Integrating assessment innovations in medical education

The role assessment plays in driving learning should not be underestimated. An understanding of assessment processes is clearly important for all those involved in teaching, as well as examiners and candidates. So how do we best assess learning outcomes? This article provides a balanced foundation in the theory of modern medical education for those involved in assessment at any level.

Progression depends upon passing agreed standards of summative assessment (examination of learning). The examiner is acting as a gatekeeper and must be diligent in upholding the highest standards in order that the qualification has value. Examinations must not only be valid and reliable in their ability to discriminate between different degrees of performance, but they must also be transparent and fair. Candidates' extenuating circumstances should be taken into account, and a rigorous appeals process needs to be available.

To provide familiarity formative assessments (practice examinations) should be made available which are representative of summative assessments in terms of form, breadth and depth of any examination. These can be both online mock and online past papers, so that candidates can assess the adequacy of their own preparation and progress in advance of examinations. Examinations should also be viewed as additional opportunities for learning, and wherever possible assessments should be constructed to facilitate learning. Furthermore, assessment questions should be blueprinted (i.e. cover a representative sample of learning objectives), and where possible be integrated and applied to demonstrate higher order thinking (Bloom, 1956; e.g. within clinically relevant contexts).

Types of question

There are many types of question that can be used in assessment. These include written (essay, long answer, short answer), oral, and questions requiring practical demonstration (i.e. objective structured examinations), as well as computer-marked question types. Mixed format-type assessments are objectively marked and provide a plurality of question types to provide a balanced approach within a curriculum, affording the opportunity to match learning objectives with the most appropriate form of assessment, while also limiting the potential for advantaging one learning style at the expense of another.

Everything that can be done on paper (single best answer, extended matching, multiple extended matching, multiple response, ranking, matrix, fill-in-the-blank, labelling, and dichotomous such as true/false/abstain) can also be examined online, and more: high quality radiographs and colour photographs can be coupled with hot-spot question types which lack the cues associated with paper equivalents. If desired, questions can be presented unidirectionally and concepts or cases elaborated, by being able to reveal and incorporate the answers to preceding questions without the candidate being able to go back and change his/her earlier answers. Unique calculation questions are possible with candidates starting with randomly generated numbers within a predetermined range for different parameters that form the variables within a pre-programmed equation.

Objective structured examinations can be of various types, including:

- OSCEs = clinical (e.g. testing clinical skills)
- OSPEs = practical (e.g. anatomy spotters)
- OSSEs = seminar (e.g. testing communication skills).

Quality control and item analyses

Questions need thorough proof checking before any examination, followed by rigorous item analyses afterwards to determine the difficulty (facility) and discrimination of each question. The whole process of vetting examination question papers and addressing external examiner comments can now be dealt with online. Statistics on performance can be monitored in real time during the examination, and presented by question, individual candidate or class cohort. Sequential item analyses on reused questions enables criterion-referenced comparison of performance between different cohorts.

Item analyses can be calculated as:

- Facility value = fraction of all candidates sitting the exam who answered the question correctly (usually expressed as a percentage)
- Discrimination value = fraction of top scoring candidates who answered the question correctly minus the fraction of bottom scoring candidates who answered the question correctly (e.g. take top and bottom thirds of the cohort according to their overall performance in the exam, then take the fraction of the best third who answered correctly minus the fraction of the worst third who answered correctly; varies from +1.0 to -1.0).
- A question with a facility value of less than 50% and a discrimination value of less than 20% should be removed (N.B. a question is inadequate if more than half the candidates fail to get the correct answer and the question does not discriminate between the best and worst candidates).

Standard setting

The thresholds of achievement within assessment need to be rigorously evaluated. Some form of standard setting must be used (Norcini, 2003), whereby examiners work together to qualitatively judge and agree upon the difficulty and importance

Dr Steven Burr *is Deputy Director of Teaching in the School of Biomedical Sciences, Faculty of Medicine and Health Sciences, Queen's Medical Centre, University of Nottingham, Nottingham NG7 2UH*

of all examination questions. Standard setting formulae then enable quantification of the 'cutscore' (i.e. the pass mark), to define the minimally acceptable candidate for that particular examination. Pass marks can then vary between different examinations.

Item-centred approaches to standard setting include:

- Criterion-referenced = absolute; competent or not
- Norm-referenced = relative; performance compared with cohort
- Angoff's method = derived by estimating the proportion of borderline candidates who would answer each item correctly
- Ebel's method = derived by assigning items to categories of difficulty (easy/medium/hard) and relevance (essential/important/'nice-to-know') and then estimating the proportion of items in each category that borderline candidates would answer correctly
- Retrospective item analyses = uses the known past performance of used questions to predict the pass mark, or to change the balance of questions in a paper to reach a desired pass mark (useful when attempting to retain fixed thresholds, such as university honours degree grades where the pass mark is usually 40% and first class threshold 70%).

Marking

For written summative assessments, marking involves making categorical judgments relative to written marking scheme descriptors, and must be checked by more than one examiner.

Methods of checking the marking of written work include:

- Moderation = review to evaluate consistency and standard
- Second marking = single-blind
- Double marking = double-blind.

For computer-based summative assessments, marking can be automated with appropriate mark modification depending on the type of questions involved. Dichotomous questions will be answered correctly 50% of the time by chance. For dichotomous questions, the adoption of negative marking nullifies the possibility of passing by random responses, and provides the option of abstaining to replace blind guessing.

Negative marking implicitly tests confidence, and fosters the development of an appropriate professional attitude that does not encourage guessing blindly, by penalizing risk-taking behaviour. In contrast, with other question types 'correction for guessing' can be used. With correction for guessing the average number of marks that would be gained by randomly responding to every question are deducted from each candidate's overall score, and also from the total number of marks available for the paper; the candidate's revised mark is then expressed as a percentage of the revised paper mark. Correction for guessing does not penalize wrong answers nor test confidence, but candidates must answer all questions to avoid being disadvantaged (i.e. always guess when the answer is unknown).

Security

Following an examination, it may not be prudent to release computer-marked papers for formative use. The reuse of tried and tested 'bank' questions increases the reliability of an examination standard, provided that the questions are kept secure. Electronic files bring with them particular security vulnerabilities, although these are easily overcome. Email attachments can be misdirected and portable media such as USB drives are particularly vulnerable to being lost. Clearly if questions become compromised it takes considerable effort to create new questions. Therefore all electronic files pertaining to summative examinations should be password protected and encrypted.

Blueprinting can be used to provide post-exam feedback to candidates on their strengths and weaknesses without compromising the details of bank questions. Thus once the item analyses and marks have been checked and approved, candidates can automatically electronically receive their individual scores broken down according to the session level learning objectives that were assessed.

Evaluation

Evaluation occurs in two directions, by both anonymous student feedback to teachers and teacher feedback to students. In addition both groups should reflect periodically on their individual strengths, weaknesses, opportunities, and threats (SWOT analysis). Students learn most from each other and fastest by making mistakes. Thus it is important to have the opportunity to make mistakes in a safe or simulated environment, without harming others. To this end, Pendleton's rules can be applied (where appropriate), to emphasize the importance of making the learner receptive to constructive criticism by accentuating good points about their work before revealing things that could be improved (Pendleton et al, 1984). **BJHM**

The author would like to thank Dr Simon Wilkinson for his pioneering virtual learning environment developments.
Conflict of interest: none.

Bloom BS (1956) *Taxonomy of Educational Objectives: The Classification of Educational Goals.* Susan Fauer Company Inc, Chicago: 201–7

Norcini JJ (2003) Setting standards on educational tests: The metric of medical education. *Med Educ* **37**(5): 464–9

Pendleton D, Scofield T, Tate P, Havelock P (1984) *The Consultation: An Approach to Learning and Teaching.* Oxford University Press, Oxford

KEY POINTS

Setting standards depends on:

- Assessing transparent learning outcomes.
- Marking fairness.
- Item difficulty and discrimination, both within and between cohorts.
- Question security.
- Making informed improvements based on evaluation feedback.