Setting and maintaining standards in multiple choice examinations
André F. De Champlain, National Board of Medical Examiners, Philadelphia, USA

One of the central activities of any organization involved in assessment, especially within a criterion-referenced framework, is to set a passing standard which presumably reflects mastery of the domains targeted by an examination. Standard setting is especially critical with examinations used in medical education as we need to assure the public that license and certificate holders possess the skills and knowledge necessary for safe and effective patient care (De Champlain, 2004). Furthermore, some research has also suggested that performance on medical licensing examinations is related to the probability of committing egregious acts in (future) practice (Tamblyn et al., 1998; 2002; Wenghofer et al., 2009). Therefore, standard setting has implications not only for the “here & now” but also for future quality of care.

The AMEE guide published by Bandaranayake (2008) provides a useful overview of norm-referenced and criterion-referenced methods that have been proposed for setting a standard with multiple-choice examinations. The latter two include test-centered approaches (the Angoff and Ebel methods) and examinee-centered approaches (the borderline and contrasting groups methods). Additionally, so-called compromise methods (the Hofstee, De Gruijter and Beuk methods) are described. As the author underscores, criterion-referenced methods should be preferred to norm-referenced approaches with medical certification and licensure examinations. Though it is possible to account for differences in test form difficulty via equating to ensure that cut-scores represent the same proficiency across versions, the use of a norm-referenced approach for setting a standard on a certification or licensure examination is ill-advised as the resulting cut-score does not reflect the level of knowledge and skills required for acceptable performance in the profession, due to the absence of expert input. Additionally, the use of a norm-referenced approach also precludes calculating classification errors (false positive and false negative decisions), a central concern in high-stakes testing. However, unlike what is stated by the author, a criterion-referenced standard is not necessarily set prior to the administration of the examination. In fact, most exercises which use a variation of the Angoff method that incorporates performance data are completed, by necessity, after the test has been given and prior to reporting scores and pass/fail decisions to examinees. Related to this point is the contentious issue of whether or not performance data should be used as part of a “modified” Angoff standard setting exercise. The rational that initially supported the use of such data had to do with the often unrealistic standards that were obtained without this “reality check”. However, recent research has suggested that these performance data may not be used to augment judgment, as originally conceived, but in a mechanical fashion that de facto reduces the exercise to a norm-referenced approach (Clauser et al., 2009a). More research is needed to address this important issue.

Following this presentation of common standard setting models, it might have been useful to also include a brief user’s guide to aid the practitioner in adopting the most appropriate method in light of a number of considerations, including the nature of the decision and the type of assessment. For example, the Angoff method is compensatory in nature as the final standard is obtained by summing or averaging individual item judgments. As such, candidates can compensate for doing poorly in certain sections of
the examination by doing well in other parts of the test. Therefore, the Angoff method is inappropriate in instances where the examinee needs to pass certain key items or sections on the examination, regardless of how they might have fared on the rest of the test, i.e., with conjunctive standards. Also, examinee-centered approaches are rarely used with selected-response examinations but are highly recommended for performance assessments as they entail a direct judgment about the (pass-fail) status of candidates with respect to the construct of interest (Kane et al. 1999).

The paper also clearly reiterates two issues that any researcher must contend with when undertaking a standard setting exercise: (1) standard setting is a judgmental process; and (2) resulting standards are method dependent. As stated in Bandaranayake (2008), in addition to selecting a representative set of judges to constitute the panel (with respect to age, sex, specialty, etc.), offering extensive training and opportunities for discussion can help to ensure that panelists are in agreement on the nature of the task as well as the key characteristics of the borderline or minimally proficient candidate. It is also critical to not only clearly document all stages of the standard setting process for transparency purposes but also to provide both reliability and validity evidence to support its use. With respect to reliability, both generalizability theory and the many faceted Rasch model are useful to identify the extent to which the cut-score varies as a function of the facets or sources of variation in the design, including judges, panels (if multiple panels of experts participated), item sets (if a sample of items is used to set the standard), etc. (Kim & Wilson, 2009). In regard to validating the standard, the suggestions provided by Bandaranayake (2008) are excellent ones and include looking at the relationships between passing the examination and other criteria as well as assessing the consequences and policy implications of implementing a given standard.

The section on equating seems out of place and might be better served by being included in a general linking and scaling monograph (Mislevy, 1992). This section does nonetheless underscore the importance of not confounding standard setting with equating as these are two distinct phases of examination processing. As stated by the author, equating is the statistical process by which scores are adjusted to account for slight variations in difficulty across test forms to ensure that we are completely fair towards examinees that may have completed slightly easier or more difficult examinations. Despite our best intentions to assemble forms that are comparable, not only with respect to content but also statistical targets, examinations will invariably differ slightly in difficulty. Equating includes a number of models that aim to (statistically) adjust test scores to ensure that the latter are comparable, irrespective of the examination that may have been completed. Additionally, equating scores enables large-scale testing programs to track performance longitudinally; an activity that cannot be legitimately undertaken with un-equated scores. Standard setting should also not be used to make these adjustments as research has repeatedly shown that experts are very poor at judging item difficulty (Impara & Plake, 1998; Clauser et. al, 2009b). Conversely, equating scores across test forms does not negate the importance of periodically reviewing the appropriateness of a standard in light of changes in practice, knowledge and other factors. Both activities have distinct purposes and should be completed as part of routine examination processing.

Bandaranayake (2008) concludes his guide by making a strong case for targeting examinations at the cut-score, using previously banked item difficulties. He continues by stating that “Surprisingly, this advice seems to have been unheeded by testing bodies” (p. 843). However, this approach to test assembly is commonplace with several large-scale certification and licensure programs, both within and outside of medicine. Optimization techniques, implemented in automated test assembly (ATA) software, were proposed to facilitate the development of multiple test forms according to a number of content and statistical targets (van der Linden, 2005). For example, using item response theory,
the user can select items from the bank that maximize information near the cut-score value, both to ensure that pass/fail decisions are rendered with the highest level of accuracy as well as to minimize classification errors (Luecht, 2006).

In closing, Bandaranayake’s guide is a valuable addition to the medical education literature, especially for those who possess rudimentary knowledge of standard setting and crave a gentle introduction to what constitutes a voluminous literature. In this regard, the guide accomplishes its goal admirably.

References


Note on Contributor

Andre F. De Champlain, Ph.D. is Principal Measurement Scientist at the National Board of Medical Examiners.

Dr A. De Champlain, National Board of Medical Examiners, Measurement Consulting Services, 3750 Market Street, Philadelphia, PA 19104, USA. Tel: 1 215 5909500; fax: 1 215 5909449; email: adechamplain@nbme.org


This AMEE Guide Supplement was published in Medical Teacher 2010. 32:436–437.