Advances in Psychiatric Treatment (2008), vol. 14, 122–130 doi: 10.1192/apt.bp.107.003871

Assessing psychiatric competencies: what does the literature tell us about methods of workplace-based assessment?

Chris Fitch, Amit Malik, Paul Lelliott, Dinesh Bhugra & Manoharan Andiappan

Abstract Workplace-based assessment (WPBA) is becoming a key component of post-graduate medical training in several countries. The various methods of WPBA include: the long case; multisource feedback (MSF); Mini-Clinical Examination (mini-CEX); Direct Observation of Procedural Skills (DOPS); case-based discussion (CbD); and journal club presentation. For each assessment method, we define what the approach practically involves, then consider the key messages and research evidence from the literature regarding their reliability, validity and general usefulness.

The philosophy and mechanics of workplace-based assessment have been discussed in an earlier issue of APT: see Brown & Doshi (2006), which may be downloaded free from the APT website (http://apt.rcpsych.org).

Major changes in post-graduate medical training in several countries have given rise to the development of workplace-based assessment (WPBA). This typically incorporates three main components:

- the direct observation and assessment of a trainee's performance while interacting with real patients in actual clinic settings
- the collation of assessment data on trainee performance from, for example, peers, colleagues and patients
- the assessment of key competencies such as case presentation and note-taking, or the reflective analysis of a trainee's logbook.

The Royal College of Psychiatrists is developing an assessment framework that incorporates WPBA in addition to standardised national examinations. To inform this framework, an international literature review on different methods of WPBA has been undertaken, and this article summarises its findings. Box 1 lists instruments and approaches commonly used in trainee assessment, some of which we focus on here, and Box 2 gives an overview of key terms used in the literature evaluation of these methods.

Box 1 Key methods and instruments of WPBA in medical training

- Long case
- Multisource feedback (MSF)
- Mini-Clinical Evaluation Exercise (mini-CEX) rating scale
- Direct Observation of Procedural Skills (DOPS)
- Assessment of Clinical Expertise (ACE)
- Mini-Assessed Clinical Encounter (mini-ACE)
- Case-based discussion (CbD)
- Journal club presentation
- Mini-Peer Assessment Tool (mini-PAT)
- Team Assessment of Behaviour (TAB)

Chris Fitch is a Research Fellow at the Royal College of Psychiatrists' Research and Training Unit (CRTU) (4th Floor, Mansell Street, London E1 8AA, UK. Email: cfitch@cru.rcpsych.ac.uk). He is a sociologist with an interest in methodological innovation in assessment, and the challenges of living with a mental health problem in the community. Amit Malik is Chair of the Royal College of Psychiatrists' Collegiate Trainees' Committee and a consultant psychiatrist employed by Hampshire Partnership NHS Trust. Paul Lelliott is Director of the CRTU and a consultant psychiatrist employed by Oxleas NHS Trust, where he works as a member of a community mental health team. Dinesh Bhugra is Dean of the Royal College of Psychiatrists, and Professor of Mental Health and Cultural Diversity and Head of the Section of Cultural Psychiatry at the Institute of Psychiatry, London. His research interests include psychosexual medicine, cross-cultural psychiatry, schizophrenia, pathways into psychiatric care, self-harm and primary care. Manoharan Andiappan is a statistician at the CRTU and the Institute of Psychiatry. He is interested in assessment methods in psychiatry and mental health services research.

Box 2 Key terms used in the assessment of methods of WPBA

Reliability
Reliability refers to the reproducibility of assessment data or scores, over repeated tests under identical conditions. The aim is to achieve scores that consistently reflect student performance, while reducing the amount of distortion due to random and systematic errors.

- *Interrater reliability* The degree of agreement between different observers. Research shows that examiners differ in their ratings when assessing the same event. Problems with interrater reliability can be offset by using multiple examiners and ensuring that they are effectively trained.
- *Internal consistency* A measure of whether a set of items in an assessment tool that propose to measure the same 'thing' or construct actually do so.

Reproducibility/generalisability
Although multiple measures of reliability exist (e.g. observer, situation, case, assessee, and their interactions), these measure only one source of error variance at a time. Benefits therefore exist in combining these multiple measures into a single reliability coefficient. Importantly, this can be used to predict how many observations are required with different test formats to achieve a given level of reliability.

- *Reliability coefficients* A measure of reliability, ranging from 0 to 1. The coefficient expresses the consistency of the assessment measure and the degree of random error associated with it. A higher reliability coefficient indicates greater consistency/lower random error.

Validity
The degree of confidence we have that an assessment measures what it is intended to measure.

- *Content validity* Whether an assessment tool systematically includes items that will adequately cover the domain under investigation. This coverage is usually assessed by experts.
- *Construct validity* Whether a particular assessment item/tool adequately captures the theoretical concept that it was designed to measure.
- *Criterion validity* A comparison of the findings of one assessment approach with a gold-standard measurement.

Assessment
Assessment is a systematic process of collecting and interpreting information about an individual in order to determine their capabilities or achievement from a process of instruction.

- *Formative assessment* Occurs during the teaching process and provides feedback to the trainee for their further learning.
- *Summative assessment* Occurs at the end of the learning process and assesses how well the trainee has learned.

The long case

The traditional long case has historically occupied a central and critical role in the evaluation of clinical skills in most medical specialties (Weisse, 2002). In the traditional long-case assessment, trainees are given 30–60 min to interview and examine a non-standard patient unobserved, and up to 1 h to present and discuss the case with one or more examiners.

For examination purposes, the underlying belief is that following assessment of a single long case, active and usually unstructured questioning by an experienced examiner can determine a trainee's competence. The key strength of this approach is that trainees are required to formulate differential diagnosis and management plans for real patients in an actual clinical setting. However, the long case has been criticised for the poor reliability of examiner assessments and the lack of direct observation by the examiner of the trainee–patient encounter (reducing the validity of assessments). Consequently, a new instrument for undertaking long-case assessments with psychiatric trainees has been developed. This instrument is called the Assessment of Clinical Expertise (ACE).

Reliability

Concerns have repeatedly been voiced about the reliability of information generated from the long-case assessment. This is because it is usually based on a single patient encounter and unstructured examiner questioning.

Surprisingly few published data reflect these concerns. However, Norcini (2002) reported that in a study by the American Board of Internal Medicine of assessment in a cardiovascular sub-specialty in the 1970s, two long cases (each with two examiners)

https://doi.org/10.1192/apt.bp.107.003871 Published online by Cambridge University Press

generated a combined reproducibility coefficient of just 0.39, whereas one case resulted in a coefficient of 0.24. For the former, this effectively meant – in strict psychometric terms – that 39% of the variance in trainees' scores was attributable to trainees' ability and the remaining 61% was a result of error measurement (Norcini, 2002). Kroboth *et al* (1992) studied the Clinical Evaluation Exercise (CEX) and reported that two long cases (again with two examiners) had an overall generalisability coefficient below 0.1 and an overall interrater reliability coefficient of 0.40. Weisse (2002) reported that the 1972 decision of the American Board of Internal Medicine to stop using the long case was because of an unacceptably low interrater agreement (measured at 43% – just 5% higher than agreement occurring by chance alone).

Validity

The second concern with the long case relates to its validity. This might appear unusual, given that an argument for retaining the long case is that it accurately replicates the type of situation trainees will encounter in their future professional life. For example, as Wass & van der Vleuten (2004) note, testing the trainee's ability to engage with real patients, collect relevant data and propose an appropriate course of action, the long case represents 'a highly authentic task … [that] comes very close to a candidate's actual daily practice'.

However, because the long case does not typically involve the direct observation of trainees during the patient interview and examination, it can mask weaknesses in their basic skills. Wass & Jolly (2001) undertook a prospective study comparing examiners who observed only the history-taking component of a long case with examiners who observed only the case presentation component. They found a lack of correlation between scores given for long-case observation compared with presentation. In essence, the examiners who directly observed trainees during the history-taking component marked trainees' competence differently from those who only observed the case presentation.

Improving the long case

Attempts to improve the reliability of the long case fall into three categories. First, studies have considered how many additional long cases would be required, with Kroboth *et al* (1992) suggesting that 6–10 long cases (each of 1–1.5 h) would achieve a generalisability coefficient of 0.8.

Second, commentators have attempted to increase the number of long cases, but have done so by employing a format that draws on shorter assessments (20–45 min) and multiple cases (4–6) assessed directly one after another in a single session (McKinley *et al*, 2000; Wass & Jolly, 2001; Hamdy *et al*, 2003; Norcini *et al*, 2003).

Third, elements of the discussion and questioning aspects of the long case have been standardised in an attempt to improve reliability and student perceptions of fairness (Olson *et al*, 2000).

There have also been attempts to improve the validity of the long case. One such has been the introduction of examiners who observe trainee performance throughout the long case. This appears to have been a more recent development in the UK (Wass & Jolly, 2001) than in the USA (US Clinical Evaluation Exercise instrument; Kroboth *et al*, 1992) and Australia (Australian Direct Clinical Examination; Price & Byrne, 1994). Content validity has been addressed through attempts to sample 'types' of patient for the long case, rather than random patient selection (Hamdy *et al*, 2003). This approach has been criticised on the grounds that trainees should be competent enough to deal with most types of patient problems that they encounter (Dugdale, 1996).

The Assessment of Clinical Expertise

The ACE rating scale was developed by the Royal College of Psychiatrists and incorporates the direct observation of trainees throughout the patient encounter. This avoids the basing of judgements on a trainee's skills in case presentation rather than in actual patient contact. The instrument also recognises that the strength of the long case lies in this direct observation of trainee performance, rather than in its reliability coefficient. Since the ACE is one assessment tool in a portfolio of multiple instruments (which have greater reproducibility), this may be less of a concern than when the long case might have been the sole method of assessment. However, direct observation is never a guarantee of accurate observation – assessors will require training and support. The literature suggests that brief training interventions may not be sufficient to achieve the required accuracy (Noel *et al*, 1992).

Multisource feedback

Multisource feedback (MSF) involves the assessment of aspects of a medical professional's competence and behaviour from multiple viewpoints. This can include peer review, where peers at the same level within the organisation and usually within the same medical discipline, are asked to assess the professional. It can include co-worker review, where

https://doi.org/10.1192/apt.bp.107.003871 Published online by Cambridge University Press

other co-workers, who may operate at a higher or lower level in the organisation or may work in a different medical discipline, are asked to assess the professional. It can also incorporate self-assessment, where the professional assesses their own competence and behaviour for comparison with assessments from other sources, and patient review, where patients are asked to assess a professional, typically using a different instrument from that used for peer, co-worker or self-assessment.

The increasing use of MSF is based on the assumption that assessments from multiple viewpoints offer a fairer and more valid description of performance than those based on a single source and that MSF allows the assessment of aspects of professional performance (such as humanistic and interpersonal skills) that are not captured by written or clinical examinations.

The Royal College of Psychiatrists' approach to MSF assessment of psychiatric trainees incorporates co-worker and patient review. It is the only method of WPBA employed by the College that considers the viewpoint of the patient, through the Patient Satisfaction Questionnaire. This assessment tool is distinct from the Mini-Peer Assessment Tool (mini-PAT) and the Team Assessment of Behaviour (TAB), two tools for MSF that are being piloted by the College. This specific approach for the assessment of psychiatric trainees emphasises that MSF is a term used to describe an approach to assessment and not a specific instrument. Hence, we need to be very careful in concluding that what has worked in one programme will also work in another, because different MSF programmes will use different instruments, with different sources, and will measure different behaviours and competencies.

Key research messages

Despite the fact that each approach will be different, a number of general points can be made. First, the number of sources targeted by different approaches ranges from 8 to 25 peers, 6 to 14 co-workers and 25 to 30 service users.

Second, data from evaluations of different instruments indicate that between 8 and 11 peer raters can generate generalisability coefficients between 0.7 and 0.81 (Ramsey *et al*, 1996; Lockyer & Violato, 2004).

Third, some studies conclude that allowing participants to select their own raters does not necessarily bias assessment (Violato *et al*, 1997; Durning *et al*, 2002), contrary to the belief that trainees would nominate raters who they felt would give them a higher score (bias). However, Archer *et al* (2006) found that the profession and seniority of assessors significantly influences assessment. Consequently, from this year the College has constrained psychiatry specialty registrars' discretion over the selection of assessors for MSF, requiring them to nominate their assessors from a broad range of co-workers, not just medical staff (follow link from Workplace Based Assessments – Frequently Asked Questions at www.rcpsych.ac.uk/training/wbpa.aspx).

Fourth, the acceptance of MSF assessment is typically associated with the source of the data – participants tend to value feedback from peers and supervisors more than that from co-workers (such as nurses), particularly when clinical competence is being assessed (Ramsey *et al*, 1993; Weinrich *et al*, 1993; Higgins *et al*, 2004).

Finally, rater groups frequently do not agree about an individual's performance – self-assessments typically do not correlate with peer or patient ratings, and differences in ratings have been found between peers with differing levels of experience (Hall *et al*, 1999; Thomas *et al*, 1999). This disagreement can be seen as a technical threat to interrater reliability, or more practically as measuring different aspects of performance from the position of the rater (Bozeman, 1997).

MSF with psychiatric trainees

When implementing MSF with psychiatric trainees a number of actions can be taken to improve assessment. This section considers some of these.

Instruments can be employed that better reflect the fact that psychiatry differs in its daily practice from other medical specialties, with a far greater emphasis on communication, interpersonal skills, emotional intelligence and relationship building. Generic instruments for MSF should be revised to reflect these differences.

The use of shorter instruments, central administration and alternatives to pen and paper (such as the computer or telephone) is a possible means of countering the view that MSF involves 'too much paperwork' (Lockyer *et al*, 2006).

Multisource feedback plays an important role in making trainees aware of how their performance is perceived by a range of stakeholders and in addressing weaknesses in competence (Violato *et al*, 1997; Lipner *et al*, 2002). However, this is dependent on the quality of the feedback provided. Research shows that highly structured feedback (oral and written) is important (Higgins *et al*, 2004), as is trainee education in appreciating feedback from non-clinical sources.

We know that MSF can bring about changes in practice. It is important that these changes are carefully monitored, both for individual trainee development and also to demonstrate to potential participants/sources that MSF is worthwhile.

https://doi.org/10.1192/apt.bp.107.003871 Published online by Cambridge University Press

Finally, an additional difficulty in a multi-ethnic country such as the UK is finding a way in which non-English speakers can be included, especially for the Patient Satisfaction Questionnaire. One method for achieving this has been to conduct interviews with patients using interpreters (Mason *et al*, 2003), but other approaches will need to be developed to avoid sampling bias.

Mini-Clinical Evaluation Exercise

The mini-CEX is a focused direct observation of the clinical skills of a trainee by a senior medical professional. It involves a single assessor observing a trainee for about 20 min during a clinical encounter. This is followed by 5–10 min of feedback. The mini-CEX was partly developed as one of the 'solutions' to the problems posed by the traditional long case (as discussed above).

Although the assessor observes the trainee while they engage with real patients in real-life clinical situations, critically the assessors are required to focus on how well the trainee undertakes specific clinical tasks, rather than attempting to evaluate every aspect of the patient encounter. This means that one mini-CEX may consider a trainee's skills in history-taking and communication and a later one may focus on clinical judgement and care. Consequently, multiple mini-CEX assessments are undertaken with each trainee. The College approach to mini-CEX has been termed the Mini-Assessed Clinical Encounter (mini-ACE).

Key research messages

Of all the tools for WPBA, the mini-CEX has the largest research evidence base. It has been shown to have a strong internal consistency (Durning *et al*, 2002; Kogan *et al*, 2003) and reproducibility, with a generalisability coefficient of 0.8 for 12–14 assessments (Norcini *et al*, 1995) and 0.77 for 8 assessments (Kogan *et al*, 2003). It has also been argued that the mini-CEX has pragmatic reproducibility, where the scores from 4 assessments can indicate whether further assessments are required (Norcini *et al*, 2003). It has reasonable construct validity, being able to distinguish between different levels of trainee performance (Holmboe *et al*, 2003).

However, the mini-CEX does have limitations. The use of direct observation is not a guarantee of accurate observation (Noel *et al*, 1992) – there is evidence that assessors do make observational errors, which makes in-depth training for assessors vital.

Moreover, the feedback component of the mini-CEX is underdeveloped (Holmboe *et al*, 2004); assessor feedback to trainees is critical for trainee development. Research indicates that assessors do not employ basic feedback strategies such as inviting trainees to assess themselves or using feedback to develop an action plan.

Direct Observation of Procedural Skills

Direct Observation of Procedural Skills (DOPS) is an instrument that allows an educational supervisor to directly observe a trainee undertaking a practical procedure, to make judgements about specific components of the observed procedure and to grade the trainee's performance in carrying out the procedure (Wilkinson *et al*, 2003).

This assessment method was originally developed by the Royal College of Physicians and is based on a large body of work on the rating of technical and procedural skills, including the Objective Structured Assessment of Technical Skills (OSATS; Martin *et al*, 1997). It has primarily focused on technical and psychomotor surgical skills used in operating rooms, laboratories and, more recently, virtual environments (Moorthy *et al*, 2005). Proficiency in basic clinical procedures remains central to good patient care in many medical specialties, but there is good evidence that some doctors lack such proficiency (Holmboe, 2004). For this reason, direct observation and evaluation of competence in clinical procedures should be a core part of training. Studies from the USA suggest that this is not currently the case and report that educational supervisors do not routinely make such observations (Holmboe, 2004).

Key research messages

Studies that consider the reliability or validity of DOPS are scarce. However, studies of the use of OSATS and similar instruments indicate that observation checklists are less reliable than global rating scales (Regehr *et al*, 1998). Moreover, the DOPS approach has been reported to be resource- and time-intensive (Moorthy *et al*, 2005) – raters need to be present during procedures, and if multiple raters of the same procedure are needed then this can be difficult to arrange. Consequently, some commentators have suggested that OSATS may be better conducted using retrospective evaluation of the procedure on videotape (Datta *et al*, 2002).

Undertaking DOPS with psychiatric trainees

Psychiatric practice has fewer practical procedures than other medical specialties. In psychiatry, DOPS could be used in its current form for physical procedures such as administering electroconvulsive

https://doi.org/10.1192/apt.bp.107.003871 Published online by Cambridge University Press

therapy (although this may be infrequent), control and restraint techniques, cardiopulmonary resuscitation and physical examination. However, if these procedures are too infrequent or difficult to schedule, the definition of a 'practical procedure' might be stretched to include practice such as administering a Mini-Mental State Examination or assessing suicide risk. Clearly, this second option raises important questions about the relationship between DOPS and instruments such as the mini-CEX, which also directly observe and assess aspects of these 'procedures'.

A number of actions can be taken to improve DOPS when implementing the approach with psychiatric trainees. Observational training programmes can address basic errors that have been documented in assessor observations (Holmboe, 2004) and can therefore avoid critical trainee performance issues being overlooked. Brief educational interventions were shown to be ineffective in one study, and it has been argued that in-depth observational training is required for all assessors (Noel *et al*, 1992). Given that direct observation features in three methods for WPBA (the long case, mini-CEX and DOPS) this is a clear issue for action.

Strategies are needed for observing procedures that are performed infrequently (Morris *et al*, 2006), with the contexts in which these events occur being identified in advance and made known to assessors and trainees.

Finally, further research is needed to evaluate the use of DOPS and to generate data on the reliability, validity and feasibility of the instrument when used with psychiatric trainees.

Case-based discussion

Case-based discussion uses a written patient record to stimulate a trainee's account of how they managed a case clinically and to allow the examiner to evaluate the decisions taken (and also those ruled out) by the trainee. Through assessing the notes that the trainee has added to the written patient record, CbD can provide useful structured feedback to the trainee.

In practice, CbD (or chart-stimulated recall (CSR), as it is known in North America) involves the trainee pre-selecting several written case records of patients with whom they have recently worked. One of these pre-selected cases is then chosen by the assessor, with detailed consideration being given to a limited number of aspects of the case (rather than an overall case description). During this discussion, trainees explain the clinical decisions they made in relation to the patients, and the medical, ethical, legal and contextual issues they considered in the process. This is followed by assessor feedback. The entire process usually takes 20–30 min.

Key research messages

Research data on the use of CbD as a trainee assessment tool are extremely scarce. There is also an absence of discursive papers about its practical implementation and psychometric strengths and limitations. This is surprising, since CbD arguably is subject to the same psychometric 'yardstick' as the mini-CEX or DOPS. However, four key messages can be identified. First, approaches using CbD are reported to have reasonable validity. Norman *et al* (1993) conducted a comparative study of five assessment methods of physician competency (CbD, standardised patients, structured oral examinations, OSCEs and multiple choice questions) and reported that CbD was among the three methods with 'superior' reliability and validity. Moreover, Maatsch *et al* (1984), in a study of competence in emergency medicine, reported concurrent validity in the relationship between physicians' CSR scores and results from the American Board of Emergency Medicine.

Second, approaches using CbD have reasonable reliability. Solomon *et al* (1990) compared CSR with a simulated patient encounter and concluded that it was a reliable form of assessment when examiners had received adequate training.

Third, Maatsch *et al* (1984) reported that 3–6 cases are required to assess physician competence in emergency medicine.

Fourth, CbD is positively related to student knowledge and observational skills. Goetz *et al* (1979) reported that, although student performance on chart reviews was affected by time pressures, performance improved with clinical experience.

However, CbD does have important limitations. Jennet & Affleck (1998) noted that its reliance on self-report raises questions about the accuracy of trainee recall and rationalisation of a case. There is the potential for linking CbD with other assessments of the same case under consideration (such as the mini-CEX or DOPS).

Journal club presentation

A medical journal club is made up of individuals who regularly meet to discuss the strengths, weaknesses and clinical application of selected articles from the medical literature (Lee *et al*, 2005). Modern medical journal clubs have evolved from being primarily a discursive means for trainees to keep abreast of new literature into a forum where critical appraisal skills and evidence-based medicine are taught and applied (Ebbert *et al*, 2001). This has resulted in increasing interest in the role and effectiveness of journal clubs in informing academic and clinical practice, and several systematic and thematic literature reviews

https://doi.org/10.1192/apt.bp.107.003871 Published online by Cambridge University Press

have been undertaken (Alguire, 1998; Norman & Shannon, 1998; Green, 1999; Ebbert *et al*, 2001; Lee *et al*, 2005). These indicate that journal clubs may improve participants' knowledge of clinical epidemiology and biostatistics, their reading habits and their use of medical literature in clinical practice. Interestingly, with the exception of Green (1999), there is no evidence, however, that journal clubs have a proven role in improving critical appraisal skills. Successful journal clubs are organised around structured review checklists, explicit written learning objectives, and formalised meeting structures and processes.

A number of these reviews have also recommended that journal clubs could serve as a tool for teaching and assessing practice-based competence. Lee *et al* (2005), for example, contend that the journal club has a familiar format, requires little additional infrastructure for assessment and has low start-up and maintenance costs.

Key research messages

A potential role for the journal club in assessing practice-based competence is now taking shape. To our knowledge, however, no studies have specifically considered trainee presentations as a method for assessing competence, with a greater emphasis instead being placed on studies of the wider membership of the journal club. Consequently, to consider this method of assessment we must turn to the large published literature on the assessment and evaluation of oral and student presentation. Not surprisingly, numerous criteria and checklists have been proposed, including criteria for specific disciplines and methods (trainee presentations can cover a range of different research studies, each with different research methodologies). This may require examiners to have access to generic critical appraisal guidelines (Greenhalgh, 1997), criteria for particular methods to assess the quality of the trainee's presentation (Critical Appraisal Skills Programme at www.phru.nhs.uk/Pages/PHD/CASP.htm; Canadian Centers for Health Evidence at www.cche.net) and delivery and oratory guidelines (criteria developed to evaluate 'non-content' issues of presentations such as structure, voice audibility and body language).

A cautionary note

Although standardised tools and processes have been used for assessments in medical education for a while, recent expert commentaries caution against over reliance on these, as they are mainly summative rather than formative tools (Schurwirth & van der Vleuten, 2006). This is reflected in the current stage of development of the Royal College of Psychiatrists' tools for WPBA. Concern has also been expressed regarding the dangers of compromising the construct validity of assessment processes for the sake of increasing reliability, as this does not mirror 'real-life' situations. Some authors stress the importance of professional trust and the judgement of educational supervisors and how these play a significant role in assessing 'performance', a step higher than 'competence' on Miller's pyramid of assessments (ten Cate, 2006). It is therefore essential that any future assessment framework takes into account these important principles and does not end up being a glorified competence checklist. Any such move would reduce psychiatrists to mere technicians and would have a significant long-term impact on the performance of the profession as a whole.

Conclusions

Each of the assessment methods described above has been developed for a specific purpose. In practice, however, they will be used in combination to assess trainees' competence. The relationship between different instruments (and trainees' scores on them) therefore needs to be carefully considered, including the interpretation, comparison and any weighting of scores. This is particularly important in relation to the ACE, mini-ACE and DOPS, which all involve the direct observation of trainee proficiency in basic psychiatric skills. In addition, as many of the assessment tools have not been designed specifically for psychiatry or for postgraduate medical education in the UK, context-specific evaluation of these tools is required to inform their further development. The literature for WPBAs in postgraduate psychiatric training is fairly limited, but as these assessments are implemented widely this is likely to change.

Declaration of interest

None.

References and related articles

Alguire, P. C. (1998) A review of journal clubs in postgraduate medical education. *Journal of General Internal Medicine*, **13**, 347–353.

Archer, J., Norcini, J., Southgate, L., *et al* (2006) mini-PAT (Peer Assessment Tool): a valid component of a national assessment programme in the UK. *Advances in Health Sciences Education*, online content (http://springerlink.metapress.com/content/v8hth3h06507ph56/?p=a3e3005a289e47e386a81bb976abef58&pi=2).

Bozeman, D. (1997) Interrater agreement in multisource performance appraisal: a commentary. *Journal of Organizational Behavior*, **18**, 313–316.

https://doi.org/10.1192/apt.bp.107.003871 Published online by Cambridge University Press

Brown, N. & Doshi, M. (2006) Assessing professional and clinical competence: the way forward. *Advances in Psychiatric Treatment*, **12**, 81–89.
Datta V., Chang A., Mackay S., *et al* (2002) The relationship between motion analysis and surgical technical assessments. *American Journal of Surgery*, **184**, 70–73.
Dugdale, A. (1996) Long-case clinical examinations. *Lancet*, **347**, 1335.
Durning, S. J., Cation, L. J., Markert, R. J., *et al* (2002) Assessing the reliability and validity of the mini-clinical evaluation exercise for internal medicine residency training. *Academic Medicine*, **77**, 900–904.
Ebbert, J. O., Montori, V. M. & Schultz, H. J. (2001) The journal club in postgraduate medical education: a systematic review. *Medical Teacher*, **23**, 455–461.
Goetz, A. A., Peters, M. J., Folse, R., *et al* (1979) Chart review skills: a dimension of clinical competence. *Journal of Medical Education*, **54**, 788–796.
Green, M. L. (1999) Graduate medical education training in clinical epidemiology, critical appraisal, and evidence-based medicine: a critical review of curricula. *Academic Medicine*, **74**, 686–694.
Greenhalgh, T. (1997) How to read a paper: assessing the methodological quality of published papers. *BMJ*, **315**, 305–308.
Hall, W., Violato, C., Lewkonia, R., *et al* (1999) Assessment of physician performance in Alberta: the Physician Achievement Review. *Canadian Medical Association Journal*, **161**, 52–57.
Hamdy, H., Prasad, K., Williams, R., *et al* (2003) Reliability and validity of the direct observation clinical encounter examination (DOCEE). *Medical Education*, **37**, 205–212.
Higgins, R. S. D., Bridges, J., Burke, J. M., *et al* (2004) Implementing the ACGME general competencies in a cardiothoracic surgery residency program using a 360-degree feedback. *Annals of Thoracic Surgery*, **77**, 12–17.
Holmboe, E. S. (2004) Faculty and the observation of trainees' clinical skills: problems and opportunities. *Academic Medicine*, **79**, 16–22.
Holmboe, E. S., Huot, S., Chung, J., *et al* (2003) Construct validity of the Mini Clinical Evaluation Exercise (MiniCEX). *Academic Medicine*, **78**, 826–830.
Holmboe, E. S., Yepes, M., Williams, F., *et al* (2004) Feedback and the mini clinical evaluation exercise. *Journal of General Internal Medicine*, **5**, 558–561.
Jennett, P. & Affleck, L. (1998) Chart audit and chart stimulated recall as methods of needs assessment in continuing professional health education. *Journal of Continuing Education in the Health Professions*, **18**, 163–171.
Kogan, J. R., Bellini, L. M. & Shea, J. A. (2003) Feasibility, reliability, and validity of the mini-clinical evaluation exercise (mCEX) in a medicine core clerkship. *Academic Medicine*, **78**, S33–S35.
Kroboth, F. J., Hanusa, B. H., Parker, S., *et al* (1992) The inter-rater reliability and internal consistency of a clinical evaluation exercise. *Journal of General Internal Medicine*, **7**, 174–179.
Lee, A. G., Boldt, C., Golnik, K. C., *et al* (2005) Using the journal club to teach and assess competence in practice-based learning and improvement: a literature review and recommendation for implementation. *Survey of Ophthalmology*, **50**, 542–548.
Lipner, R. S., Blank, L. L., Leas, B. F., *et al* (2002) The value of patient and peer ratings in recertification. *Academic Medicine*, **77**, S64–S66.
Lockyer, J. M. & Violato, C. (2004) An examination of the appropriateness of using a common peer assessment instrument to assess physician skills across specialties. *Academic Medicine*, **79**, S5–S8.
Lockyer, J., Blackmore, D., Fidler, H., *et al* (2006) A study of a multisource feedback system for international medical graduates holding defined licenses. *Medical Education*, **40**, 340–347.
Maatsch, J. L., Huang, R. R., Downing, S., *et al* (1984) The predictive validity of test formats and a psychometric theory of clinical competence. *Research in Medical Education*, **23**, 76–82.
Martin, J. A., Regehr, G., Reznick, R., *et al* (1997) Objective structured assessment of technical skill (OSATS) for surgical residents. *British Journal of Surgery*, **84**, 273–278.
Mason, R., Choudhry, N., Hartley, E., *et al* (2003) Developing an effective system of 360-degree appraisal for consultants: results of a pilot study. *Clinical Governance Bulletin*, **4**, 11–12.
McKinley, R. K., Fraser, R. C., van der Vleuten, C., *et al* (2000) Formative assessment of the consultation performance of medical students in the setting of general practice using a modified version of the Leicester Assessment Package. *Medical Education*, **34**, 573–579.
Moorthy, K., Vincent, C. & Darzi, A. (2005) Simulation based training. *BMJ*, **330**, 493–495.
Morris, A., Hewitt, J. & Roberts, C. M. (2006) Practical experience of using directly observed procedures, mini clinical evaluation examinations, and peer observation in pre registration house officer (FY1) trainees. *Postgraduate Medical Journal*, **82**, 285–288.
Noel, G. L., Herbers, J. E., Capow, M. P., *et al* (1992) How well do internal medicine faculty members evaluate the clinical skills of residents? *Annals of Internal Medicine*, **1**, 757–765.
Norcini, J. J. (2002) The death of the long case? *BMJ*, **324**, 408–409.
Norcini, J. J., Blank, L. L., Arnold, G. K., *et al* (1995) The mini-CEX (clinical evaluation exercise): a preliminary investigation. *Annals of Internal Medicine*, **123**, 795–799.
Norcini, J. J., Blank, L. L., Duffy, D., *et al* (2003) The mini-CEX: a method for assessing clinical skills. *Annals of Internal Medicine*, **138**, 476–481.
Norman, G. R. & Shannon, S. I. (1998) Effectiveness of instruction in critical appraisal (evidence-based medicine) skills: a critical appraisal. *Canadian Medical Association Journal*, **158**, 177–181.
Norman, G. R., Davis, D. A., Lamb, S., *et al* (1993) Competency assessment of primary care physicians as part of a peer review 270 program. *JAMA*, **9**, 1046–1051.
Olson, L. G., Coughlan, J., Rolfe, I., *et al* (2000) The effect of a Structured Question Grid on the validity and perceived fairness of a medical long case assessment. *Medical Education*, **34**, 46–52.
Price, J. & Byrne, G. J. A. (1994) The direct clinical examination: an alternative method for the assessment of clinical psychiatry skills in undergraduate medical students. *Medical Education*, **28**, 120–125.
Ramsey, P. G., Wenrich, M. D., Carline, J. D., *et al* (1993) Use of peer ratings to evaluate physician performance. *JAMA*, **13**, 1655–1660.
Ramsey, P. G., Carline, J. D., Blank, L. L., *et al* (1996) Feasibility of hospital-based use of peer ratings to evaluate the performances of practising physicians. *Academic Medicine*, **71**, 364–370.
Regehr, G., MacRae, H., Reznick, R. K., *et al* (1998) Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. *Academic Medicine*, **73**, 993–997.
Schurwirth, L. W. T. & van der Vleuten, C. P. M. (2006) A plea for new psychometric models in educational assessment. *Medical Education*, **40**, 296-300.
Solomon, D. J., Reinhart, M. A., Bridgham, R. G., *et al* (1990) An assessment of an oral examination format for evaluating clinical competence in emergency medicine. *Academic Medicine*, **65**, S43–S44.
ten Cate, O. (2006) Trust, competence, and the supervisor's role in postgraduate training. *BMJ*, **333**, 748–751.
Thomas, P. A., Gebo, K. A. & Hellmann, D. B. (1999) A pilot study of peer review in residency training. *Journal of General Internal Medicine*, **14**, 551–554.
Violato, C., Marini, A., Toews, J., *et al* (1997) Feasibility and psychometric properties of using peers, consulting physicians, co-workers, and patients to assess physicians. *Academic Medicine*, **72**, S82–S84.
Wass, V. & Jolly, B. (2001) Does observation add to the validity of the long case? *Medical Education*, **35**, 729–734.
Wass, V. & van der Vleuten, C. P. M. (2004) The long case. *Medical Education*, **38**, 1176–1180.
Weinrich, M. D., Carline, I. D., Giles, L. M., *et al* (1993) Ratings of the performances of practising internists by hospital-based registered nurses. *Academic Medicine*, **68**, 680–687.
Weisse, A. B. (2002) The oral examination. Awful or awesome? *Perspectives in Biology and Medicine*, **45**, 569–578.
Wilkinson, J., Benjamin, A. & Wade, W. (2003) Assessing the performance of doctors in training. *BMJ*, **327**, S91–S92.

https://doi.org/10.1192/apt.bp.107.003871 Published online by Cambridge University Press

MCQs

1 The Assessment of Clinical Expertise (ACE):
a involves trainees making a presentation about a patient they have interviewed in the past
b involves trainees being assessed by their co-workers
c requires trainees to critically assess a clinical article they have read
d involves trainees being directly observed while interviewing and examining a patient, and discussing this with examiners
e requires the trainee to perform a specific medical procedure.

2 The Mini-Peer Assessment Tool (Mini-PAT)
a involves trainees being directly observed and assessed during part of an actual clinical encounter
b requires the trainee to perform a medical procedure
c involves trainees being assessed by their co-workers
d involves trainees discussing with examiners a patient they have managed in the past
e requires trainees to critically assess a clinical article they have read.

3 The Mini-Assessed Clinical Encounter (mini-ACE):
a involves trainees interviewing a patient unobserved, and then discussing this with examiners
b involves trainees being directly observed and assessed about a specific part of an actual clinical encounter
c involves trainees discussing with examiners a patient they have managed in the past
d requires trainees to critically assess a clinical article they have read
e involves trainees being assessed by their co-workers.

4 Case-based discussion (CbD):
a involves trainees being directly observed and assessed about a specific part of an actual clinical encounter
b involves trainees interviewing a patient unobserved, and then discussing this with examiners
c involves trainees discussing with examiners a patient they have managed in the past
d requires trainees to critically assess a clinical article they have read
e involves trainees being assessed by their co-workers.

5 Journal club presentation:
a involves trainees being directly observed and assessed about a specific part of an actual clinical encounter
b requires trainees to critically read a clinical article
c involves trainees discussing a paper the examiner presents to them
d requires trainees to critically read and present a clinical article they have read
e involves trainees being assessed by their co-workers.

MCQ answers

1	2	3	4	5
a F	a F	a F	a F	a F
b F	b F	b T	b F	b F
c F	c T	c F	c T	c F
d T	d F	d F	d F	d T
e F	e F	e F	e F	e F

https://doi.org/10.1192/apt.bp.107.003871 Published online by Cambridge University Press