2010; 32: 154–160 MEDICAL TEACHER

A standard setting method with the best performing students as point of reference: Practical and affordable

JANKE COHEN-SCHOTANUS[1] & CEES P. M. VAN DER VLEUTEN[2]

[1]University of Groningen and University Medical Center Groningen, The Netherlands, [2]Maastricht University, The Netherlands

Abstract

Background: Teachers involved in test development usually prefer criterion-referenced standard setting methods using panels. Since expert panels are costly, standards are often set by a pre-fixed percentage of questions answered correctly or norm-referenced methods aimed at ranking examinees.

Aim: To discuss the (dis)advantages of commonly used criterion and norm-referenced methods and present a new compromise method: standards based on a fixed cut-off score using the best scoring students as reference point.

Methods: Historical data from 54 Maastricht (norm-referenced) and 52 Groningen (criterion-referenced) tests were used to demonstrate huge discrepancies and variability in cut-off scores and failure rates. Subsequently, the compromise model – known as Cohen's method – was applied to the Groningen tests.

Results: The Maastricht norm-referenced method led to a large variation in required cut-off scores (15–46%), but a stable failure rate (about 17%). The Groningen method with a conventional, pre-fixed standard of 60% led to a large variation in failure rates (17–97%). The compromise method reduced variation in required cut-off scores as well as failure rates.

Conclusion: Both the criterion and norm-referenced standards, used in practice, have disadvantages. The proposed compromise model reduces the disadvantages of both methods and is considered more acceptable. Last but not least, compared to standard setting methods using panels, this method is affordable.

Introduction

There is a vast amount of literature on standard setting in assessment. One thing is clear: there is no gold standard (Friedman Ben-David 2000; Norcini 2003; Downing et al. 2006). Two main categories of standard setting methods can be distinguished: (1) *criterion-referenced* or *absolute* methods, where standard setting is independent of test results (Angoff 1971; Ebel 1979) and (2) *norm-referenced* or *relative* methods, where standard setting is based on test results (Norcini & Guille 2002; Downing et al. 2003). Norm-referenced standards are considered the method of choice when the aim is to rank examinees. Criterion-referenced standards are considered most appropriate when the aim is to ascertain whether examinees' mastery of a specific domain meets the pre-set requirements (Norcini 2003). Unfortunately, these two different approaches can yield widely divergent results on the same test (Norcini & Shea 1997; Cusimano & Rothman 2003; Downing et al. 2003; Reckase 2006).

Teachers tend to favour criterion-referenced standards because they seek safeguards to ensure that their students have attained the required level of mastery to certify competence. Criterion-referenced standard setting procedures typically require panels to determine the minimally acceptable level per item (Bandaranayake 2008). However, panel procedures are time consuming and, therefore, often *too costly* to use for in-house tests. The generally limited resources prohibit the regular use of panels for standard setting procedures. As a consequence, cut-off scores are often established in the form of a *pre-fixed percentage* of test questions that is to be answered correctly. Different countries have different traditions in defining minimum requirements.

Practice points

- Different standard setting methods lead to different test results. A gold standard does not exist.
- Criterion-referenced methods with a pre-fixed cut-off score lead to a large variation in failure rates. Norm-referenced methods lead to a large variation in cut-off scores. These disadvantages diminish the credibility and defensibility of the methods.
- A compromise method, combining a pre-fixed cut-off score with a relative point of reference, reduces the disadvantages of conventional criterion and norm-referenced methods, whilst making optimal use of the advantages. The method is more acceptable than conventional criterion and norm-referenced methods, and, last but not least, affordable for in-house tests.

Correspondence: J. Cohen-Schotanus, Center for Research and Innovation in Medical Education, University of Groningen and University Medical Center Groningen, A. Deusinglaan 1, 9713 AV Groningen, The Netherlands. Tel: 00 31 50 363 2884; fax: 00 31 50 363 2884; email: j.cohen-schotanus@med.umcg.nl

ISSN 0142–159X print/ISSN 1466–187X online/10/020154–7
DOI: 10.3109/01421590903196979

A cut-off score that is commonly used on the European continent is 55% or 60% correct answers after correction for guessing, but practices vary widely. In the UK, the norm is 50% (Nuffic 2006).

Unfortunately, the use of such pre-fixed cut-off scores can cause unexpected and substantial variation in pass/fail decisions, simply as a function of test difficulty. Norm-referenced standards, on the other hand, can result in unacceptably low cut-off scores and fixed numbers of examinees passing and failing the test, irrespective of the ability of the specific group that is being assessed. In both the cases, there is an undesirable mismatch between test results and expectations (Norcini & Guille 2002).

To minimise the disadvantages of both standard setting methods, we propose a compromise: a conventional, pre-fixed cut-off score with high performers as relative point of reference (Cohen-Schotanus et al. 1996). We have used this method for some years and, in our opinion, it is practical, acceptable and, last but not least, much more affordable than standard setting methods using panels. In the Netherlands, this method has come to be known as Cohen's method.

Because we feel that this standard setting method may be of use to a wider audience – especially faculty who prefer to certify competence using a criterion-referenced standard, but do not have the resources to convene standard setting panels – we will describe and explain why it was developed. After illustrating how conventional standards can lead to widely differing cut-off scores and highly variable failure rates, we compare the results of both conventional methods with those of our new method and infer the advantages of the new method as a compromise between the two approaches.

Two practices compared

To underline the large variation in cut-off scores and failure rates between, and even within, different standard setting practices, we will present the historical data used to develop Cohen's method. We examined the *cut-off scores* and *failure rates* of tests related to the same six block courses (6 weeks) administered to the first-year medical students at Maastricht University over a 9-year period (54 tests, administered between 1985 and 1994) and of 9 discipline-related tests administered to medical students of Groningen University over a 6-year period (1990–1995). Because we were only interested in tests relating to courses that had remained unchanged during the study period, the total number of included tests was 52.

In Maastricht, a norm-referenced standard setting method was applied at that time. The block-related tests consisted of 200 newly constructed true/false items. In keeping with the self-directed learning approach of the Maastricht problem-based curriculum, a question mark option was added and formula scoring was used where the percentage of incorrect answers was subtracted from the percentage of correct answers to yield the final score. The mean score minus one standard deviation was used to set the standard (Wijnen 1971). Cut-off scores varied between 15% and 46% (Figure 1), both within and between block tests over the years. There was no consistency or pattern that could explain this variation (i.e. across cohorts). Even within the same block course and with unchanged course delivery across years, cut-off scores varied substantially. The only logical explanation that remains is the use of new items with concomitant variation in test difficulty. The overall average percentage of students failing a test was stable at approximately 17%.

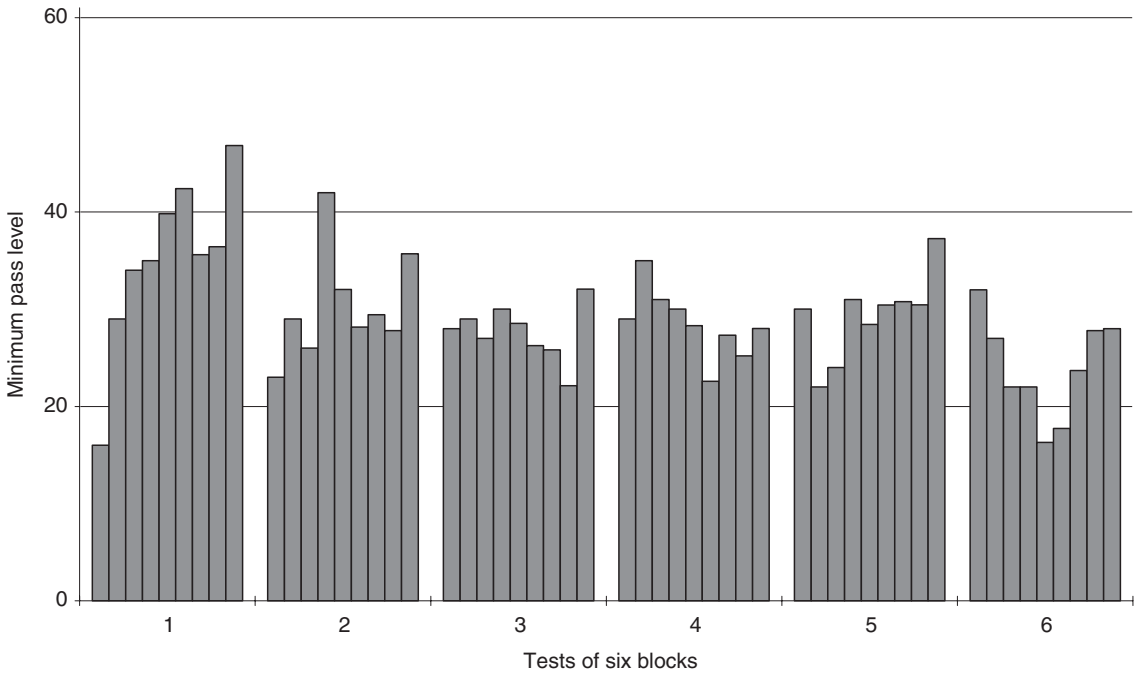

Figure 1. Minimum pass levels (%) of medical students on six written block-related tests in the first year of the Maastricht medical curriculum over a period of 9 years. The passing score was based on a relative standard setting procedure (average score minus one standard deviation).

In the same period, a conventional, criterion-referenced standard setting method was applied in Groningen. The discipline-oriented tests consisted on an average of 75 newly constructed multiple-choice questions (1 point for the single best answer). A pre-fixed cut-off score of 60% with correction for guessing was the standard. The resulting failure rates varied from 17% to 97% (Figure 2). So, here we find variation as we did in the Maastricht data, but, in this case, it is the failure rates that are not stable. Again, this variation appears to be rather random and occurs between and within disciplines, and again, it seems to point to a major effect of variation in test difficulty.

The standard setting practices in Groningen resulted in an average failure rate of 53% in contrast to an average failure rate of 17% in Maastricht with relative standards. Please note that these are real-life historical data and, moreover, that the Groningen practice is widely used in the European Higher Education Area.

What can we learn from this?

The differences between the outcomes of the two standard setting practices are larger than expected. First of all, there is the variability in cut-off scores (norm-referenced standard) and pass/failure rates (pre-fixed standard), even within the same courses across years. It is highly unlikely that these results can be attributed to course or test content, because we only included tests if the course had remained unchanged. Thus, the most probable cause is variability in test difficulty across different tests, both within and across courses. We can learn from this that any standard setting method that does not take test difficulty into account in some way is in danger of damaging credibility. From this perspective, norm-referenced standards are to be preferred over pre-fixed standards. It should be noted that criterion-referenced standards based upon panel judgements (such as Angoff and Nedelsky) co-vary with item and test difficulty (Verhoeven et al. 1999) supporting their credibility.

The second outcome is the large *average* difference in failure rates between the two methods, with the norm-referenced standard being much more lenient than the pre-fixed one. This begs the question whether this difference between the two standard setting methods could be associated with differences in knowledge gains. A plausible answer to this question can be found in the progress test results of the students from both universities. Progress testing is assumed to be curriculum independent (van der Vleuten et al. 2004). Test items are tailored to the Dutch national objectives for undergraduate medical education (Metz 1999). The obvious expectation would be for the Groningen students to outperform the Maastricht students, because they had to meet higher cut-off scores. However, we found no significant recurring differences, not even at graduation level, i.e. in year 6 of the two curricula (Bender et al. 1984; Verhoeven et al. 1998; Muijtjens et al. 2008). Apparently, low cut-off scores for in-house block-related tests have no linear negative impact on students' performance on progress tests. This leads us to the conclusion that the norm-referenced standards are *not* too lenient. Because of the numerous re-sit tests taken by the Groningen students, they took 1 year longer to graduate than their Maastricht colleagues of the same entering cohorts (Cohen-Schotanus 1999). If we were to express this in terms of effort and resources expended, we would come to the rather shocking conclusion that the use of standards, set by a conventional, pre-fixed cut-off score (percentage of items answered correctly), leads to a potential substantial waste of resources. Thus, we have evidence that pre-fixed cut-off

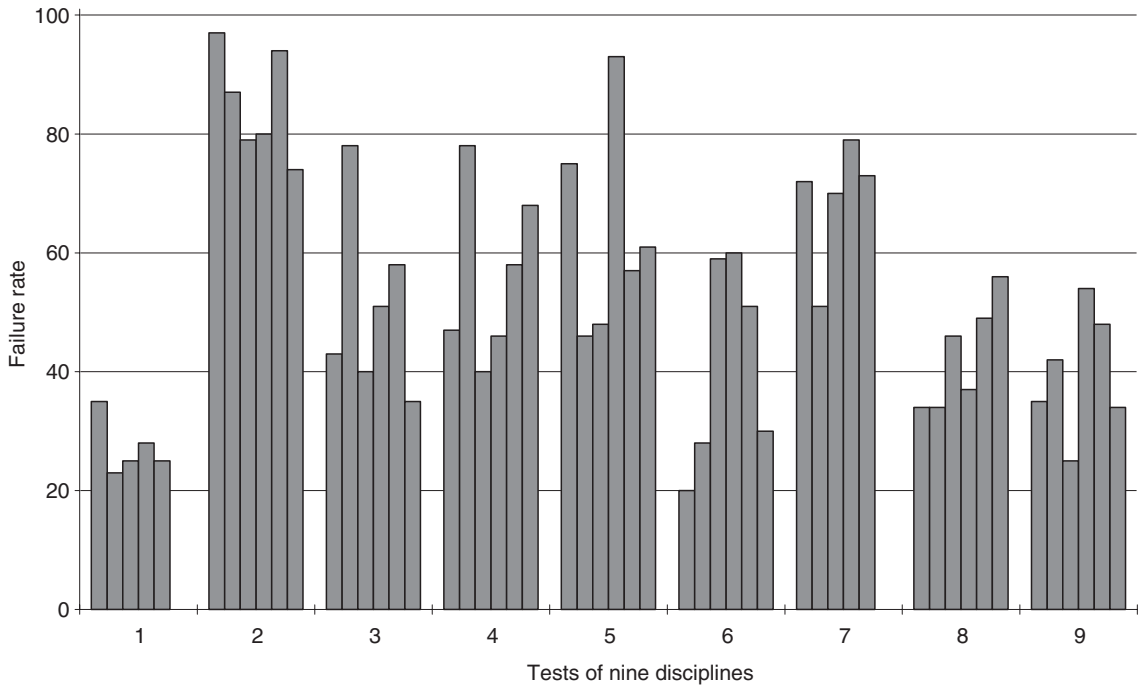

Figure 2. Failure rates (%) of 52 undergradute medical tests relating to nine disciplines based on the 60% absolute standard, corrected for guessing.

scores too often result in unrealistically high failure rates. So, again, we can only conclude that the use of such a standard setting method is less credible.

Finding the right balance

Despite the absence of a gold standard to identify the 'best' standard setting methods, some conditions have been identified as being able to enhance the quality of standard setting methods: (1) stakeholders agree on the standard setting method; (2) the method has credibility; (3) the method is based on sound research; (4) the method is easy to understand and use (transparency) and (5) the method yields realistic outcomes (Norcini & Guille 2002). Realistic outcomes combine a credible cut-off score with an acceptable failure rate. If a standard setting method produces unrealistic outcomes, it cannot be considered credible. Our experience has taught us that teachers have a tendency to blame unrealistic outcomes on students' shortcomings. Over the years, we have heard the same excuses and explanations for unexpectedly high failure rates time and again: students do not study hard enough; class attendance is low; the previous cohort was much more intelligent or the students were preoccupied with the Soccer World Championship in the run-up to the test. Students, on their part, complain that this year's test is substantially more difficult than last year's test; several test items are flawed; the items are not representative of course content or the handouts were not available in time. All these factors are a threat to the validity of a test (Downing & Haladyna 2004). Research also shows that flawed items make a test more difficult (Downing 2005).

It is an advantage of norm-referenced standards that they will correct for errors in the educational programme and the test. However, such standards bring their own idiosyncratic problems. The first problem is that students who do not prepare for the test and/or whose course attendance is low tend to depress the mean score and, consequently, the cut-off score. Because of this problem, most teachers consider norm-referenced standards as not acceptable. The second problem is that, regardless of the group's ability, a fixed percentage of examinees will always fail (Muijtjens et al. 1998), which seems rather unfair and compromises the credibility of this method.

Another way to correct for test difficulty and flaws in the educational or assessment process is to use a compromise method. An elegant example of such a method is the Hofstee method (Hofstee 1983). Judges are asked to specify two levels: a maximum cut-off score where a 0% failure rate is acceptable and a minimum cut-off score where even a 100% failure rate is acceptable. As a result, cut-off scores and failure rates will vary depending on the difficulty of the test. Unfortunately, we found that this method was too complicated and not sufficiently transparent to be acceptable to teachers and students. Too often, this method did not prevent high failure rates (Cohen-Schotanus et al. 1996). Once again, we found ourselves faced with the challenge to devise a method that combines the fairness of norm-referenced standards with the undeniable attraction of pre-fixed cut-off scores representing the minimally required standards of mastery.

The Cohen method

In an attempt to combine the advantages of both standard setting methods and diminish the disadvantages, we searched for alternatives. Pondering the problems posed by the instability of cut-off scores and failure rates, it dawned on us that there is one stable factor in the complicated process of standard setting: the best performing students. We decided to include their performance in the standard setting process by taking the highest student's scores as a point of reference instead of the mean score of all students. The argument supporting this idea is that, in large groups in any case, the results of the best performing students are the best estimate of the outcome to be attained on that particular test. The brightest students who understand and have mastered the course material and irregularities in educational practices are unlikely to have a substantial impact on their performance. Nevertheless, variation in test difficulty will inevitably affect the results of the group, including those of the brightest students. Therefore, we decided to set the standard at 60% of the highest achiever's score instead of the cut-off score of correct answers for 60% of the total number of items. The rationale underlying this compromise method turned out to be easy to explain to both teachers and students and proved highly acceptable. Moreover, the method is transparent: students can be sure in advance that 60% correct answers (after correction for guessing) is a guaranteed pass. Naturally, 60% correct answers is an arbitrary minimum pass rate, which is tailored to the Dutch conventional way of grading exams. In the following section, we will compare the outcomes of the proposed compromise method with those of the other standard setting methods.

Application of Cohen's method

In order to compare the methods, we applied the norm-referenced standard, the pre-fixed standard and Cohen's method to the 52 Groningen tests to determine the differential effects on cut-off scores and failure rates. For Cohen's method, we used two points of reference: (1) the highest scoring student; and (2) the 95th percentile point.

For Cohen's method, the following formula was applied:

$$\text{Standard} = cN + 60(N^* - cN)$$

with c being the proportion of chance performance in the test, N the total number of items in the test and N^* the score of the best performing student(s).

The norm-referenced standard and Cohen's method yield lower cut-off scores (Figure 3) and considerably lower failure rates (Figure 4) than do the pre-fixed 60% standards. The outcomes are summarised in Table 1. Application of the norm-referenced standard to the Groningen tests leads to average failure rates of 15%, which is comparable to the Maastricht's 17%. Unlike the Maastricht students, the Groningen students could not use the question mark option. They had to answer all the questions and were not penalised for incorrect answers. Therefore, the Groningen tests had higher cut-off scores with the norm-referenced standard than their Maastricht counterparts.

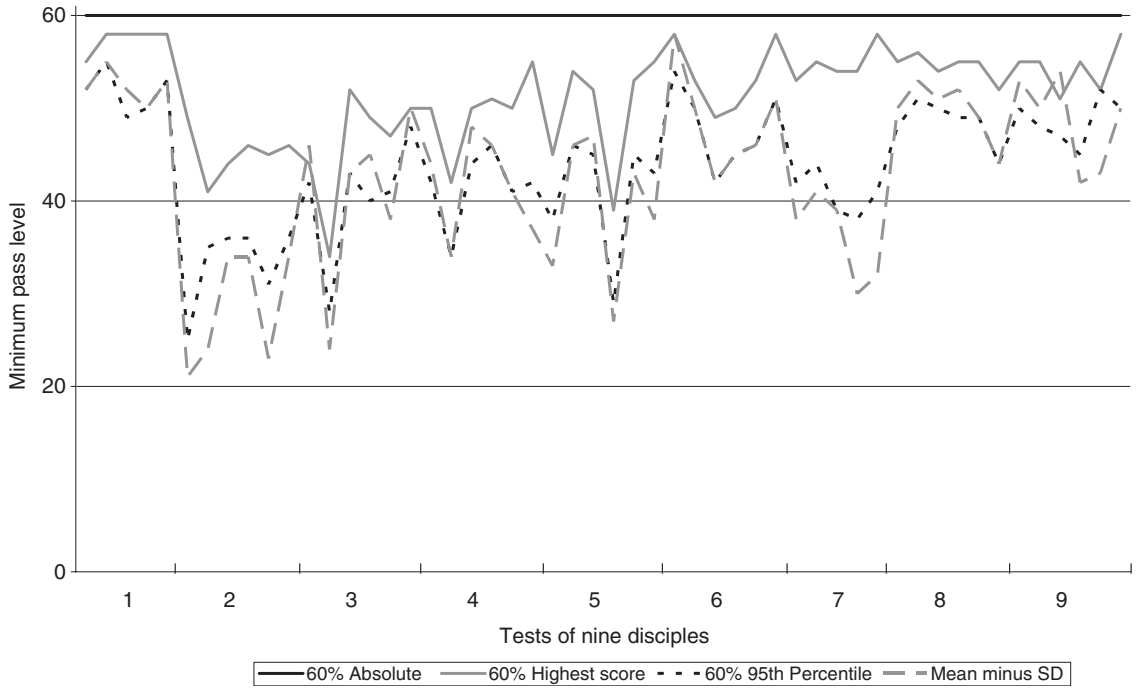

Figure 3. Minimum pass levels of 52 tests of third and fourth-year medical students based on different standard setting procedures: (1) 60% absolute standard, (2) 60% of the highest score, (3) 60% of the 95th percentile and (4) the relative standard of mean score minus one standard deviation.

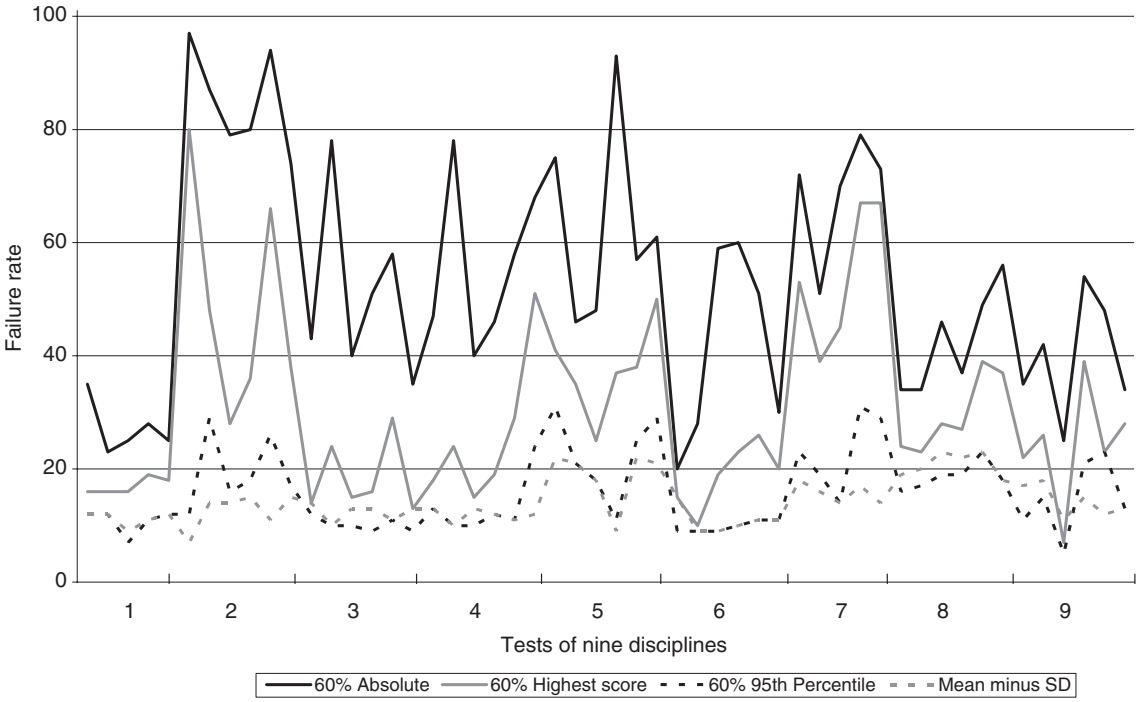

Figure 4. Failure rates (%) on 52 tests of third- and fourth-year medical students based on different standard setting procedures: (1) 60% absolute standard, (2) 60% of the highest scoring student, (3) 60% of the 95th percentile (all corrected for guessing) and (4) the relative standard of mean score minus one standard deviation.

Application of Cohen's method based on both points of reference results in less fluctuation in failure rates (SD = 16.07 and 6.80, respectively) compared to the pre-fixed 60% standard (SD = 20.47). At the same time, Cohen's method diminishes the fluctuation in cut-off scores (SD = 5.40 and 6.85, respectively) compared to the relative standard (SD = 9.17). Inspection of the standard deviations of pass levels and failure rates across methods (Table 1) reveals that

Table 1. Average minimum pass levels and failure rates (%) for 52 Groningen tests based on four standard setting methods.

Standard setting method	Average competence level (%)	Standard deviation	Range competence level	Average failure rates (%)	Standard deviation	Range failure rates (%)
60% absolute	60	0.00	60	53	20.47	17–97
60% of the highest score	52	5.40	34–60	30	16.07	13–80
60% of the 95th percentile	44	6.95	25–57	17	6.80	7–39
Relative mean minus SD	43	9.17	21–58	15	4.17	9–23

the best balance is reached with Cohen's method using the 95th percentile as point of reference.

Discussion

We compared the results of three affordable standard setting methods: the traditional pre-fixed standard setting method in Europe, a norm-referenced standard setting method and a new, compromise method, Cohen's method. The norm-referenced standard setting method as used in Maastricht is characterised by large fluctuation of cut-off scores; the pre-fixed standard as used in Groningen typically shows large variations in failure rates. In our opinion, these fluctuations are mainly attributable to variations in test difficulty. Cohen's method succeeds in reducing the fluctuations in both variables.

Cohen's method requires a point of reference to be chosen before the students sit for the test. We analysed two options: the highest scoring student and the 95th percentile point. With the highest scoring student as the reference point, the average failure rate is 31%, a decrease of 22% compared to the original pre-fixed standard, whereas the average cut-off score shows a decrease of no more than 8%. However, although the considerable gain achieved with this point of reference is undeniable, the fluctuations in failure rates are still (too) large, probably due to the low reliability of a single individual's score as reference point. With the 95th percentile as point of reference, variation in failure rates is reduced but minimum pass levels rise. The 95th percentile outcomes most closely resemble those of the relative standard.

So, why do not we just apply a conventional norm-referenced standard? In our experience, teachers find it more acceptable to certify students' competence in relation to the best performing students than in relation to the group average. In our opinion, the choice of a point of reference has to be guided by the intentions and motives of faculty and has to be in alignment with local practices.

In conclusion, for high quality, standard setting panels are preferable. However, for in-house tests, when resources for such expensive methods are lacking, our compromise method has the advantage of being simple and combining the advantages of widely used traditional pre-fixed standards and norm-referenced methods. While attenuating the disadvantages of these methods, it makes optimal use of the advantages, thereby enhancing its acceptability. Last, but not least, the method is *affordable* and has been proven to function well in educational practice at different Dutch universities. We would like to advocate the use of Cohen's method and are very interested in reading about the outcomes of its use in other countries.

Declaration of interest: The authors report no conflicts of interest. The authors alone are responsible for the content and writing of this article.

Notes on contributors

JANKE COHEN-SCHOTANUS, PhD, is a Professor and Head of the Center for Research and Innovation in Medical Education, University of Groningen and University Medical Center Groningen, The Netherlands.

CEES P.M. VAN DER VLEUTEN, PhD, is a Professor and Chair of the Department of Educational Development and Research, Maastricht University, The Netherlands.

References

Angoff WH. 1971. Scales, norms and equivalent scores. In: Thorndike RL, editor. Educational measurement. 2nd ed. Washington, DC: American Council on Education. pp 508–600.

Bandaranayake RC. 2008. Setting and maintaining standards in multiple choice examinations: AMEE Guide No. 37. Med Teach 30:836–845.

Bender W, Cohen-Schotanus J, Imbos T, Versfelt WA, Verwijnen GM. 1984. Medisch kennis bij studenten uit verschillende medische faculteiten: Van hetzelfde laken een pak? [Medical knowledge of students from various medical schools: being served with the same sauce?] Ned Tijdschr Geneesk 128:917–921.

Cohen-Schotanus J. 1999. Student assessment and examination rules. Med Teach 21:318–321.

Cohen-Schotanus J, van der Vleuten CPM, Bender W. 1996. Een betere cesuur bij tentamens [A better standard setting method for written tests]. Onderzoek van Onderwijs September:54–55.

Cusimano MD, Rothman AI. 2003. The effect of incorporating normative data into a criterion-referenced standard setting in medical education. Acad Med 78(10 Suppl):S88–S90.

Downing SM. 2005. The effects of violating standard item writing principles in tests and students: The consequences of using flawed test items on achievement examinations in medical education. Adv Health Sci Educ 10:133–143.

Downing SM, Haladyna TM. 2004. Validity threats: Overcoming interference with proposed interpretations of assessment data. Med Educ 38:327–333.

Downing SM, Lieska NG, Raible MD. 2003. Establishing passing standards of classroom achievement tests in medical education: A comparative study of four methods. Acad Med 78(10 Suppl):S85–S87.

Downing SM, Tekian A, Yudkowsky R. 2006. Procedures for establishing defensible absolute passing scores on performance examinations in health professions education. Teach Learn Med 18:50–57.

Ebel RL. 1979. Essentials of educational measurement. 3rd ed. Englewood Cliffs, NJ: Prentice Hall.

Friedman Ben-David, M. 2000. AMEE Guide No. 18: Standard setting in student assessment. Med Teach 22:120–130.

Hofstee WKB. 1983. The case for compromise in educational selection and grading. In: Anderson SB, Helminck JS, editors. On educational testing. San Francisco: Jossey-Bass. pp 107–127.

Metz JC. 1999. "Blueprint 1994": Common objectives of medical education in the Netherlands. Neth J Med 55:165–167.

Muijtjens AMM, Hoogenboom RJI, Verwijnen GM, van der Vleuten CPM. 1998. Relative or absolute standards in assessing medical knowledge using progress tests. Adv Health Sci Educ 3:81–87.

Muijtjens AMM, Schuwirth LWT, Cohen-Schotanus J, Thoben AJNM, van der Vleuten CPM. 2008. Benchmarking by cross-institutional comparison of student achievement in a progress test. Med Educ 42:82–88.

Norcini JJ. 2003. Setting standards on educational tests. Med Educ 37:464–469.

Norcini J, Guille R. 2002. Combining tests and setting standards. In: Norman GR, van der Vleuten CPM, Newble DI, editors. International handbook of research in medical education. Dordrecht: Kluwer Academic Publishers.

Norcini JJ, Shea JA. 1997. The credibility and comparability of standards. Appl Meas Educ 10(1):39–59.

Nuffic. 2006. Cijfers ontcijferd [Grades graded]. The Hague: Netherlands Organization for International Cooperation in Higher Education–Nuffic.

Reckase MD. 2006. A conceptual framework for a psychometric theory of standard setting with examples of its use for evaluating the functioning of two standard setting methods. Educ Meas: Issues Pract 25(2):4–18.

van der Vleuten CPM, Schuwirth LWT, Muijtjens AMM, Thoben AJNM, Cohen-Schotanus J, van Boven CPA. 2004. Cross institutional collaboration in assessment: A case of progress testing. Med Teach 26:719–725.

Verhoeven BH, van der Steeg AF, Scherpbier AJ, Muijtjens AM, Verwijnen GM, van der Vleuten CP. 1999. Reliability and credibility of an Angoff standard setting procedure in progress testing using recent graduates as judges. Med Educ 33:832–837.

Verhoeven BH, Verwijnen GM, Scherpbier AJJA, Holdrinet RSG, Oeseburg B, Bulte JA, van der Vleuten CPM. 1998. An analysis of progress test results of PBL and non-PBL students. Med Teach 20:310–316.

Wijnen WHFW. 1971. Onder of boven de maat [To be or not to be up to the mark] [Dissertation]. Lisse: Swets and Zeitlinger.