Who should set the standards for surgical assessments?

In light of national policy and educational theory, this article addresses the question of who should set the standards for surgical assessments, highlighting the multidisciplinary nature of those currently (and potentially) involved.

Surgical education in the UK has been influenced by major changes over the last two decades. These include the establishment of Calman reforms to training (Calman, 1993), the New Deal to improve working conditions, Modernising Medical Careers to restructure training (Department of Health, 2003), and the European Working Time Directive to limit working hours. This has lead to key reforms in the delivery of surgical education and patient care, as these are inextricably linked in the NHS.

In 1889 William Halsted introduced a German-style residency training system at the Johns Hopkins Hospital, Baltimore, with an emphasis on graded responsibility (Carter, 1952). In the USA, this system remains the cornerstone of surgical training. However, advances in educational theory, in addition to mounting pressures in the clinical environment, have advocated a change in this traditional approach to the teaching and acquisition of surgical skills, both technical and non-technical, as defined by the Intercollegiate Surgical Curriculum Programme in the UK.

Modern surgical education thus aims to cultivate surgeons with the appropriate operative and clinical skills, knowledge depth, professional values (such as team working and communications skills) and judgement to be members of the multiprofessional surgical team (Intercollegiate Surgical Curriculum Programme, 2010).

Mr H Sadideen *is Registrar in the Department of Plastic and Reconstructive Surgery, Birmingham Children's Hospital, Birmingham B4 6NH and* **Professor R Kneebone** *is Professor of Surgical Education in the Department of Surgery and Cancer, Imperial College London, London*

Correspondence to: Mr H Sadideen (hazim.sadideen@doctors.org.uk)

However, it can be difficult to establish when a trainee is competent in performing these skills. There should be a measurable outcome that can be assessed. Describing assessment methods is far beyond the scope of this article, but key theories will be touched upon to highlight their importance and role in helping to establish cut-off points for trainees, distinguishing the 'competent' from the 'non-competent'.

Furthermore, an integral question is: 'Who defines these cut-off points?' Given the drive to improve surgical education in the current environment, it follows suit that there should be a drive to choose appropriate assessment methods and judges. Hence the notion of who should set surgical assessment standards could not be more topical. This article will address this question in light of educational theory and national policy in the UK, highlighting the current and potential multidisciplinary nature of standard setting.

Surgical assessments

Surgical assessments are a core component of undergraduate and postgraduate surgical education.

In the UK, undergraduate medical education standards have been the responsibility of the General Medical Council, as set out in the Medical Act 1983. Specifically, Tomorrow's Doctors identified certain skills students were to attain by the time of graduation, which universities and the NHS are expected to comply with. It also paid particular attention to the standards for the delivery of assessments. Medical schools in the UK construct their own exams for students in line with the undergraduate curriculum and General Medical Council standards. This increased awareness of assessments was highlighted in February 2011 when a press release from the General Medical Council outlined a supplementary document 'Assessment in undergraduate medical education' setting out different methods which a medical school could use to assess medical students and providing advice to medical schools on setting standards and marking examinations (General Medical Council, 2011).

Before April 2010, the Postgraduate Medical Education Training Board set standards for postgraduate surgical education in the UK. The Postgraduate Medical Education Training Board has now merged with the General Medical Council, so that the General Medical Council now also sets the standards for postgraduate education, potentially creating a more seamless and consistent approach to education and training throughout doctors' careers. Thus, the General Medical Council is effectively now responsible for surgical education from admission to medical school, through postgraduate training, to continued practice until retirement; hopefully this will help to ensure consistency of expectations and standards. Through the merger, the General Medical Council acquired the legal functions formerly performed by Postgraduate Medical Education Training Board in relation to the regulation of specialty training, including standard setting and quality assurance of the delivery of specialty training against those standards (General Medical Council, 2010).

In line with the above and with General Medical Council definitions, the Intercollegiate Surgical Curriculum Programme (which highlights the curriculum for all nine postgraduate surgical specialties) highlights an assessment as a 'systematic procedure for measuring a trainee's progress or level of achievement, against defined criteria to make a judgement about a trainee'.

The purpose of a surgical assessment is thus multi-fold. It can define whether a trainee meets the standards of compe-

tence required at various stages in the curriculum, provide comprehensive and systematic feedback as part of the learning cycle, and address the domains of good medical practice laid down by the General Medical Council and professional education bodies.

Assessment tools

The Postgraduate Medical Education Training Board (2007) guidance document acknowledges that optimizing any assessment tool or programme is about balancing the components of the utility index. The utility index described by van der Vleuten (1996) serves as an excellent framework for assessment design and evaluation (Schuwirth and van der Vleuten, 2006). This was originally composed of five indices: reliability, validity, educational impact, cost efficiency and acceptability. Feasibility was advocated as a sixth component by the Postgraduate Medical Education Training Board, and rightly so (although technically it was implicit within cost efficiency and acceptability).

The choice of assessment instruments and aspirations for high validity and reliability are undoubtedly limited by the constraints of feasibility, such as resources to deliver the tests, and its acceptability to the trainees. Hence, the relative importance of the components of the utility index for a given assessment will depend on both the purpose and nature of the assessment system. For example, a high stakes examination, such as the Membership of the Royal College of Surgeons exam, on which progression into higher specialist training is dependent, will need high reliability and validity, and may focus on this at the expense of educational impact. In contrast, an assessment which focuses largely on providing a trainee with feedback to inform his/her own personal development planning would focus on educational impact, with less of an emphasis on reliability.

Trainees are often deemed competent to perform procedures based on non-validated, subjective or objective global assessments and case logs (Reznick, 1993). However, such forms of assessment do not provide direct evidence of a trainee's competence at performing a skill, because case logs lack content validity and many forms of global assessment have poor reliability and unknown validity (Jelovesek et al, 2010).

Therefore, the challenge continues for surgical educators to establish more robust methods of determining surgical competence to truly reflect effective, safe surgical care. Competency-based learning begins with standard setting.

Setting standards

In its most essential form, standard setting refers to the process of establishing one or more cut-off scores on a test. It has been postulated that a more accurate term would be 'standard recommending' in that the role of the panels (e.g. professional bodies, educational experts or patients) engaging in a process is technically to provide informed guidance to those actually responsible for the act of setting, approving or implementing any cut-off scores. However, for the sake of argument, it appears that such 'recommending' by default becomes 'setting' when incorporated into a curriculum and used for assessment purposes (Cizek, 2006, 2007).

Often the difficulty in assigning scores lies in differentiating those students or trainees on the borderline between passing and failing. The purpose of standard setting is to devise some meaningful way to identify the point of differentiation, i.e. whether the knowledge and/or skills assessed are 'enough'.

Standard setting really should be practical rather than philosophical and so the Postgraduate Medical Education Training Board published a guidance document entitled 'Developing and maintaining an assessment system' (Postgraduate Medical Education Training Board, 2007). Since the priority was to ensure that passing standards were set with due diligence and at sufficiently robust levels to ensure patient safety, it advised assessors to choose appropriate methods they would be happy with. It also advocated the importance of assessors being appropriately trained for their involvement in assessment.

Methods of standard setting

Standard setting can be broadly categorized as absolute or relative (criterion or norm referencing respectively).

Criterion referencing (absolute standard setting) provides a clear definition of what trainees should be able to perform, providing a standard to indicate a level of performance deemed as competent. In effect, this is responsive to the subject matter being taught, allowing the trainee and trainer to clearly identify and align capabilities. For most high-stakes competence assessment, such as the Membership of the Royal College of Surgeons exam in the UK, or the United States Medical Licensing Exams, there is inherent appeal in developing a criterion-referenced or absolute standard for passing (Schindler et al, 2007).

Norm referencing (relative standard setting) describes an individual's performance relative to his or her position within a group. A surgical trainee, for example, is judged by comparison to the scores achieved by colleagues at the same level for the same performance. Although this is the most common method of referencing, it aims to rank trainees and allows trainees to be compared with one another. However, it may not provide a clear assessment of the trainee's abilities, because there will always be a fixed number who fail. Moreover, norm referencing encourages competition, not cooperation, and can appear somewhat unstable, as it will shift according to the performance of the norm group (Nungester et al, 1991).

The Angoff method was the first of the absolute methods for standard setting and has the longest history of successful use, even in high-stakes testing situations such as the United States Medical Licensing Exams (Angoff, 1971; Jelovesek et al, 2010). The modified Angoff method allows experts to have actual performance data from a scale as an additional source of information to help inform their decision. Most standard setting in medicine uses it, as it is a reasonable, practical and defensible standard-setting method (Kaufman et al, 2000). It asks participants to determine the borderline trainee's performance on specific items within the different scales. This is often referred to as a test-centred, or item-centred, method, as it requires experts to make judgments about the expected performance of borderline competent examinees on selected items (Cusimano, 1996).

Such item-centred methods seem useful for setting a pass-fail score on surgical skill assessments.

However, this method does not appear to be ideal for assessing performance in an entire surgical clerkship (undergraduates) or rotation (postgraduates). In this case, a method that combines elements of a relative standard with other elements of an absolute standard may be more appropriate (Hofstee, 1983). At undergraduate level, the Hofstee method has been used and published to set a standard for the overall grade for a surgery clerkship for American undergraduates (Schindler et al, 2007).

The ideal standard setter

It has been noted that a crucial characteristic of each standard-setting method is that it depends on value judgments of experts. In effect, this means that even if the most rigorous standard-setting method was followed meticulously, it would be somewhat arbitrary. Different standard-setting methods and different judges will inevitably produce different passing scores; there is no gold standard (Downing et al, 2006). Thus, the terms credible or defensible, rather than valid, are more appropriate when discussing of a cut-off point because a performance standard is by nature arbitrary and subject to educational or social judgments.

Norcini and Guille (2002) identify that credible standards share three important characteristics. The first is most relevant here, which states they are set by appropriate numbers and types of 'judges'. This reiterates the importance of carefully selecting those involved in standard setting. Downing et al (2006) further describe the necessary qualities for judges involved in standard setting. They advocate that assessors should be content experts, know the target population, and understand the task and the assessment tool. In addition, they should be fair-minded, willing to follow directions and give their full attention to the process. It is important to avoid bias by identifying demographically diverse judges.

The range of standard setters

Perhaps a key point here involves who is to set the surgical standard for high-stake exams. In the UK, the Specialist Advisory Committee is a subcommittee of the Joint Committee on Surgical Training which oversees training in all surgical specialities. Each subspecialty (e.g. plastic surgery) has its own specialist advisory committee which oversees higher surgical training. The specialist advisory committee sets standards on behalf of the certifying authority, the General Medical Council. The specialist advisory committee is usually made up of consultants, some of whom are clinical supervisors, and experts in the specialty, who can highlight aspects of the curriculum that must be met for trainees to progress.

The key to defending acceptable standards thus lies in the choice of credible judges and in the use of a systematic approach to collecting their judgements. If a surgical skill is to be assessed on a national scale, such as that of a specialty fellowship exam, for instance the FRCS(Plast), which is the plastic surgery 'exit' exam in the UK, the experts selected to assess standards are 'calibrated' to have a realistic expectation of actual trainee performance. This is important because the exam (in this example) tests numerous sub-specialties within the field of plastic and reconstructive surgery (e.g. hands, burns, breast, congenital deformities, head and neck, and aesthetics). Experts chosen thus have a clear understanding of the assessment, task and trainee involved in order to avoid the implications of the ha-ha effect, described by Kneebone (2009); a metaphor to account for the differing perspective between expert and novice. For example, an expert's perception may be radically different from a novice's, and a novice may struggle with difficulties that the expert can no longer see. Misrepresentations in such scenarios may set cut-off scores that are either too high, resulting in higher failure rates on the assessments, or vice versa.

The communications aspect of the exam (e.g. breaking bad news or gaining informed consent) may benefit from direct patient input into standard setting. This is because clinically patients represent the target population who are on the receiving end of the consultation or treatment, and can thus provide the standard-setting panel with direct feedback as to what is essential from a patient perspective. Patient involvement is more common in the undergraduate curriculum, and learning from these experiences and its application to the postgraduate forum is essential, especially in examining the non-technical skills domain. This is extremely important in the current era of the patient-centred approach to management. Involving patients will require careful, unbiased selection.

Most experts agree that because there is no 'true' standard, many standard setters with appropriate qualifications should be involved. Therefore, before institutions or advisory bodies such as the specialist advisory committee make formal decisions on competence, it may be appropriate to repeat assessments using a selection of assessors in order to establish reliable standards to measure against.

It appears that different assessment types will ideally require different standard setters. This can best be explained by the role of workplace-based assessments. To monitor a trainee's progress, the Intercollegiate Surgical Curriculum Programme has encouraged trainees to undertake workplace-based assessments such as surgical direct observed procedures and case-based discussions. The information acquired during a workplace-based assessment not only provides immediate feedback for the trainee and trainer, but can also provide evidence of progression and therefore contribute evidence suitable for recording in the trainee's learning portfolio. This can then be compared to the agreed outcomes set by the clinical (or educational) supervisor and trainee in the educational agreement.

Therefore, educational supervisors will play a role of being both tutors and assessors. By being regularly involved with the assessment and feedback process, and with adequate training, clinical supervisors are ideal to set standards to monitor a trainee's progression during a specific rotation. Having said that, in the current climate, not all educational supervisors are trained to set standards and assess trainees, and hence this will need to be targeted to optimize assessment modalities and outcomes.

It must be mentioned that there is a perception that consultants display a varying enthusiasm for undertaking workplace-based assessments as there is an appreciable amount of time required to

undertake them. A systematic review also highlighted that although subjective reports on the educational impact of the use of alternative workplace-based assessments (e.g. case-based discussions and surgical direct observed procedures) are positive, there is no evidence that they lead to improvement in performance (Miller and Archer, 2010).

Conclusions

Standard setting in surgical education is challenging. Standards are inherently arbitrary and in certain instances it can be difficult to agree what constitutes the pass/fail cut-off. There are currently no universally accepted standards for some aspects of surgical education, such as the non-technical skills domain.

As surgical education moves towards competency-based models, particular attention to standard setting is crucial. Standard setting can be applied to surgical assessments to establish minimum cut-off scores in order to determine surgical competence. Experts should be carefully chosen and rigorous methods should be used to determine the most appropriate method for different types of assessment. It only seems prudent to involve appropriate standard setters, and in many instances a range of standard setters, for different assessments.

The General Medical Council as a professional body and its expert panel comprising multidisciplinary members is key in overseeing surgical undergraduate and postgraduate education. Regulatory bodies, education experts, clinical supervisors and patients all have a role to some extent in standard setting. Querying stakeholders and revisiting assessment strategies and standard setters will help ensure that 'judges' and content remain appropriate in the future. A credible standard as per both Norcini and Guille's (2002) and Downing et al's (2006) descriptions of characteristics are essential.

Training institutions will need to use objective evidence to reassure health-care regulators and the public that trainees have reached a minimum level of competence during their training. It must be recognized that introducing appropriate assessment tools with high utility indices and selecting ideal standard setters from a range of backgrounds for appropriate assessments will require a multidisciplinary team effort. The knowledge that surgical competence is determined using systematic approaches of standard setting rather than traditional, unsystematic approaches will ensure a robust system for assessments is in place, deeming trainees as competent before progressing. This will help reassure trainers, colleagues, patients and the public that future surgeons have graduated as safe and competent. Appropriate standard setters should thus be used to augment current assessment strategies, in order to enhance the learning experience, ensure trainee competence and ultimately improve patient care. **BJHM**

Conflict of interest: none.

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

Calman K (1993) *Hospital doctors- training for the future.* Department of Health, London

Carter BN (1952) The fruition of Halsted's concept of surgical training. *Surgery* **32**: 518–27

Cizek GJ (2006) Standard setting. In: Downing SM, Haladyna TM, eds. *Handbook of Test Development.* Lawrence Erlbaum, Mahwah, NJ: 225–58

Cizek GJ (2007) What is standard setting? In: Cizek GJ, Bunch MB, eds. *Standard Setting: A Guide to establishing and Evaluating Performance Standards on Tests.* Sage, Thousand Oaks, CA: 13–34

Cusimano MD (1996) Standard setting in medical education. *Acad Med* (suppl 10): S112–S120

Department of Health (2003) *Modernising Medical Careers. The Response of the Four UK Health Ministers to the Consultation on 'Unfinished' Business – Proposals for Reform of the Senior House Officer Grade.* Department of Health, London

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**: 50–7

General Medical Council (2010) *Standards for Curricula and Assessment Systems.* www.gmc-uk.org/Standards_for_Curricula__Assessment_Systems.pdf_31300458.pdf (accessed 16 June 2011)

General Medical Council (2011) GMC launches new supplementary advice for medical schools. www.gmc-uk.org/news/8828.asp (accessed 18 June 2011)

Hofstee WK (1983) The case for compromise in educational selection and grading. In: Anderson SB, Helmick JS, eds. *On educational testing.* Jossey-Bass, San Francisco: 109–27

Intercollegiate Surgical Curriculum Programme (2010) ISCP/GMP Blueprint. www.iscp.ac.uk/static/public/overarching_blueprint2010.pdf (accessed 15 November 2011)

Jelovesek E, Walters MD, Corn A et al (2010) Establishing cutoff scores on assessments of surgical skills to determine surgical competence. *Am J Obstet Gynecol* **203**(1): 81.e1–81.e6

Kaufman DM, Mann KV, Muijtjens AM, van der Vleuten CP (2000) A comparison of standard-setting procedures for an OSCE in undergraduate medical education. *Acad Med* **75**: 267–271

Kneebone R (2009) Perspective: Simulation and transformational change: the paradox of expertise. *Acad Med* **84**(7): 954–7

Miller A, Archer J (2010) Impact of workplace based assessment on doctors' education and performance: a systematic review. *BMJ* **341**: c5064

Norcini JJ, Guille RA (2002) Combining tests and setting standards. In Norman G, van der Vleuten C, Newble D, eds. *International Handbook of Research in Medical Education.* Kluwer Press, Dordrecht: 811–34

Nungester RJ, Dillon GF, Swanson DB, Orr NA, Powell RD (1991) Standard-setting plans for the NBME comprehensive part I and part II examinations. *Acad Med* **66**: 429–33

Postgraduate Medical Education Training Board (2007) *Developing and maintaining an assessment system - a PMETB guide to good practice.* www.gmc-uk.org/Assessment_good_practice_v0207.pdf_31385949.pdf (accessed 25 June 2011)

Reznick RK (1993) Teaching and testing technical skills. *Am J Surg* **165**: 358–61

Schindler N, Corcoran J, DaRosa D (2007) Description and impact using a standard-setting method for determining pass/fail scores in a surgery clerkship. *Am J Surg* **193**(2): 252–7

Schuwirth L, van der Vleuten C (2006) *How to design a useful test: the principles of assessment.* ASME, Edinburgh

van der Vleuten C (1996) The assessment of professional competence: developments, research and practical implications. *Adv Health Sci Edu* **1**: 41–67

KEY POINTS

- The General Medical Council is responsible for undergraduate and postgraduate surgical education.
- Standard setting for surgical assessments is complex; it requires defining 'cut-off points' for trainees, distinguishing the competent from the non-competent.
- Optimizing assessment tools requires balancing the components of the 'utility index'.
- The key to defending acceptable standards lies in the choice of credible judges and in the use of a systematic approach to collecting their judgements.
- A credible standard as per both Norcini and Guille's and Downing's descriptions of characteristics is essential.
- Different assessment types ideally require different standard setters.