Some time ago, I wrote a paper about the assessment of competence in which I argued that any assessment situation is inevitably a compromise between what is desirable and what is achievable (Van der Vleuten, 1996). There are no fixed and firm strategies that guarantee the perfect compromise. Things strongly depend on specific assessment contexts and local conditions. The challenge is to take a firm grasp of the assessment situation, including its psychometrics and other relevant factors, and then weigh these arguments and combine them into an informed compromise that one feels can be justified when challenged. I have no doubt that this same perspective applies to standard setting. In principle, with standard setting we have the technology to do a very good job indeed, although outcomes may vary considerably and some arbitrariness is inevitable. What needs to be considered, however, is how defensible procedures are and how much effort one is prepared to put into due process. How to do an excellent job on standard setting is admirably described in AMEE guide no. 37 (Bandaranayake, 2008). A preference is expressed for the criterion-referenced approach, with a well-trained and preferably sizable panel of judges, who use one of several techniques to reach a collective judgment of the passing score. Although there is no gold standard for this approach, the (modified) Angoff method comes close. If one should wish to do an even better job, one can always turn to a compromise method that enables calibrating for test difficulty by taking account of panel outcomes as well as group performance data. A still better job can be done by calibrating test scales using test equating methods such as those described in the AMEE guide. One step further along the best practice road takes one to Item-Response Modelling for statistical equation (Downing, 2003). Using these procedures one can estimate item and test difficulty independently of the ability of the group on which the estimates are based. In fact, such a calibrated bank of questions enables one to select items that best discriminate around the preselected standard, as is proposed in this Guide. Finally, the most sophisticated approach allows one to tailor test content to individual test takers' abilities. After a candidate has answered the first item, every next item is selected (by computer) in alignment with the candidate's performance on the previous items (tailored testing). And all this technology is available today! But what we need to ask here is: who can afford it? Yes, certification agencies around the world all have sufficient resources to make their professional standard setting practice nearly perfect. But what about the run-of-the-mill medical school, or small colleges of specialty training? What can they do to raise the standards of their exams? Where should and can they compromise? Which valid arguments are suitable to underpin their 'informed' compromises?

In the AMEE guide, Raja Bandaranayake rightfully starts by explaining the concept of noise affecting any measurement. It is good measurement practice to aim for maximum noise reduction. I would argue that by far the biggest source of noise around standard setting is variation in test difficulty. A while ago, I carried out a variance component estimation on test data of 7 cohorts of medical students of Maastricht University (unpublished). Having calculated the average test result across all the course exams over four years of preclinical training, I was able to estimate
variation not only across cohorts, but also across tests and even across examinees within tests, because variance component estimation allows one to express the contributions of these sources of variance relative to each other. The results of my experiment were interesting. Fortunately, the largest variance (82.60%) was attributable to variation between students, which is good, because after all the prime purpose of the entire assessment exercise is to discriminate between students. The second largest component was unexplained variance (12.42%) or general error. This is not surprising considering that general error is frequently the main variance component in any form of measurement. The third component was variation between tests (4.97%). Finally, explaining only 0.01% of the variance, the smallest component was the variance associated with cohort (cohort size varied between 150 and 200 examinees)! It is hard to compete with unexplained variance without further information on its sources, but test difficulty effects are much more impressive than cohort effects. During our continuous monitoring of progress testing, we have consistently observed a marked effect of test difficulty on group averages (Muijtjens, Schuwirth, Cohen-Schotanus, Thoben, & van der Vleuten, 2008), despite very rigorous quality assurance procedures around the construction of every test. An important implication of this notion is that any standard setting method that does not adjust for test difficulty one way or another is hard to defend (and the correlation of item-pass marks with item p-values is an important external validity criterion). And if we were to push this line of reasoning to its ultimate conclusion, we would probably be forced to disapprove of the vast majority of standard setting practices in smaller organizations such as the ones I referred to above. For my own country I think I can safely say that 98% of all multiple choice (and other) exams have an arbitrary, pre-set passing score of 60% (after correction for guessing) without any correction of the scoring scale (for example through test equation). Unsurprisingly, variation in pass/fail rates is huge (Cohen-Schotanus & Van der Vleuten, In press). I would even go so far as to generalize this state of affairs to the whole of Europe. So what it all boils down to is that the least defensible approach to standard setting is also the one that is most commonly used! Indeed I very much hope that this AMEE guide will be read and taken to heart by many, leaving a strong imprint on educational practice.

So, now we have to tackle the next question, that of how to reach a workable compromise now that we have decided that the most popular approach is to be rejected as unacceptable. Test difficulty being the general problem, I am inclined to argue that simple norm-referenced procedures are defensible ones. Although I am aware that the resulting standards are subject to fluctuations due to group differences and consequently are likely to have poor credibility among the teaching community at large, I would like to point out, in defence of these standards, that with large enough groups these adverse effects are negligible compared to those of variations in test difficulty. Moreover, there are strategies one can use to include the measurement error in the relative standard. For example, a method proposed by Wijnen uses the mean test score as the starting point (the average student is assumed to be competent) and the mean score minus two times the standard error of measurement (SEM) as the cut score (Wijnen, 1971). The fact that the SEM includes test reliability as a parameter leads to more stringent cut scores with more reliable tests. Admittedly, this procedure may be seen as inherently unfair since it produces a more or less fixed percentage of failures regardless of the ability of the group as a whole. To combat this potential flaw, Cohen has recently proposed an alternative approach with the highest scoring students as the reference point for setting the pass/fail score at a fixed percentage (60%) of mastery (Cohen–Schotanus & Van der Vleuten, 2010). In addition to this cut score, fixed percentages may be pre-set for either minimum or maximum competence (or failure rates), as is done in Hofstee’s method. My central argument in support of relative standards is that in assessment contexts where highly sophisticated methods are unaffordable,
relative standards are more defensible than the commonly used method of a fixed domain percentage. As professionals in assessment we may have good reasons for preferring criterion referencing but, when proper and costly measures to accommodate variations in test difficulty are just not feasible, norm-referenced procedures are likely to be the more defensible, informed compromise.

So why not go for the best of both worlds and massively opt for compromise methods such as the Hofstee one? As a matter of fact, compromise methods seem rather ideal and I think Hofstee’s method was a true stroke of genius (Hofstee, 1983). That is why I would applaud initiatives to research its status in educational practice as well as experiences associated with these practices. In my own experience, the utility of Hofstee’s method is jeopardized by its being hard to explain to our stakeholders and compromise being hard to attain because the distribution curve of score versus failure rate lies outside the quadrant of boundary values \((c_{min}, c_{max}, f_{min}, f_{max})\). Some sound research is badly needed here. Another problem is that Hofstee’s publications are not easily accessible. Typically his method is explained graphically, such as in this AMEE guide. However, he has also published a mathematical solution (Hofstee, 1977), which many may find convenient for doing calculations. I once developed an SPSS syntax for calculating pass/fail decisions using the Hofstee procedure which I am happy to share with anyone interested (but will not accept any liability for any errors!).

Finally, there is an anecdote worth telling, if only to preserve this piece of educational history in print. In 1971, Wijnen, my former boss, proposed the above method prompted by his practical experiences with arbitrary, fixed standards at the University of Groningen, the Netherlands. He did so in his doctoral dissertation. Hofstee, despite having supervised Wijnen’s dissertation, was not satisfied with the solution proposed therein. His misgivings kept nagging him until, some time later, he found and published his own solution in a paper titled: ‘the standard setting problem resolved’ (Hofstee, 1977). After that Hofstee lost interest in the subject. It is all the more intriguing that his name has come to be associated with a standard setting method. Hofstee is a retired personality psychologist, who has written extensively and brilliantly, including a masterpiece on judgment in assessment situations (Hofstee, 1999).

To sum up, this AMEE guide gives an excellent overview of standard setting methods for multiple choice items. The technology for good standard setting is widely available. Nonetheless, the key question to be answered is whether it is affordable in all contexts, leading on to the question of which compromise is most appropriate in which context. I realize that I may be accused of taking a rather indecisive, typically Dutch, consensus approach. Still, I am convinced that this view will pave the way for informed compromises that will help us to develop standards that are fair and feasible in realistic contexts.

References


Notes on Contributor

CPM van der Vleuten is Professor of Education and Chair of the Department of Educational Development and Research, Maastricht University.

Prof. Cees Van der Vleuten, Universiteit Maastricht, Educational Development and Research, P.O. Box 616, Maastricht, 6200 MD, Netherlands. Tel: þ31 43 3885725; fax: þ31 43 3885779 email: c.vandervleuten@educ.unimaas.nl


This AMEE Guide Supplement was published in Medical Teacher 2010. 32:174–176.