Taylor&Francis healthsciences

Setting defensible performance standards on OSCEs and standardized patient examinations

JOHN R. BOULET[1], ANDRÉ F. DE CHAMPLAIN[2] & DANETTE W. MCKINLEY[1]
[1]Educational Commission for Foreign Medical Graduates, Philadelphia, USA;
[2]National Board of Medical Examiners, Philadelphia, USA

SUMMARY *Recently, standardized patient assessments and objective structured clinical examinations have been used for high-stakes certification and licensure decisions. In these testing situations, it is important that the assessments are standardized, the scores are accurate and reliable, and the resulting decisions regarding competence are equitable and defensible. For the decisions to be valid, justifiable standards, or cut-scores, must be set. Unfortunately, unlike the body of research specifically dedicated to multiple-choice examinations, relatively little research has been conducted on standard-setting methods appropriate for use with performance-based assessments. The purpose of this article is to provide the reader with some guidance on how to set defensible standards on performance assessments, especially those that utilize standardized patients in simulated medical encounters. Various methods are discussed and contrasted, highlighting the relevant strengths and weaknesses. In addition, based on the prevailing literature and research, ideas for future studies and potential augmentations to current performance-based standard setting protocols are advanced.*

Introduction

The use of standardized patient (SP) methodology, both for educational and evaluative purposes, has evolved considerably over the past 30 years. The seminal work of Harden, Stevenson, Downie and Wilson (Harden *et al.*, 1975) and Barrows and Abrahamson (Barrows & Abrahamson, 1964), among others, contributed significantly to a broader acceptance and use of objective structured clinical examinations (OSCEs) and standardized patient examinations (SPEs) in medical education. Initially, OSCEs and SPEs were used primarily for formative assessment purposes, that is, within a pedagogical context where the primary aim was to provide feedback to students about their clinical skills and offer remediation, where necessary. Here, the main purpose was to evaluate and teach students under simulated, yet realistic and standardized, conditions. More recently, however, SPEs and OSCEs have become a key component of several high stakes certification and licensure programs (Brailovsky *et al.*, 1992; Whelan, 1999; Medical Council of Canada, 2002). In this context, candidates are usually assessed over a series of simulated patient encounters and individual judgments concerning clinical competence are made. The consequences of these judgments can be quite severe, with failing candidates often being restricted, at least temporarily, from practising medicine.

The use of OSCEs and SPEs for summative assessment purposes has brought several key psychometric issues to the forefront. First, it is well known that the scores or ratings from OSCEs/SPEs can be subject to a wide array of potential, often interacting, sources of measurement error (Swanson & Norcini 1989; van der Vleuten *et al.*, 1991; Boulet *et al.*, 1998; Swanson *et al.*, 1999; Floreck & De Champlain 2001). As a result, much greater care must be taken in the development and validation of test materials, the training of SPs, and the overall administration of the exam. Second, especially where accuracy is crucial, scoring SP examinations can be complex, involving the intertwined issues of the choice of raters and the selection of test forms that they complete (Martin *et al.*, 1996; Hodges *et al.*, 2002). Some groups use physician raters while others use the SPs to capture examinees' performances. Similarly, many groups ask their raters to complete checklists while others have advocated the use of global ratings. Third, for delivery models where the exam is given over time and/or across multiple test sites, it is often necessary to use different versions or forms of the same SPE or OSCE. If the forms are not identical in difficulty or the ability to discriminate along the competence continuum, the resulting scores will not be equivalent, nor will the conclusions that emanate from them (e.g. pass–fail decisions) be directly comparable. Finally, and more importantly, while numerous standard-setting protocols have been developed for multiple-choice or selected-response examinations, there has been relatively little research undertaken to assess how best to set a passing standard on performance-based assessments such as OSCEs or SPEs. Once reliability, scoring and equivalence concerns have been addressed, one critical subsequent step in the validation process is to identify appropriate and defensible standard-setting methods for use in setting the cut-score.

Purpose

The purpose of this paper is to provide an overview of some standard-setting methods that have been proposed for use with performance-based assessments, including OSCEs and SPEs. To aid the reader in understanding the key standard-setting issues, a brief summary of the standard-setting process used for the Educational Commission for Foreign Medical Graduates' (ECFMG®) Clinical Skills Assessment (CSA®) will also be provided. Finally, some future directions for standard-setting studies and associated research are proposed.

Correspondence: John R. Boulet, PhD, Assistant Vice-President, Research and Evaluation, Educational Commission for Foreign Medical Graduates, 3624 Market Street, Philadelphia, PA, 19104, USA. Tel: 215-823-2227; fax: 215-386-3309; email: jboulet@ecfmg.org

ISSN 0142–159X print/ISSN 1466–187X online/03/030245–5 © 2003 Taylor & Francis Ltd
DOI: 10.1080/0142159031000100274

Standard setting for performance-based assessments

The increasing use of performance-based assessments as part of professional certification and licensure activities has necessitated the development of specific standard-setting methodologies. More recently, considerable effort has been aimed at developing and validating standard-setting methods that can be used to make reliable decisions regarding the clinical skills of examinees (Margolis *et al.*, 1998; Kane *et al.*, 1999; Meara *et al.*, 2001; Chinn & Hertz, 2002). For locally administered OSCEs/SPEs (e.g. at a medical school), the purpose of the assessment is usually formative rather than summative. The stakes are typically low, with the assessment being used primarily for training and remediation. As a result, there is often little need to determine a point on the score scale that can be used to classify examinees into defined proficiency cohorts. Where summative assessment is required in this lower stakes setting, relative or norm-referenced standards (e.g. passing the top 90% of examinees, failing students who score below the 5th percentile) are often used. While this strategy may be appropriate in some situations (e.g. when the consequence of making a classification error is not that severe), the additional effort required to set meaningful, absolute standards is not prohibitive, and should definitely be considered.

In a high-stakes context, absolute or criterion-referenced standards are preferred because relative standards may result in passing examinees with little regard for how much they *actually know or can do*. For example, if all the students in a class were exceptionally skilled, the use of norm-referenced standards (e.g. passing the top 90% of the class) would result in failing (i.e. misclassifying) a certain cohort that in reality possesses adequate ability with regard to overall clinical skills. This is certainly at odds with the purpose of a test of competence. Moreover, since relative standards will vary over time with the ability of the examinees being assessed (Norcini & Guille, 2002), the reliability of any competence-based classifications could be questionable. Therefore, if valid measures of competence are desired, it is essential to set standards with reference to some defined performance measure, or criterion.

Criterion-referenced standard-setting procedures

There are a variety of absolute standard-setting procedures that have been applied with SP assessments (Norcini *et al.*, 1993; Ross *et al.*, 1996; Dauphinee *et al.*, 1997; Kent, 2001; Wilkinson *et al.*, 2001; Humphrey-Murto & MacFadyen, 2002). These methods can be broadly classified as either test centered or examinee centered. For test-centered methods, hypothetical decisions based on test content are used to derive a standard. For examinee-centered methods, judgments regarding actual examinee performance are used to determine appropriate cut-points.

Test-centered methods

In general, test-centered methods require subject matter experts (panelists) to make judgments regarding the expected performance of minimally competent examinees on select tasks. The Angoff procedure (Angoff, 1971), and associated modifications, is a prominent example of a test-centered method. It can be, and has been, used to set standards on the history taking and physical examination checklist items that are often used for scoring cases in SP examinations. Here, the panelists are required to make judgments as to the probability of a minimally qualified examinee asking the particular question or performing (correctly) the indicated physical examination maneuver. Equivalently, the panelists could be asked to estimate the number of examinees (say, out of 100) that would receive credit on the particular item. These judgments are then usually averaged over panelists and checklist items to obtain a standard for the case. Unfortunately, this method is laborious, especially when there are multiple checklists. More importantly, since the resulting standard is a mean judgment across items and/or tasks, the use of this method makes the implicit assumption that ratings on tasks are independent. This assumption is often untenable with SPEs and OSCEs given the well-known phenomenon of case specificity (Linn & Burton, 1994). That is, individual SP checklist items are often interrelated within a task. As a result, the raters' judgments are not totally independent, potentially invalidating the use of this method for setting cut-scores on cases, at least where checklists are being used. One alternate strategy is to have the panelists make judgments regarding the number, or percentage, of checklist items that a minimally qualified examinee would obtain credit for. While this may alleviate the problem of checklist item dependencies, and substantially reduce the amount of time required to set standards, the task of deciding how many items constitutes minimally qualified can be quite difficult. For example, some panelists may entertain conjunctive competence rules, deciding that certain combinations of credited items (history questions asked, physical examination maneuvers performed), rather than a total, should be used to ascertain minimal competence. If this were the case, and total checklist performance is used for scoring, then it would possible for individuals judged to be minimally qualified to have lower scores than those judged not to be. As a result, the precision of any derived standard could be compromised. For these and other reasons, examinee-centered methods are usually preferred.

Examinee-centered methods

Examinee-centered methods can also be used to establish standards for SPEs/OSCEs. Instead of providing judgments based on test materials, the panelists are asked to review a series of examinee performances, or suitable proxies, and make judgments about the demonstrated level of proficiency. For example, the task could involve distinguishing qualified from unqualified examinees, or simply identifying 'borderline' performances. For the former, referred to as the contrasting groups method, the intersection of the distribution of qualified and unqualified examinee scores can be used to delimit the standard. Here, it is extremely important that the panelists be oriented to the definition of a 'qualified' examinee. For the latter, referred to as the borderline group method, the mean of the scores for the borderline group, or some other measure of central tendency, would define the cut-point. While 'borderline' can be defined in many ways, panelists can simply identify those performances where they are unsure as to whether the examinee is qualified or unqualified. This subset of performances would constitute the borderline group.

In contrast with test-centered methods, the standard-setting panelists must view actual performances, or reasonable facsimiles (e.g. videotapes, completed checklists) and make a direct judgment concerning competence or qualifications. For subject matter experts, this task is intuitively appealing, as it capitalizes on their clinical experience. However, one potential shortcoming of examinee-centered methods, usually attributable to insufficient training of panelists, is the propensity to attribute ratings based on proficiencies or factors that are not directly targeted by the examination. The attribution of positive ratings based on irrelevant factors (halo effects) is one such phenomenon. This, and other potential sources of rater bias, can be readily addressed by offering extensive training to judges, including the opportunity to complete performance exemplars prior to the actual standard-setting exercise. In addition, it will help to clarify what is being judged and immediately rectify any misinterpretation of the task at hand.

ECFMG Clinical Skills Assessment (CSA)

In this section we present an overview of the process used to set standards on the checklist component of the ECFMG CSA.

The ECFMG CSA was designed to assess the readiness of graduates of international medical schools, individuals who did not attend medical schools in the United States or Canada, to enter graduate medical education programs in the United States. The assessment consists of 10 standardized patient encounters. The candidates have up to 15 minutes to interview and examine each SP. This time frame has been shown to be adequate (Chambers *et al.*, 2000), and is typical of that commonly encountered for other clinical skills assessments that employ SPs. Following the patient encounter, candidates write a patient note that summarizes the pertinent positive and negative history and physical examination findings, provides a list of up to five possible diagnoses, and outlines a diagnostic workup plan. The time allotted for this post-encounter exercise is 10 minutes but individual candidates who leave the patient encounter before the end of the 15 minutes may begin composing the note immediately. The assessment can be considered 'high stakes' in that failing candidates will not meet ECFMG certification requirements and therefore cannot be admitted to graduate medical education programs.

Scoring

Although multiple skills are measured in the ECFMG CSA, for the purposes of explaining standard-setting methods for SP examinations, the data gathering, or checklist, component will be highlighted here. The SPs document each candidate's data-gathering ability in each patient encounter via a case-specific checklist. For the CSA, the checklist items reflect the relevant history taking (Hx) questions and physical examination (PE) maneuvers that a graduating medical student would be expected to ask or perform. The candidate data-gathering score for a case is simply the percentage of checklist items credited.

Setting standards on case checklists

There are a number of steps that must be completed to ensure a successful standard-setting exercise, including choosing the right panelists, adequately defining the performance standard, and selecting an appropriate method for setting the cut-score. These issues are discussed in detail elsewhere (Friedman Ben-David, 2000). For the ECFMG CSA, the panelists were chosen to represent a broad sampling of medical specialties and geographic areas. The panelists were oriented to the performance standard (i.e. readiness to enter graduate medical education ('readiness'), viewed videotapes of case portrayals, and, to familiarize themselves with the difficulty of the task, were required to perform one SP encounter under exam-like conditions. They then used their expert judgment to determine whether select candidate performances, as depicted in listings of actions taken (i.e. completed checklist scores), denoted readiness. They were told to look at the completed checklists and visualize the corresponding candidate in terms of the history questions asked and the physical examination maneuvers correctly performed. For each case (task), the percentage of readiness decisions was calculated. Regression modeling was then used to predict the point on the score scale (based on the sum of checklist items) where the panelists were in maximal disagreement regarding readiness; that is, the score at which 50% of the panelists judged the performance as ready to enter graduate medical education. While the regression modeling can be complex, the method is similar to that of contrasting groups. Essentially, the aim is to find the point in the score distribution that maximally discriminates between candidates who are ready, or qualified, and those who are not ready to enter graduate medical education. Once the individual case standards are set, they can be averaged to derive an examination-level standard.

While the method described above is relatively simple, it provides panelists with an opportunity to use their expert judgment in determining the cut-score. This process is reliant upon the panelists' ability to envision the candidate whose scores were depicted on the checklist. Ideally, videotape review could also be used. However, given the length of the standardized patient encounters (15 minutes), the number of cases and the requirement for panelist judgments across the performance continuum, this strategy would be impractical and costly, at least for large-scale assessments.

Future directions

Recently, several standard-setting studies have been conducted specifically for SPEs or OSCEs (Kent, 2001; Wilkinson *et al.*, 2001; Humphrey-Murto & MacFadyen, 2002; Martin & Jolly, 2002). Although promising, the amount of standard-setting research conducted specifically in this domain pales in comparison with the body of research devoted to setting standards for multiple-choice examinations. Given the increased use of OSCEs/SPEs for high-stakes decisions, it is hoped that the number of studies aimed at enhancing standard-setting methods and protocols with SPEs and OSCEs will continue to grow. From a technological perspective, the use of the Internet (e.g. online performance rating) would be of great benefit and might minimize the costs associated with convening panelists. One could also simulate performance samples, highlighting specific interviewing and examination patterns, thereby eliminating the need to collect data from live assessments. From a methodological perspective, there are several key areas that appear to warrant further investigation. First, the setting of a meaningful standard by any panel requires a clear definition of the target examinee group (e.g. master/non-master, competent/not competent, ready/not ready, etc.). In the absence of a clear definition, each individual panelist will use his or her own beliefs, potentially compromising the standard-setting process. Consequently, it might be useful to gather relevant data on performance criteria, either from surveys, focus groups or analyses of practice patterns, as a means of formulating operational definitions of the targeted skills. This, in turn, could lead to a higher level of consistency in judgments and enhance the defensibility of any derived cut-point. Second, there is often no defined protocol for training panelists, or identifying aberrant rating patterns. Orientation to the assessment, including completion of some of the exercises by panelists under exam-like conditions, might prove to be a valuable component of the standard-setting process. Likewise, providing rapid, tailored feedback to panelists, especially during the standard-setting exercises, will probably enhance the validity of the resulting standard. Third, the application of innovative statistical techniques to model a standard seems worthwhile (De Champlain *et al.*, 2001). Here, it would also be informative to provide an estimate of the amount of error associated with a given cut-score. Generalizability theory (Brennan, 2001) could be used to help design standard-setting studies by determining conditions (e.g. number of panelists, number of performance samples) that would minimize sources of measurement error and result in a more defensible cut-score.

For OSCEs/SPEs, there has also been relatively little work done in the area of standards validation. This line of research seems appropriate in that the true measure of any standard is its ability to validly discriminate between those who are competent and those who are not. For certification or licensure decisions, it is important that decisions based on passing scores achieve the purposes of the organization while avoiding any serious negative consequences. For example, for certification exams such as the ECFMG CSA, a central goal is to protect the public from poorly trained physicians. If the standard is set too low, some candidates might pass the CSA without adequate clinical skills. Conversely, if the standard is set too high, and access to the residency programs is unduly restricted, then the profession—and public—will not be served. In either case, it is extremely important that research be conducted to investigate the appropriateness of the standard. Validation studies can take many forms, including comparing the results of different standard-setting methods, creating parallel standard-setting panels and contrasting results, gathering procedural evidence pertaining to the standard setting process (e.g. surveys of panelists) and, most importantly, investigating the skills and knowledge that characterize the passing candidates (Whelan *et al.*, 2001). While the adequacy of criterion measures is often questioned, it is still essential to do follow-up studies of passing candidates, documenting their abilities relative to the standard.

Conclusion

While defensible standard-setting methods are available for use with constructed-response examinations, there has been comparatively less research focused on methods applicable to performance-based assessments. Although the usefulness of some standard-setting methods has been evaluated with SPEs and OSCEs, given the dearth of research conducted in this area, conclusions on which methods are preferable are tenuous at best. The choice of standard-setting method to be used for an SPE/OSCE will, to some extent, depend on the purpose of the assessment and the availability of resources to conduct the exercise. While test-centered approaches have often been used for selected-response examinations, the limitations associated with these approaches often preclude their use with OSCEs. Also, panelists often express a level of difficulty in making judgments about the checklist or station performance of a hypothetical group of minimally competent examinees. In contrast, examinee-centered methods are more intuitively appealing to panelists given the greater ease with which they can make judgments about specific performances. Additionally, panelists find the process and results more credible because the standard is derived from judgments based on the actual test performances. This probably accounts for a growing use of examinee-centered standard-setting methods to set cut-scores on OSCEs and SPEs (Kane *et al.*, 1999).

Although examinee-centered standard-setting procedures may be preferred for SPEs/OSCEs, technological and methodological advancements may lead to more efficient protocols and, eventually, more defensible standards. The increased use of computers, both for rating performances and for providing rater feedback, is certainly indicated. Likewise, embracing new statistical techniques and estimation procedures may provide for more consistent and reproducible standards.

Regardless of the method used to set standards on SPEs/OSCEs, it is imperative that data be collected both to support the system that was used and to establish the credibility of the standard. Where SPEs/OSCEs are used for credentialing decisions, the responsible organization must ensure that passing scores achieve the intended purposes (e.g. public protection) and avoid any serious negative consequences.

Notes on contributors

John Boulet is the Assistant Vice-President for Research and Evaluation at the Educational Commission for Foreign Medical Graduates (ECFMG). He is responsible for research activities related to clinical skills and international medical education.

André De Champlain, PhD, is Program Director of Global Initiatives at the National Board of Medical Examiners. His professional interests include cross-cultural assessment and clinical skills examinations.

Danette McKinley is a Research Scientist at ECFMG. She is actively involved in psychometric research, including standard setting, scale construction and score validation.

References

Angoff, W.H. (1971) Scales, Norms and Equivalent Scores, in: R.L. Thorndike (Ed.) *Educational Measurement, American Council on Education*, pp. 508–600 (Washington, DC, American Council on Education).

Barrows, H.S. & Abrahamson, S. (1964) The programmed patient: a technique for appraising student performance in clinical neurology, *Journal of Medical Education*, 39, pp. 802–805.

Boulet, J.R., Friedman Ben-David, M., Hambleton, R.K., Burdick, W.P., Ziv, A. & Gary, N.E. (1998) An investigation of the sources of measurement error in the post-encounter written scores from standardized patient examinations, *Advances in Health Sciences Education*, pp. 89–100.

Brailovsky, C.A., Grand'maison, P. & Lescop, J. (1992) A large-scale multicenter objective structured clinical examination for licensure, *Academic Medicine*, 67(10 Suppl.), pp. S37–S39.

Brennan, R.L. (2001) *Generalizability Theory* (New York, Springer-Verlag).

Chambers, K.A., Boulet, J.R. & Gary, N.E. (2000) The management of patient encounter time in a high-stakes assessment using standardized patients, *Medical Education*, 34(10), pp. 813–817.

Chinn, R.N. & Hertz, N.R. (2002) Alternative approaches to standard setting for licensing and certification examinations, *Applied Measurement in Education*, 15(1), pp. 1–14.

Dauphinee, W.D., Blackmore, D., Smee, S., Rothman, A.I. & Reznick, R. (1997) Using the judgments of physician examiners in setting the standards for a national multi-center high stakes OSCE, *Advances in Health Sciences Education*, 2, pp. 201–211.

De Champlain, A.F., Margolis, M.J., Macmillan, M.K. & Klass, D.J. (2001) Predicting mastery level on a large-scale standardized patient test: a comparison of case and instrument score-based models using discriminant function analysis, *Advances in Health Sciences Education*, 6(2), pp. 151–158.

Floreck, L.M. & De Champlain, A.F. (2001) Assessing sources of score variability in a multisite medical performance assessment: an application of hierarchical linear modeling, *Academic Medicine*, 76(10 Suppl.), pp. S93–S95.

Friedman Ben-David, M. (2000) AMEE Guide No. 18: Standard setting in student assessment, *Medical Teacher*, 22(2), pp. 120–130.

Harden, R.M., Stevenson, M., Downie, W.W. & Wilson, G.U. (1975) Assessment of clinical competence using objective structured examination, *British Medical Journal*, 1, pp. 447–451.

Hodges, B., Mcnaughton, N., Regehr, G., Tiberius, R. & Hanson, M. (2002) The challenge of creating new OSCE measures to capture the characteristics of expertise, *Medical Education*, 36(8), pp. 742–748.

Humphrey-Murto, S. & Macfadyen, J.C. (2002) Standard setting: a comparison of case-author and modified borderline-group methods in a small-scale OSCE, *Academic Medicine*, 77(7), pp. 729–732.

Kane, M.T., Crooks, T.J. & Cohen, A.S. (1999) Designing and evaluating standard-setting procedures for licensure and certification tests, *Advances in Health Sciences Education*, 4, pp. 195–207.

Kent, A. (2001) Setting standards for an objective structured clinical examination: the borderline group method gains ground on Angoff, *Medical Education*, 35, pp. 1009–1010.

Linn, R.L. & Burton, E. (1994) Performance-based assessment: implications of task specificity, *Educational Measurement: Issues and Practice*, 13, pp. 5–8.

Margolis, M.J., De Champlain, A.F. & Klass, D.J. (1998) Setting examination-level standards for a performance-based assessment of physicians' clinical skills, *Academic Medicine*, 73(10 Suppl.), pp. S114–S116.

Martin, I.G. & Jolly, B. (2002) Predictive validity and estimated cut score of an objective structured clinical examination (OSCE) used as an assessment of clinical skills at the end of the first clinical year, *Medical Education*, 36(5), pp. 418–425.

Martin, J.A., Reznick, R.K., Rothman, A., Tamblyn, R.M. & Regehr, G. (1996) Who should rate candidates in an objective structured clinical examination? *Academic Medicine*, 71(2), pp. 170–175.

Meara, K.C., Hambleton, R.K. & Sireci, S.G. (2001) Setting and validating standards on professional licensure and certification exams: a survey of current practices, *CLEAR Exam Review*, Summer, pp. 17–23.

Medical Council of Canada (2002) *Qualifying Examination Part II, Information Pamphlet* (Ottawa, Ontario, Canada, Medical Council of Canada).

Norcini, J. & Guille, R. (2002) Combining tests and setting standards, in: G.R. Norman, C.P.M. Van Der Vleuten & D.I. Newble (Eds) *International Handbook of Research in Medical Education* (Part Two), pp. 811–834 (Dordrecht, The Netherlands, Kluwer Academic).

Norcini, J.J., Stillman, P.L., Sutnick, A.I., Regan, M.B., Haley, H.L.A., Williams, R.G. & Friedman, M. (1993) Scoring and standard setting with standardized patients, *Evaluation and the Health Professions*, 16(3), pp. 322–332.

Ross, L.P., Clauser, B.E., Margolis, M.J., Orr, N.A. & Klass, D.J. (1996) An expert-judgment approach to setting standards for a standardized-patient examination, *Academic Medicine*, 71(10, Suppl.), pp. S4–S6.

Swanson, D.B., Clauser, B.E. & Case, S.M. (1999) Clinical skills assessment with standardized patients in high-stakes tests: a framework for thinking about score precision, equating, and security, *Advances in Health Sciences Education*, 4, pp. 67–106.

Swanson, D.B. & Norcini, J.J. (1989) Factors influencing reproducibility of tests using standardized patients, *Teaching and Learning in Medicine*, 1(3), pp. 158–166.

Van Der Vleuten, C.P., Norman, G.R. & De Graaff, E. (1991) Pitfalls in the pursuit of objectivity: issues of reliability, *Medical Education*, 25(2), pp. 110–118.

Whelan, G.P. (1999) Educational commission for foreign medical graduates: clinical skills assessment prototype, *Medical Teacher*, 21(2), pp. 156–160.

Whelan, G.P., Mckinley, D.W., Boulet, J.R., Macrae, J. & Kamholz, S. (2001) Validation of the doctor–patient communication component of the Educational Commission for Foreign Medical Graduates Clinical Skills Assessment, *Medical Education*, 35(8), pp. 757–761.

Wilkinson, T.J., Newble, D.I. & Frampton, C.M. (2001) Standard setting in an objective structured clinical examination: use of global ratings of borderline performance to determine the passing score, *Medical Education*, 35(11), pp. 1043–1049.