Workplace-based assessments in psychiatry: setting the scene
Amit Malik and Dinesh Bhugra

Background: changing socio-political context

Over the last decade the socio-political world within which medicine exists has changed significantly. The expectations of medicine as a profession, and doctors as its constituent members, have morphed drastically and the role of the physician has transformed from an autocratic decision-maker to a collaborative facilitator for patients in supporting them to make complex healthcare decisions. National and international high-profile healthcare scandals have also meant that governments, and through them the public, now expect greater transparency in the standards to which doctors are trained and demand evidence that trained doctors have actually met these standards before they are allowed into independent practice. Additionally, there are public and professional expectations that fully trained doctors are evaluated at regular intervals to ensure that they are up to date with the latest developments and have maintained the professional skills that enable them to practise medicine.

Moreover, all this is happening when, on the one hand, quite appropriately, an emphasis on patient safety has necessitated a reduction in the working hours of trainees around the world and, on the other hand, many healthcare training systems are being encouraged to shorten the length of training, to meet healthcare workforce needs and to reduce the overall cost of training. Therefore, one can no longer rely on the traditional apprenticeship model where trainees were expected to spend long hours over a number of years in training with the hope that they would have enough experience and would have learnt sufficiently by osmosis to enable them to practise independently. Instead, training has had to become more structured, efficient and outcome-orientated, to ensure that trainees engage in educationally valuable activities to attain competencies that have been defined in a curriculum that is ‘fit for the purpose’ to prepare the needs of their professional
practice on completion of training. Additionally, trainees must undergo assessment to demonstrate that they have attained these competencies.

However, socio-political factors are not the only reason for changes in the world of postgraduate medical education. For decades, assessments in medical training have relied on the evaluation of a postgraduate trainee’s knowledge or skills in one-off high-stake encounters rather than focusing on their ability to acquire competencies and perform in the workplace. This has created perverse incentives for learning amongst postgraduate medical trainees and fostered a medical education system that, some would argue, has created ‘competent’ doctors despite of rather than because of it.

This introductory chapter discusses some of the rationale for the changes in postgraduate medical assessments, including the introduction and implementation of workplace-based assessments as components of assessment systems. It also outlines the formative and summative role of all assessments and contextualizes assessments within educational programmes based on competency-based curricula. Finally, it introduces some of the technical and psychometric considerations that must be taken into account when developing assessment systems for postgraduate medical training.

Rationale for workplace-based assessment

For decades, postgraduate medical assessments have focused on assessing knowledge in one-off high-stake examinations. Since the 1980s, in addition to assessments of knowledge, these examinations have progressed to include skills assessments, such as the long case examination and objective structured clinical examinations (OSCEs). However, as Miller outlined in his landmark discourse in 1990, clinical competencies need to be assessed at four different levels, which he draws on a pyramid: ‘knows’, ‘knows how’, ‘shows how’ and ‘does’. Workplace-based assessments (WPBAs) target ‘does’ – the highest level of Miller’s pyramid, as this involves assessing a trainee’s actual performance in the workplace (Miller, 1990). It is particularly important to assess performance in the workplace throughout the duration of training, for a variety of reasons.

Firstly, workplace-based assessments allow for a more authentic evaluation of how a trainee would respond in a real-life situation compared with the artificial circumstances of an examination. Secondly, assessing trainees on a regular basis within the context of the workplace provides greater opportunities for diagnosing developmental needs at an early stage and delivering regular formative feedback throughout training rather than at a few discrete points in high-stake situations. Thirdly, having a number of assessments over the training period allows for greater sampling of the curriculum than would be feasible in periodic examinations. In addition, having assessments from a number of assessors enables a greater triangulation of information to assess a trainee’s competence and mitigates some of
the concerns around inter-rater reliability in examination settings with a limited number of assessors assessing over a short period of time. Finally, it is now clear that assessments drive learning and that assessing trainees for their performance in all professional domains (clinical and non-clinical) in real-life situations will enable them to develop such professional skills as teamworking, forming relationships with colleagues and leadership, which are often not easily tested in artificial examination contexts.

All this is not to say that workplace-based assessments are superior to examinations, as examinations provide a degree of externality and standardization that plays a crucial role in providing confidence in the very high-stake decisions that allow a medical trainee to undertake independent clinical practice with patients. The above rationale merely supports a case for developing an assessment programme that includes both workplace-based assessments and external examinations, where the overall high-stake pass–fail decisions are made by taking evidence from both these sources in addition to the professional judgement of experienced educational and clinical supervisors in the workplace.

Purpose of workplace-based assessments

To engender stakeholder trust within an assessment programme, the purpose of its various component assessments should be explicit from the outset. The two main purposes of assessment are for learning (formative) and of learning (summative) (Wass et al., 2001). This distinction, however, is a lot more artificial and complex in actual practice: some of the evidence from formative assessments, such as workplace-based assessments, is often used for making summative decisions, and the feedback from summative assessments, such as examinations, can often clarify developmental needs. Rather than attempting to create an artificial and impossible distinction between formative and summative assessments, assessment programme developers should ensure that the different purposes of assessments are clarified to examiners from the outset and all efforts are made to ensure that individual assessments are psychometrically, logistically and educationally fit for their intended purposes.

Curriculum development and blueprinting

Whilst it is beyond the scope of this book to discuss the principles and process of curriculum development, suffice it to say that outcome-based curricula must be developed to define the competencies that a fully trained medical professional must possess in a particular specialty at a particular time. Assessment systems must be embedded within these outcome-based curricula and must consist of a range of assessments that assess the various domains of the curriculum.
Two principles should be taken into account whilst doing this. Firstly, the test content should be planned in relation to the learning outcomes by the process of blueprinting. A blueprint is a matrix in which the test designer decides how many items/tasks are to be assessed for each subject or category (Holsgrove et al., 2009). Secondly, the blueprinting should also ensure that the format of an assessment is appropriate for the domain (such as knowledge, skills or attitudes) of the curriculum that it intends to assess (Wass et al., 2001). Both these steps are crucial to ensure the validity of individual assessments and the overall assessment system.

Utility of assessments

In his seminal paper, Cees van der Vleuten (1996) defined the concept of utility as a multiplicative product of reliability, validity, educational impact, acceptability and cost. Describing perfect utility as a utopia, he used the model to describe the compromises and trade-offs between the constituent variables of utility, depending on the context and purpose of the assessment. In designing assessment tools and programmes, each of these variables must be considered and it would be prudent to discuss them in a bit more detail here.

Reliability

Reliability is a measure of how reproduducible and consistent a test’s results are (Wass et al., 2001). Many factors influence the overall reliability of an assessment, including assessor factors, clinical case and the context of the assessment. Cronbach’s $\alpha$ (alpha) is a well-recognized contemporary measure of internal consistency used by test developers. Its value ranges from 0 to 1; the requisite value for a particular assessment depends on the purpose of the assessment. As a general principle, the higher the stakes, the greater the value of $\alpha$ that is required for an assessment. For instance, whilst $\alpha \geq 0.9$ would be desirable in high-stake examinations, $\alpha \leq 0.8$ might be acceptable for workplace-based assessment (Holsgrove et al., 2009).

Different assessors rate trainees differently for similar performances and some of this error can be minimized by assessor training and standardization of rating scales. However, to ensure greater inter-rater reliability of workplace-based assessments, several assessors should be used to triangulate the assessment information across a range of contexts. It is now also known that trainees perform differently across clinical cases and that, notwithstanding generic skills such as communication, performance measurement is task-specific and doctors perform differently at different tasks. Therefore, there should be adequate sampling from the curriculum to assess trainees across a range of clinical cases (van der Vleuten, 1996).

However, it has recently been argued that it is not sufficient merely to regard assessors as *interchangeable measurement instruments*, whereby lack of reliability
or consistency can be resolved simply by enhancing the standardization and objec-
tivity of assessment tools or increasing inter-rater reliability by assessor training. Such a great reliance solely on a psychometric approach might ignore the influence of assessors’ personal and organizational perspectives on the overall performance measurement, as well as the clinical, organizational and social context within which the assessment takes place, especially in workplace-based assessments. Assessors must be seen as actively making complex professional judgements and it must be recognized that contextual factors have a significant influence on this process. It is very important to understand the assessors’ judgement and decision-making processes from a social-cognitive perspective and the environmental factors that affect assessors’ motivation and goals (Govaerts et al., 2007).

Validity
Validity is a measure of how thoroughly a test assesses what it purports to assess. Messick (1995) defines validity as:

The degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other models of assessment.

Traditionally, validity has been classified into three types—content, construct and criterion, where criterion validity is further divided into predictive and concurrent validity. There have been criticisms of the concept of face validity (does a test appear to assess what it claims to?) (Streiner and Norman, 2003). A more contemporary view also sees all validity as being construct validity. The current state-of-the-art view is that validity is a process of hypothesis testing, wherein the aim of a validation study is to formulate a hypothesis about the inferences that can be drawn from the results of an assessment and then to collect evidence to prove or disprove the hypothesis (Downing and Haladyna, 2004).

Appropriate blueprinting of assessments of a curriculum supports the content validity of the assessment framework. Concurrent validity models (e.g., workplace-based assessment with national written examinations) must be evaluated, especially where the same curriculum domains are being assessed.

Predictive validity studies using longitudinal data (e.g., success in workplace-based assessments predicting clinical or examination success) are important in enhancing the validity and acceptability of assessment programmes. Qualitative evaluation methods (such as surveys and focus groups) can be used to assess the consequential validity (e.g., educational impact) of workplace-based assessment (Holsgrove et al., 2009).

Educational impact
Assessments strongly influence learning in a variety of ways and therefore have a significant educational impact (Wass et al., 2001). The content, format, feedback
Chapter 1: Setting the scene

and scheduling of assessments all influence what and how trainees learn from the curriculum (van der Vleuten, 1996). The challenge, therefore, for developers of assessments and assessment systems is to utilize this knowledge in developing assessment systems that promote the curriculum objectives and outcomes.

Especially in busy vocational learning environments, such as postgraduate medical training, trainees will always prioritize learning components of the curriculum that are assessed most frequently, or most thoroughly, or to which greatest importance is attached within the overall programme outcome (van der Vleuten, 1996). Therefore, whilst blueprinting is important to ensure adequate sampling of the curriculum, it is also crucial to ensure that trainees learn the most important components of the curriculum most effectively. The content of the assessment programme should give consistent messages regarding this.

Ideally, a variety of assessment formats should be used within an assessment system with each format validated for a particular purpose. Additionally, the same format can be used to assess different aspects of the curriculum (Holsgrove et al., 2009). It is important to recognize the intended and unintended consequences of assessment formats, and their evaluation and ongoing analysis are crucial in managing and responding to these consequences when they occur (van der Vleuten, 1996).

Feedback must constitute an indispensable part of assessments and assessment programmes as it contributes significantly to the educational impact of assessments and the trainee’s learning experience. It is postulated that in the context of postgraduate medical training, facilitative rather than directive feedback is more effective. Feedback should relate to tasks, be specific instead of general and relate to the individual (Archer, 2010). To achieve this, workplace-based assessments should be viewed as structured and regular feedback opportunities and feedback given should link to action planning and the trainee’s overall learning plan. All this can only be attained if a culture of reflection is developed within training healthcare organizations and trainers are trained to facilitate effective, formative and action-orientated feedback (Holsgrove et al., 2009).

Finally, the frequency, timing and relation to progression have a huge effect on the educational impact of assessments, and careful and conscious consideration should be given to these factors by the designers of assessment programmes (van der Vleuten, 1996).

Acceptability

Assessment tools and programmes that do not enjoy stakeholder acceptability, especially from trainees and assessors, will not succeed. Dijkstra et al. (2010) outline the prerequisites for trust in and acceptability of assessment systems previously detailed in the literature. These include authenticity, fairness, honesty, transparency of procedures (due process), well-defined roles and high-quality feedback (Govaerts et al., 2007). Additionally, including assessors in the development of
assessment tools and programmes enhances the acceptability of the assessment tools and programmes. It is also important that there is clear communication between the programme developers and the assessors and trainees regarding the outcome measures set out within the curriculum, their relative significance and their correspondence to the domains on the assessment tools.

A clear, unambiguous and explicit message needs to be relayed to assessors and trainees regarding the purpose of assessments and the manner in which assessment results will be used. All stakeholders must be actively encouraged to provide feedback on the assessment programme and this should be utilized to make positive modifications to the programme. Finally, employing organizations must provide support for assessments to be carried out not only by ensuring that time and other resources are made available — an area that will be covered next when considering the issue of cost — but also by promoting an organizational culture that values honest feedback and reflection in order to promote the development of competent and reflective specialists (Govaerts et al., 2007).

Cost

There are significant overt costs involved in developing, delivering and evaluating assessment methods and programmes. It is important, right at the outset of assessment programme development, to address this issue in two ways. Firstly, the assessment programme developers must identify the funding resources for the various stages of assessment development and delivery and, secondly, they must explore innovative ways in which these costs can be minimized, for example by using information technology systems rather than paper assessments and by sharing assessment delivery at a regional or national level. For instance, in the UK, workplace-based assessments are delivered across all training programmes using a national IT system.

However, there are significant covert costs involved, especially in the delivery of assessment systems, including assessor and trainee time to undertake assessments and provide or receive regular feedback. As already stated, it is crucial that employing organizations make time and resources available to assessors and trainees to ensure that detailed and valid assessments can take place and structured feedback can be provided to trainees.

Assessments as components of educational programmes

The medical education literature is flooded with evaluations of different workplace-based assessment tools. However, concerns have been raised in more recent work, regarding the relative absence of a systematic approach to assessments. This involves having a clear set of outcomes (or competencies) that are expected of a postgraduate medical training programme. Various assessment methods and their content
should then be clearly specified, to ensure that there is adequate sampling across all outcomes (blueprinting) and appropriate assessment formats are utilized for assessing different competencies. Addressing assessments in this way will also help in reducing the overlap between the areas assessed by different assessments; allowing weaknesses of some assessments to be compensated by strengths of others and assisting in the triangulation of information from a range of sources, especially whilst making high-stake decisions (Dijkstra et al., 2010).

Conclusion

This chapter has aimed to set the context for the assessment of trainee performance in the workplace during postgraduate psychiatric training. A clear message is that developing and delivering a successful assessment programme is not only a psychometric issue but is also a cultural, political and economic one. There is an overall note of caution in overemphasizing the psychometric principles at the expense of educational and social-cognitive ones. Moreover, the responsibility lies with us, the psychiatric profession, to ensure that we set out clear outcomes for our trainees within training programmes and then ensure that these outcomes are adequately assessed by an assessment programme. This is crucial not only to develop ‘fit-for-the-purpose’ psychiatrists for tomorrow but also to ensure that we protect our patients and maintain wider public confidence and trust within the psychiatric profession.

REFERENCES


Chapter 1: Setting the scene


Workplace-based assessments – an evidence-based overview

Amit Malik and Dinesh Bhugra

Editors’ introduction

In many parts of the world, workplace-based assessment methods now constitute an indispensable component of assessment systems in postgraduate psychiatric training. The previous chapter outlined the theoretical background for workplace-based assessments, including some of the technical considerations of their development, implementation and evaluation. This chapter aims to classify different methods by which postgraduate medical trainees can be assessed in the workplace, and a brief discussion follows of the contemporary evidence related to some of these assessment methods. It is inevitable that some assessments work better than others, and the supporting literature varies depending upon how long a tool has been used and how its validity is assessed. Both assessors and trainees need to be aware of the pros and cons of the various assessments and use them accordingly.

Classification of workplace-based assessments

Various workplace-based tools are being used and different aspects of their utility have been analyzed. It is crucial that assessment methods are developed and used as part of a pragmatic approach rather than as individual assessment tools that do not relate to each other as components of the relevant curriculum. Careful consideration also needs to be given to the content and format of assessments to ensure that they sample the curriculum efficiently and effectively and validly assess the intended aspects of a trainee’s competencies.

Reviewing the literature (C. Fitch, A. Malik and P. Lelliott, personal communication, 2009), it is possible to classify the assessment tools into three distinct categories: