Journal of MultiDisciplinary Evaluation
Volume 14, Issue 30, 2018

JMDE
Journal of MultiDisciplinary Evaluation

ISSN 1556-8180
http://www.jmde.com

Customizing the Standard to the Purpose of the Assessment

Sandra Monteiro
McMaster University

Debra Sibbald
Touchstone Institute

The purpose of standard setting in health professions education is to facilitate decisions following the assessment of competence at admissions, during training or at certification. Typically, once a standard is set, it is applied to distinguish only between those who pass and fail. Conceptually, it is relatively straightforward to identify either the consistently skilled or the consistently unskilled applicants; it is the group of individuals that are inconsistent that pose the greatest challenge. As a result, a fundamental concept that any standard setting process requires is a clear definition of that borderline candidate; the individuals who accomplish only some tasks in an acceptable manner. The characteristics of the borderline group are especially relevant to helping examiners evaluate candidates during any kind of performance assessment. Inevitably, however, the standard, which may be only a written description of acceptable and unacceptable behaviours, must be translated into a numerical cut-score.

While it is generally understood that there is no single best approach to identifying a cut-score (Swanwick, 2014), we find that in the context of assessing internationally graduated health professionals, there are a few factors that point to some methods being more appropriate than others. Additionally, in this context, there is often increased pressure to develop standards for re-training or remediation pathways, as well as pass-fail decisions. It is in this context that we offer guidance regarding how to select a standard setting method and cut-score.

At Touchstone Institute in Toronto, Canada we are involved in the development and evaluation of high stakes assessments of internationally graduated health professionals (IGHP). We support the needs of regulators and professional colleges as they seek to create standards for entry to practice- and skills-bridging programs. These standards are intended to ensure the fair selection of IGHPs who can proceed to the licensing exam and those who are referred to further training. Coming from backgrounds in education and psychology, we bring a wealth of understanding to these discussions, but what has been most interesting to us is that we could not simply apply a method directly as described in the literature; we always customize the approach for each assessment we build.

We realized that the abundance of approaches in the literature creates significant confusion about what method to use and when, particularly for the novice standard setter. Regulators not only have to select methods that are objective and valid, but also defensible and fair to candidates and local or national legislation; the pressure to select *the best* approach is very high. Without understanding the mechanics of various methods, regulators are only able to gain a sense of familiarity with methods like Angoff, Ebel, and borderline regression (Norcini, 2003) and are unsure how to decide between methods that are norm-referenced, criterion-referenced and/or panel-based. Inevitably regulators have heard about various methods but do not understand the limitations or unique characteristics of their own context and how that could influence the decision. This lack of understanding is only exacerbated in the assessment of internationally graduated health professionals; a very heterogeneous and sometimes unpredictable group. Unfortunately, this can result in unreliable, unsafe and unfair assessments. Our own experience in reviewing the literature highlighted two important points of dissonance that regulators should be aware of and that support the need for a customized approach. We feel these points are also important to all contexts of assessment and can help contextualize assessment data when making decisions about standard setting.

First, in the context of assessing local graduates, those with a homogeneous and predictable education background, panel-based standard-setting methods may be appropriate as they incorporate the experience of health professional and educator experts from the local context. Experts are able to rely on a normative process to judge the potential of the average borderline candidate given their experience as instructors in the field. In the context of assessing international graduates, however, the opinion of the expert and their norm reference process may be significantly flawed. Often experts included in standard setting processes are taken from education settings in local context, and have little to no experience understanding the needs or aptitude of the international graduate. Therefore, their predictions about item performance are also inevitably misaligned. The consequence may result in a standard that is too low or too high.

Second, the premise in most standard-setting literature is that applicants truly represent a heterogeneous sample of possible candidates. Most evidence or comparisons of standard-setting methods, come from medical school admissions data and most medical school applicants are *relatively* homogeneous; compared to Internationally Graduated Health Professional (IGHP) applicants. In our experience and context, when assessing only IGHPs, the reliability of standardized assessments is inevitably high due to the large variation in performance scores. As a concrete example, the identical version of an exam (written or performance-based) administered to a local group of Canadian graduates or in-practice professionals will likely result in a moderate Cronbach's alpha of 0.5 to 0.6, but with a group of IGHP candidates, reliability rises to 0.8 and sometimes 0.9.

> Classic reliability estimates hinge on the variance of the total scores. As this variance increases the reliability estimate will also tend to increase, due to greater theoretical confidence that we have appropriately measured the participants on the trait of interest. One implication of the role of the total score variance is that different samples will likely yield different reliabilities because the total variance will likely change. (Henson, 2001)

The classic definition of reliability is consistently reflected in our assessment data as we often compare the performance of local graduates to IGHPs. While the expected knowledge and skills may be identical for IGHPs and local graduates challenging a certification exam, there will be a broader range of scores for IGHPs. This may also translate into different indicators of unsafe behavior that are relevant to the IGHP. The use of a criterion group standard-setting method can greatly facilitate the assessment of IGHPs and help define the scores that represent pass and fail as well as scores that reflect different levels of re-training or remediation.

At Touchstone Institute, we work with regulators to establish a criterion-based standard because the goal is to ensure that the IGHP is compared to the locally trained graduate. Our process includes evaluating the quality of the assessment using locally trained senior students and recently certified graduates. We employ this process for both written and performance-based examinations. The variability between students and graduates is often sufficient to ensure variance in scores and a minimum acceptable level of reliability. This allows us to then establish the range of possible scores based on a local sample. In this way, we ensure a fair comparison against the IGHP, who must perform similarly to the local graduate but better than the local student in training. What is most important for us is to establish the lowest cut-off using these data as well since IGHPs who fall below the lowest performing local candidate may be considered for more intense re-training pathways (e.g., registration in a local undergraduate program). The small range of scores that defines the borderline local graduate may represent only the lower limit of candidates requiring intense re-training. Once this range has been established and the quality of the exam is deemed acceptable, we are able to offer it to the IGHP applicant. We then have a stable standard that can be applied to future groups of IGHPs seeking certification.

References

Norcini, J. J. (2003). Setting standards on educational tests. *Medical Education*, *37*(5), 464-469.

De Champlain (2014). Standard setting methods in medical education. In T. Swanwick (Ed), *Understanding medical education: Evidence, theory and practice* (305-316). Oxford: John Wiley & Sons.

Henson, R. K. (2001). Understanding internal consistency reliability estimates: A conceptual primer on coefficient alpha. *Measurement and Evaluation in Counseling and Development*, *34*(3), 177.