International Journal of Medical Education. 2012;3:198-200
ISSN: 2042-6372
DOI: 10.5116/ijme.506f.1aaa

Editorial

Standard Setting: the application of the Receiver Operating Characteristic method

Mohsen Tavakol, Reg Dennick

Medical Education Unit, The University of Nottingham, UK

Correspondence: Mohsen Tavakol, Medical Education Unit, The University of Nottingham, UK
Email: mohsen.tavakol@ijme.net

Medical teachers must not only continue to accommodate and assimilate constant changes in medical knowledge, they must also assimilate new approaches to assess both students' declarative knowledge and procedural skills. Medical schools, postgraduate training programs and licensing bodies (e.g. the General Medical Council) must provide and oversee valid and reliable assessments for students' competence.[1] Assessments therefore should inform medical educators that students have a minimum acceptable level of competency which is also defensible. Over the past four decades educators have developed many standard setting methods in order to discriminate between competent and incompetent students. These methods identify different passing scores, which can be classified into two main groups, the test-centred and student approaches.[2, 3]

Nedelesky, Ebel and Angoff methods, which are commonly used in medical education, are considered test-centred approaches. These traditional methods are grounded in the subjective judgments of standard setters. Sometimes standard setters use the results of classical test theory (e.g. item difficulty and item discrimination parameters) to judge the behaviour of a minimally competent (borderline) student. However, the test centred approaches are based on a mathematical consensus of standard setter's judgements rather than an analysis of the test questions.[4] A shortcoming of these methods is that if the standard setters are changed the passing mark could be changed.[5]

In the student-centred approach, student abilities (ability scores) are calculated to identify a minimally competent student. A common standard setting method of the student - centred approach is the "borderline method" which is usually used in OSCEs, where standard setters identify students who performed between competent and unsatisfactory levels using a global rating scale. The pass mark of the station is then the average of the checklist scores for borderline students.

Most objective tests currently use the test-centred approaches with more attention increasingly being given to the student-centred approaches.[6] Other standard setting methods have been developed to differentiate students into two groups in order to identify the pass mark. For example it has been shown that there is a convergence between cluster analysis and the Angoff method for setting the pass mark.[5,7] Another technique is the "bookmark method", in which exam data are analysed using item response theory models.[8]

However, a standard setting technique that is more suitable for assessors with a clinical background is the Receiver Operating Characteristic (ROC) method which uses post-examination data for setting the pass mark. Colliver and colleagues have already used the ROC method to set passing standards for a standardised-patient examination of clinical competence.[9,10] We feel that this is a useful standard setting method that should be more widely known among the medical education community and the purpose of this editorial is to describe and advocate the ROC method.

What is the ROC?

The ROC method uses concepts taken from the analysis of laboratory tests that are familiar to many clinicians. Clinicians are interested in the accuracy or credibility of laboratory tests that can correctly classify patients into negative and positive groups. This helps them to make precise diagnoses and more accurate prognoses of patients with a given disease. The ROC method is derived from decision theory and was developed by radar engineers and used later by psychologists to explain perceptual detection of stimuli. This method has been used in medicine for many decades where ROC curves provide useful information for the evaluation of laboratory tests.[11] To better explore the ROC curve it is helpful to review the concepts of sensitivity and specificity in the context of clinical tests. Sensitivity refers to the ability of a test to correctly identify disease when disease is actually present (correctly identified positives). Specificity refers to the ability of a test to correctly identify non-disease when disease is actually absent (correctly identified negatives). However, in the context of achievement tests, sensitivity (or True Positive Rate (TPR)) refers to the proportion of correct answers, predicted by standard setters, which are

correctly answered by students. Specificity refers to the proportion of incorrect answers predicted according to standard setters which are incorrectly answered by students. 1 minus the specificity also refers to the False Positive Rate (FPR). It should be noted that the item difficulty parameter influences the sensitivity and specificity of an achievement test.

In order to classify students into pass-fail states, a cut score is calculated using a test-centred method (e.g. Angoff or Ebel) and students with marks less than the cut score are labelled as failures and those with values greater than or equal to the cut score are labelled as passers. The accuracy of such a classification can be identified by calculating the TPR and FPR. If we plot the TPR (on the y axis) against the FPR (on the x axis) at all possible cut scores the ROC can be created and tells us how the TPR and the FPR vary together (see Figure 1). The ability of the test to classify correct or incorrect answers, predicted by standard setters, that are answered correctly or incorrectly by students respectively, is measured by the area under the curve (AUC). A value of AUC $\leq$ 0.5 indicates that the test does not discriminate between passing and failing students.

The Youden index (J), the sum of sensitivity and specificity minus one, is used for setting a performance standard on the test.[12] This index indicates the optimal potential for the pass mark (Pass-fail cut-off) and should have a maximum value.

An illustrative example

Consider 320 students who participated in a hypothetical multiple choice physiology test consisting of 23 questions. Students who answered a question correctly received 1 point and those who answered the question incorrectly received 0. The cut score was 11, estimated by a standard setting method and therefore students who received a mark of 11 or greater passed the physiology test. Table 1 shows the descriptive statists of student marks.

Table 1. Descriptive statistics

N	Mean	SD	Min	Max	Skewness	Kurtosis	Available mark
320	19.68	2.93	10	23	-0.81	0.13	23

The TPR was plotted against the FPR using the possible cut scores of the test (Table 2) in order to create the ROC curve (Figure 1).

Some points in the ROC curve need to be examined. The lower left point (0, 0) indicates that the test does not classify students into two groups (competent and incompetent students). The upper left point (0, 1) indicates a perfect test, where students are classified into two groups and the cut score identified by the test is perfect (the standard setters have set the pass mark perfectly according to students' abilities). The diagonal, green, line is an indicator for a test with no discrimination (the two groups are indistinguishable). If the ROC curve lies below this line the discrimination of the test is poor. The area under the green line is equal to 0.5, indicating randomness. That is if we identify passers and failures in a test by chance, we will expect that a ROC curve will fall along the diagonal line. However, for a good test we want to have an AUC that is significantly greater than 0.5, indicating that the test discriminates between the low and high ability students. The result of the test shown in Figure 1 was statistically significant from 0.5 (AUC= 0.62, 95% CI= 0.50 to 0.75). A larger AUC indicates a better differentiation among students.

Table 2. The cut scores assigned according to the sensitivity and specificity indices

Cut scores	Sensitivity (TPR)	1- Specificity (FPR)	J
10.50	1.000	0.077	0.923
11.50	0.997	0.077	0.920
12.50	0.983	0.077	0.906
13.50	0.976	0.115	0.861
14.50	0.949	0.154	0.795
15.50	0.908	0.192	0.716
16.50	0.857	0.346	0.511
17.50	0.796	0.423	0.373
18.50	0.714	0.538	0.176
19.50	0.605	0.615	-0.010
20.50	0.459	0.731	-0.272
21.50	0.361	0.731	-0.370
22.50	0.207	0.808	-0.601

But what should the pass mark be? To do this we need to calculate the Youden index for each cut score. By inspection of Table 1, we see the cut score of 10.50 has a maximum value using the Youden formula (1.000 + (1- 0.077) - 1 = 0.923. Therefore the optimum pass mark for this test is 10.50, which is slightly less than the pass mark identified by the standard setters (11.0).

In conclusion, the ROC curve provides information which can be used to calculate the cut score that creates optimal differentiation between passing and failing students. It addresses all possible cut scores and reveals relationships between the sensitivity of the test and its FPR. Although we can assume a cut score generated by a test-centred method, such as Angoff or Ebel, it is not absolutely necessary since the method calculates all possible cut scores based on the optimal relationship between sensitivity and specificity of the test. This means that the method can generate an objective pass mark, based on student performance. We believe that this student-centred approach can provide valuable and defensible feedback for standard setters in order to monitor and better identify able and weaker students. These methods can confirm minimally competent students who are identified using traditional standard settings (test-centred approaches). We also recommend that standard setters use this method for research purposes as we require further evidence concerning the validity of applying the ROC method for setting the pass

mark in medical education. Finally, we hope this editorial stimulates some arguments about this student-centred approach and we welcome any discussion and publications concerning these methods.

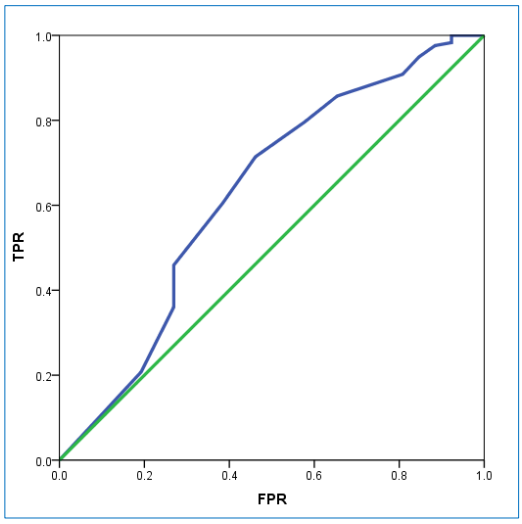

Figure 1. The ROC curve created by the possible cut scores of the test

References

1. Epstein R. Assessment in medical education. N Engl J Med. 2007;356:387-96.
2. Downing S, Tekian A, Yudkowsky R. Procedures for establishing defensible absolute passing scores on performance in health profession education. Teach Learn Med. 2006;18:50-7.
3. Jaeger R. Certification of student competence. In: Linn R, editor. Educational measurement. New York: American Council on Education and Macmillan; 1989. p. 485-515.
4. Sireci S. Standard setting using cluster analysis. In: Cizek G, editor. Setting performance standards: concepts, methods and perspective. London: Lawrence Erlbaum Associates, publishers; 2001. p. 339-65.
5. Sireci S, Robin F, Patelis T. Using cluster analysis to facilitate standard setting. Applied Measurement in Education. 1999;12:301-25.
6. Kane M. So much remains the same: conception and status of validation in setting standards. In: Cizek G, editor. Setting performance standards: concepts, methods and perspective. London Lawrence Erlbaum Associates, publishers; 2001.
7. Hess B, Subhiyah R. Convergence between cluster analysis and the setting minimum passing scores on credentialing examinations. Evaluation and the Health Profession. 2007;30:362-75.
8. Mitzel H, Lewis D, Patz R, Ross D. Standard setting using cluster analysis. In: Cizek G, editor. The bookmark procedure: psychological perspectives. London: Lawrence Erlbaum Associates, publishers; 2001. p. 339-65. P. 249-281.
9. Colliver J, VU N, Barrows H. Screening test length for sequential testing with a standardised-patient examination: a receiver operating characteristic (ROC) analysis. Acad Med. 1992;67:592-5.
10. Colliver J, Barnhart A, Marcy M, verhulst, S. Using a receiver operating characteristic (ROC) analysis to set passing standards for a standardised-patient examination of clinical competence. Acad Med. 1994;69:S37-39.
11. Zweig M, Cample G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin Chem. 1993;39:561-77.
12. Youden W. Index for rating diagnostic tests. Cancer. 1950;3:32-5.