Educ Meas 2000;19:22-34.
78. Mauer TJ, Alexander RA. Method of improving employment test critical scores derived by judging test content: A review and critique. Person Psychol 1992;45:727-62.
79. Crocker L. Assessing content representativeness of performance assessment exercises. Appl Meas Educ 1997;10: 83-95.
80. Wang N, Schnipke D, Witt EA. Use of knowledge, skill, and ability statements in developing licensure and certification examinations. Educ Meas 2005;24:15-22.
81. Chesser AM, Laing MR, Miedzybrodzka ZH, Brittenden J, Heys SD. Factor analysis can be a useful standard setting tool in a high stakes OSCE assessment. Med Educ 2004;38:825-31.
82. Wolf A, Silver R. Work Based Learning: Trainee Assessment by Supervisors. Research and Development Series Number 33 Sheffield: MSC, 1986.
83. Shepard LA. Why we need better assessments. Educ Leader 1989;46: 4-9.
84. Berk RA. Something old, something new, something borrowed, a lot to do! Appl Meas Educ 1995;8:99-109.
85. Norcini JJ, Shea JA. The credibility and comparability of standards. Appl Meas Educ 1997;10:39-59.
86. Pell G, Roberts TE. Setting standard for students assessment. Int J Res Meth Educ 2006;29:91-103.
87. Bramley T. Accessibility, easiness and standards. Educ Res 2005;47: 251-61.
88. Thorsteinson TJ. Framing effects on the setting of critical scores for content valid tests. Hum Perform 2006;19:201-17.
89. Ebel RL, Frisbie DA. Essentials of Educational Measurement. 5th ed. New Jersey: Prentice-Hall, 1991.
90. Linn RL, Gronlund NE. Measurement and Assessment in Teaching. 8th ed. New Jersey: Prentice-Hall, 2000.
91. Brown B. Trends in assessment. In: Harden R, Hart I, Mulholland H, editors. Approaches to the Assessment of Clinical Competence. Vol. 1. Dundee Centre for Medical Education, 1992.
92. Sireci SG, Parker P. Validity on trial: Psychometric and legal conceptualizations of validity. Educ Meas 2006;25:27-34.
93. Phillips SE. GI Forum V. Texas Education Agency: Psychometric evidence. Appl Meas Educ 2000;13:343-85.
94. Mehrens WA, Popham WJ. How to evaluate the legal defensibility of high-stake tests. Appl Meas Educ 1992;5:265-83.
95. Sireci SG. The construct of the content validity. Soc Indicat Res 1998;45:83-117.
96. Crocker LM, Miller MD, Franks EA. Quantitative methods for assessing the fit between test and curriculum. Appl Meas Educ 1989;2:179-94.
97. Cole NS, Nitko AJ. Measuring program effects. In: Berk RA, editor. Educational Evaluation Methodology: The State of the Art. Baltimore: Johns Hopkins University Press, 1981:32-63.
98. Wiggins G. Teaching to the (Authentic) test. Educ Leader 1989;46: 41-7.
99. Monyooe LA. On shifting sands: Exploring the curriculum and assessment dichotomy. Int J Instructional Technol Distance Learning 2004:1:11-26.
100. La Marca P, Redfield D, Winter P, Bailey A, Despriet L. State Standards and State Assessment Systems: A Guide to Alignment. Series on Standards and Assessments. Washington, DC: Council of Chief State School Officers, 2000.
101. Newble D. Assessment. In: Jolly B, Rees L, editors. Medical Education in the Millennium. 1st ed. UK: Oxford University Press, 1998:129-42.
102. Brown B, Roberts J, Rankin J, Stevens B, Tompkins C, Patton D. The objective structured clinical examination: reliability and validity. In: Hart IR, Harden RM, Walton HJ, editors. Further Developments in Assessing Clinical Competence. International Conference Proceedings; Ottawa, Canada, 1987:563-71.
103. Fraser R, McKinley RK, Mulholland H. Consultation competence in general practice: establishing the face validity of prioritized criteria in the Leicester assessment package. Br J General Pract 1994;44:109-13.
104. Bhola DS, Impara JC, Buckendahl CW. Aligning tests with states' content standards: Methods and Issues. Educ Meas 2003;22:21-9.
105. Green BF. A primer of testing. Am Psychol 1981;36:1001-11.
106. Webb NL. Research monogram No. 6: Criteria for alignment of expectations and assessment in mathematics and science education. Washington, DC: Council of Chief State School Officers, 1997:1-46.