Simulated patients and objective structured clinical examinations: review of their use in medical education

Jeremy Wallace, Ranga Rao & Richard Haslam

Simulated or standardised patients have been used in medical education and other medical settings for some 30 years (Box 1). Their use encompasses undergraduate and postgraduate learning, the monitoring of doctors’ performance and standardisation of clinical examinations. Simulation has been used for instruction in industry and the military for much longer (Jason et al., 1971) but the first known effective use of simulated patients was by Barrows & Abrahamson (1964), who used them to appraise students’ performance in clinical neurology examinations.

The objective structured clinical examination (OSCE) was first described by Harden & Gleeson as, ‘a timed examination in which medical students interact with a series of simulated patients in stations that may involve history-taking, physical examination, counselling or patient management’ (Harden & Gleeson, 1979). Because OSCEs have been shown to be feasible and have good reliability and validity (Hodges et al., 1998) their use has become widespread as the standard for performance-based assessment, particularly in undergraduate examinations.

The use of OSCEs in undergraduate examinations (‘summative’ use) occurs in every London medical college and was pioneered by the Royal London and St Bartholomew’s Hospitals. Many colleges across the UK have now adapted their examinations to include these components. In addition, there is considerable uptake in the use of simulated patients for medical student training (‘formative’ use).

Several of the medical Royal Colleges have introduced an OSCE component into their postgraduate membership examinations. For example, the Royal College of Anaesthetists includes an OSCE in Part I of the fellowship examination, and the Royal College of Obstetricians and Gynaecologists has an OSCE in Part II of their examinations. The Royal College of Surgeons, London, are introducing OSCEs and a pilot is being planned for this year. The Royal College of Physicians has a Practical Assessment of Clinical Examination Skills (PACES) in their clinical examinations, part of which comprises a communication and ethics station, in which simulated patients are used. Although the preferred method of examination is with videotape of real consultations, the Royal College of General Practitioners has a ‘simulated surgery’ that about 5% of candidates use.

The Royal College of Psychiatrists has recently proposed changes to the existing membership examinations with a view to increasing their reliability and validity (http://www.rcpsych.ac.uk/traindev/exams/exam_recent.htm). The main changes are to the Part I examination, and from spring 2003 the existing individual patient assessment will...
be replaced by an OSCE, in which simulated patients are likely to be used. The Part II examination will essentially retain the same format.

In view of the proposed changes we feel that a review of the literature surrounding the use of simulated patients and OSCEs would be both useful and timely. In this paper we focus on the following issues. First, the definition of simulated patients; second, the feasibility of standardising simulators; third, a historical overview of simulated patients in medical settings; and fourth, the reliability and validity of OSCEs. We also comment on the possible burden to the people who are playing the part of patients.

Who are simulated patients?

Despite the increasing use of simulated patients in medical education, there remains a problem when reviewing the literature regarding what is meant by a simulated or standardised patient. Some authors use the term simulated patients (Sanson-Fisher & Poole, 1980; Norman et al, 1982), but others use the term standardised patients (Rubin & Philp, 1998). Although having quite different meanings, these two terms are often used interchangeably. Others have used the terms pseudo or surrogate patients (Badger et al, 1995). Vu & Barrow’s (1994) definition of standardised patients includes ‘real or simulated patients who have been coached to present a clinical problem’. We found only one reference relating to the use of real patients (McLure et al, 1985). Therefore, in the rest of this review we use the term ‘simulated patients’, as we feel it encompasses all definitions.

In some cases simulated patients are professionally trained actors playing the part of patients (Norman et al, 1982). However, Sanson-Fisher & Poole (1980), when comparing medical students’ performance with that of real and simulated patients, used volunteer simulators who were not ‘members of the acting profession’. In a description of new medical student teaching at Michigan State University, Jason et al (1971) write that their simulated patients were ‘primarily drama students from our campus’ and additionally ‘several housewives, some of whom had no previous acting experience’. Rubin & Philp (1998), in a study of the health perceptions of simulators, state that simulated patients were recruited from ‘the allied health programmes at local colleges, community volunteer programmes, the community senior citizen programme and from clinics at a university-based department of family medicine’. Others, for example Hodges et al (1996), do not clearly state where they recruited their simulated patients. Therefore, simulated patients are not a homogeneous group and their only common characteristic is that of simulating real patients. This raises the question of whether such a diverse group can be trained to behave in a ‘standardised’ way.

Uses of simulated patients in clinical and educational settings

Teaching communication skills

This is the main use in medical education, where the use of simulation gives students the opportunity to be involved in approximations of real-world settings. Here they are confronted with the challenging task of establishing a relationship while eliciting clinical information. The major advantage of effectively devised simulations is that they can simultaneously have most of the engaging qualities of reality while being explicitly controlled and safe (Jason et al, 1971). Other advantages include the role simulators have in giving direct feedback to medical students on their performance and their being readily available for teaching purposes (Sanson-Fisher & Poole, 1980).

In a well-controlled study, Sanson-Fisher & Poole (1980) demonstrated the validity of using simulated patients in the assessment of medical students’ interpersonal skills. The simulators were psychiatry out-patients and the dependent measure was a rating of empathy based on a retrospective review of audio recordings. They demonstrated that student performance, on a rating of empathy, was not significantly different with genuine as opposed to simulated patients. In addition, they reported that students were unable to discriminate between persons simulating a patient role and those who presented a real history.

Teaching clinical skills

Norman et al (1982), in a study within a postgraduate setting, showed that simulated patients could be used in areas that went beyond communication skills alone. Using a sample of 10 residents within hospital and family medicine they compared residents’ performance on four real patients with chronic stable conditions and on four simulators coached to present the same problem. They found that there were no significant differences in the number of questions asked in the history, physical examination findings, diagnosis considered or investigations proposed by the residents. Interestingly, the residents elicited more historical information from the simulated patients, but this was found to be due to a single case in which the real patient, a woman with multiple sclerosis, had memory loss. Residents also correctly identified...
67% of the patients as real or simulated (against a chance 50%). Arguably, a problem with the studies by both Norman et al (1982) Sanson-Fisher & Poole (1980) is whether they can be generalised beyond the settings and patient problems involved.

**Monitoring doctors’ performance**

Another use of simulated patients is in the monitoring of doctors’ performance. For example, Rethans & Boven (1987) in The Netherlands introduced simulated patients into 48 general practitioners’ (GPs’) surgeries and collected data about their performance. The simulators described symptoms of a urinary tract infection and the researchers were interested in whether the GPs acted according to a consensus standard for managing a patient presenting with this condition. The GPs did not detect the simulators and GP performance was on the whole shown to be a more accurate reflection of actual practice than is elicited by written simulations.

Rethans et al (1991) repeated this study for other clinical problems such as headaches, diarrhoea, diabetes and shoulder pain and found similar results, which led them to conclude that ‘standardised patients may be the method of choice in the assessment of quality of actual care of doctors’. They state that existing methods for performance assessment, such as written tests and clinical examinations, have doubtful reliability and validity. Although audio and video recordings of consultations can be reliable and valid, they feel that the limited control over patient characteristics makes it difficult to compare performance between doctors.

A study by McLure et al (1985) introduced trained patients with uncomplicated rheumatic disease into the consulting rooms of 26 family physicians in Arizona. The physicians’ ability to collect diagnostic information and formulate a plan was investigated. They found that despite the fact that most of the physicians neglected areas such as psychosocial impact and mental health issues, the majority made an adequate assessment and virtually all developed an adequate care plan. They felt that the simulated patient method provided an unobtrusive way of auditing physicians’ encounters and that the technique can be used in peer review and the planning of training programmes. It is interesting to note, however, that additional postgraduate education does not seem to change the practice of doctors (Sibley et al, 1982).

**Simulated patients and clinical examinations**

The traditional oral examination has been criticised for its inherent unreliability and poor validity (Hodges et al, 1997). Data gathered by the National Board of Medical Examinations in the USA (1960–1963), involving over 10 000 medical students, showed that the correlation of independent evaluations by two examiners was less than 0.25 (Hubbard et al, 1963). It has also been demonstrated that the luck of the draw in selection of examiner and patient played a significant role in the outcome of postgraduate oral examinations in psychiatry (Leichner et al, 1984). Indeed, Leichner et al went on to suggest that the use of OSCEs may be a feasible means of improving the validity and reliability of oral examinations (Leichner et al, 1986).

An advantage of simulated patients over real patients is that of allowing different students to be presented with a similar challenge, thereby reducing an important source of variability (Norman et al, 1985). Other advantages include their reliable availability and adaptability, which enables the reproduction of a wide range of clinical phenomena tailored to the student’s level of skill. In addition, they can simulate scenarios that may be distressing for a real patient, such as bereavement or terminal illness (Sanson-Fisher & Poole, 1980).

A disadvantage of simulated patients is that they may become ideal ‘textbook cases’, to which real patients, with all their idiosyncrasies, do not often conform. Another disadvantage, discussed by Hodges et al (1997), is the expense of simulated patients, whose training and time spent performing accounted for the largest proportion of the direct cost of setting up OSCEs. Hodges et al do, however, go on to point out that this cost is more than made up for by the reduced faculty hours per annum that it takes to set up OSCEs compared with traditional oral examinations (Box 2).

**Standardisation and training of simulated patients**

The range of clinical problems that simulators can reproduce is wide and varied, but training is essential to making their performance as lifelike as possible. Arguably, this is of key importance with respect to the issue of standardisation. However, we have found little in the literature on what is meant by standardisation or on the benchmark standards to which the simulators are trained.

We feel that standardisation has two components: the validity or accuracy of performance and the reliability or consistency of performance when faced with different examinees. The first of these has been determined in a number of ways. Both content and face validity have been addressed by getting an
analysed the performance of 13 simulated patients over time and between trainees. Their study simulated patients in their consistency of performance across scenarios as real or very real. How real the simulations were: 80% described the realism of the portrayal and to ensure consistency across the simulated patients in their presentation of affect and in their response to questions.

**Box 2 Advantages and disadvantages of objective structured clinical examinations**

**Advantages**
- Simulations of real-life situations
- Close to reality
- Controlled and safe
- Feedback from the actors (simulators)
- Ready availability when required
- Stations can be tailored to level of skill to be assessed
- Scenarios that are distressing to real patients can be simulated
- The patient variable in examination is uniform across trainees

**Disadvantages**
- The idealised ‘textbook’ scenarios may not mimic real-life situations
- May not allow assessment of complex skills
- Cost
- Training issues in setting up the stations

A key requisite for achieving both accuracy and consistency of simulated patient performance is good training. Much of the literature refers to the standard methods described by Barrows (1971). For example, in a study to test the validity of psychiatric undergraduate OSCEs, Hodges et al (1997) describe the work of an experienced simulated patient trainer: ‘training for a role begins with the presentation of written material and where possible, video footage of real patients. Each simulated patient is then observed performing the role by the station’s author to verify the realism of the portrayal and to ensure consistency across the simulated patients in their presentation of affect and in their response to questions.’

**Objective structured clinical examinations and psychiatry**

In many disciplines and specialities OSCEs have been studied extensively and their reliability and validity established (Hodges et al, 1997). This has been more difficult to achieve in psychiatry examinations, and there are several reasons to believe that an OSCE might not be as valid in the assessment of psychiatric skills. First, the binary checklists typically used in most OSCEs are insufficiently sensitive to detect higher clinical components such as empathy, rapport and ethics (Cox, 1990). Second, although a typical OSCE station lasts up to 15 minutes, a traditional psychiatric interview is of 50 minutes, raising questions about the content validity of the short OSCE station. Finally, there are some who argue that complex psychiatric presentations such as thought disorder are difficult to simulate (Hodges et al, 1998).

Famuyiwa et al (1991), at the university of Lagos in Nigeria, investigated whether the OSCE examination in psychiatry was effective and valid. By comparing the OSCE scores of 123 students with criterion-based reference scores on multiple choice questions he found that the multiple choice marks correlated significantly with the OSCE marks.
Much of the work in this area, however, has been carried out by Hodges and colleagues at the University of Toronto. For example, in 1995, following the success of a number of pilot OSCEs, they included an eight-station OSCE into the curriculum as a criterion for medical students to pass the psychiatry clerkship. Subsequent analysis of this OSCE suggested that a psychiatric OSCE is a feasible method for assessing complex psychiatric skills. These skills were evaluated in three ways. First, behaviours identified as important to a successful interview were scored as ‘done’ or ‘not done’ on a binary checklist. For example, a station on depression might typically include the questions ‘Asks about sleep?’ and ‘Asks about suicidal ideation?’ Box 3 shows amended extracts from Guy’s, King’s and St Thomas’ psychiatric OSCE rating form. Second, a global rating scale for performance was scored using four 5-point scales that addressed each student’s organisation, rapport, interview-building skills and control of his or her emotions during the interview. Third, each examiner recorded a global impression of each student’s performance. They examined 192 medical students on two parallel forms (A and B) of the examination and found it to have a reasonably high interstation reliability (Form A: Cronbach’s $\alpha = 0.64$; Form B, $\alpha = 0.66$). However, they add that ‘Whilst it is generally desirable to have a reliability of 0.80 or greater for high stakes examinations … [t]he reliability we have found is adequate for decisions regarding a clerkship rotation’ (Hodges et al., 1997).

We therefore feel that the following issues need consideration with regard to the introduction of OSCEs into Part I of the Royal College of Psychiatrists’ membership examinations:

(a) In view of this being a ‘high-stakes’ postgraduate examination, how are reliability and validity being established?
(b) What is the added value of OSCEs over the current method of examination?
(c) What skills can be adequately tested by binary checklists?
(d) How can key psychiatric skills (e.g. empathy and building rapport with patients) be assessed?

Is there a burden to simulated patients?

It is worth bearing in mind that the often highly emotional nature of simulated patients’ roles can have a residual effect on the simulators. Hodges et al (1997) note significant sequelae when simulated patients are required to play difficult roles. These include difficulties emerging from the characters, exhaustion, euphoria and, more seriously, sleep disturbances and heightened levels of anxiety, anger or sadness. They suggest that great care be taken in the selection of simulated patients and that debriefing and monitoring of simulated patients are essential. In a 5-year longitudinal study examining the impact of participation as a simulated patient on the simulators’ own health care perceptions, Rubin & Philp (1998) found that, although overall the simulated patients’ perceptions of their interactions with their own doctors were positive both before and after participation in OSCEs, perceptions of their own health care was significantly worse at 1 year post-OSCE. Again, they suggest the need for debriefing post-OSCE.

Conclusion

The use of simulated patients has a relatively long history. Simulated patient standardisation, however, is a poorly defined term, but it involves accuracy and consistency of performance. Research in OSCE examinations has shown that reliability and validity can be achieved, but it is dependent on adequate and standardised training methods.

With careful planning, it does seem that OSCEs are a feasible means of testing postgraduates in psychiatry, and attempts to make the examination more valid and reliable are to be welcomed. There is the possibility of a burden to those who simulate patients, but this can be minimised by careful selection and debriefing following the examination.
When considering the use of OSCEs to test postgraduates we feel it is important to remember that their nature is to break down clinical skills into small ‘testable’ tasks. This runs the risk of training doctors who are very good at performing these piecemeal tasks without being able to assimilate them into a coherent assessment. An analogy might be a pianist who can play beautiful scales and arpeggios but cannot play a complete sonata. We would hope that senior house officers preparing for OSCEs do not forget how to take a history, make a diagnosis and formulate a management plan. Even though this will be tested in Part II of the membership examinations, we feel these are essential skills for all doctors, from house officer to consultant.

References


Multiple choice questions

1. With regard to the selection of simulated patients:
   a) they are always professional actors
   b) they may be real patients
   c) careful selection is unnecessary
   d) housewives have been used
   e) it is inadvisable to use students of acting.

2. The following are interchangeable with the term ‘simulated patient’:
   a) surrogate patient
   b) pseudopatient
   c) pretend patient
   d) standardised patient
   e) real patient.

3. Uses of simulated patients include:
   a) teaching communication skills to medical students
   b) reducing variability in examinations by presenting different students with the same challenge
   c) teaching clinical skills to postgraduates
   d) psychiatry examinations
   e) monitoring the performance of doctors.

4. Simulated patients have the advantage of:
   a) never being detected
   b) being more intelligent than real patients
   c) being able to play scenarios which real patients may find distressing
   d) being cheaper than real patients
   e) having no personal emotions that might influence the doctor–patient relationship.
The introduction of objective structured clinical examinations (OSCEs) into the Royal College of Psychiatrists’ membership (MRCPsych) examinations follows a comprehensive review and reform of College examinations. In responding to the paper by Wallace et al (2002, this issue), I will set out the context of the changes in the College examinations, identifying the justification for these changes, and respond to the specific queries raised by Wallace et al.

**Context**

The aim of education is to ensure that students learn and know specific facts, comprehend the principles underpinning these facts, demonstrate the ability to analyse and evaluate the source of these facts and, furthermore, show an ability to synthesise information in order to produce new (that is, original) work. Assessments in their various forms attempt to test whether students can demonstrate mastery in these domains. For example, traditional multiple choice questions (MCQs) test for factual knowledge, and newer MCQs in the form of extended matching items (EMIs) test for the application of factual knowledge to specific situations. In other words, the test methods are directed at specific domains.

In the College examinations, the critical review question paper tests the candidate’s ability to analyse and evaluate information presented in research reports and the essay paper tests the candidate’s ability to synthesise information and communicate it fluently in written format. In medicine, it is important also to test for competence in practical, clinical skills. This includes competence in particular performances such as interviewing the patient as well as competence in the application of knowledge to unique situations. The clinical examinations attempt to test mastery of skills and competence as well as application of knowledge.

**Methodological problems**

It is true that all assessment methods have weaknesses. These weaknesses are all well rehearsed.
Multiple choice question papers have the advantage of being highly reliable but concerns remain as to their validity. However, the EMI format is recognised as being more likely to test clinical reasoning, that is, the application of clinical knowledge to practical situations. This is another way of saying that all tests continue to be refined with a view to reducing their acknowledged weaknesses.

Essay papers have the advantage that in-depth knowledge and understanding of a given area can be tested. But the scope for sampling a wide area of knowledge is markedly reduced as most essay papers require the candidates to write only one or two essays, whereas with MCQ papers a wide area of knowledge can be examined. The reliability of essay papers is low compared with MCQ papers. Thus, systems have to be found to improve the reliability of essays. Training of essay markers, regular auditing of marked scripts and double-marking as appropriate are methods widely used to reduce the potential variability in the marking of essays.

Before the introduction of the critical review paper the ability to analyse and evaluate data was untested in the College examinations. The methodology for this paper was invented specifically for this purpose. There are no significant issues regarding the content validity of the paper, but it could be argued that the terrain that it currently covers is narrow. The paper could be extended to include the appraisal of qualitative research data and data from other forms of scholarship, including the analysis and evaluation of books, essays and journalistic writings.

The traditional method of examining clinical skills is the single long case. This method has been adopted by medical schools and postgraduate professional examinations all over the world since it was first introduced in Cambridge in 1842. As a method, it has face validity. A candidate sees a real patient, takes a history and conducts an examination of the patient with a view to discussing the case with examiners. The examination mirrors clinical reality and the candidate’s competence is judged by senior colleagues. The problems with the traditional long case are self-evident. A candidate’s competence is determined by his or her performance on a single case. In clinical practice we would be wary of reaching major decisions on the basis of a single case report and the argument goes that we should be equally wary of such decisions in clinical examinations. Furthermore, it is acknowledged that the outcome of the examination for the candidate can be adversely influenced by factors such as the difficulty of the case and the cooperation of the given patient. Examiners can also have undue influence: a single ‘rogue’ examiner can unfairly influence the outcome for the candidate. These concerns about the traditional long case are not necessarily fatal to the long case as a method of examining candidates. However, we are obliged to find solutions to the problems described above. One solution to the traditional long case is the OSCE.

Objective structured clinical examinations allow a candidate’s clinical competence over a relatively wide area to be sampled. Usually, in excess of 10 OSCE stations are used. In the College’s OSCEs, there will be 12 stations. Undue influence of one examiner on the outcome for any candidate is much reduced. Thus, the reliability of the examination is markedly better than it is for the single long case. There are concerns, though, about face validity. Actors take the part of patients, complex clinical tasks are deconstructed into their component parts and it is the component parts that are tested. Thus, it could be argued that the skill to examine a single case and make sense of it in all its complexity is not examined. Furthermore, especially in general medicine and surgery, physical examination findings are often normal because simulated patients are normal people. This raises issues about the test’s capacity to assess the candidate’s ability not merely to conduct a sound physical examination but also to recognise and to describe abnormality.

What is clear from the discussion above is that a fair, reliable and valid examination will have to adopt a multiplicity of methods to assess candidates. The more methods, and convergence of results from these varying methods, the more confident we will be that the outcome for candidates is an accurate reflection of their knowledge and skills. This is the so-called method of triangulation, and it is the overall approach that the College takes. Thus, in the MRCPsych examinations, Part I candidates have to satisfy the examiners in a modified MCQ paper, an EMI paper and an OSCE. Part II candidates will have to satisfy the examiners in a modified MCQ paper, an EMI paper, a critical review paper, an essay paper, a traditional long case and a structured oral examination.

Response

Wallace et al (2002) pose a number of questions directly about OSCEs. I will now consider each of these in turn.

In view of this being a ‘high-stakes’ postgraduate examination, how are reliability and validity being established?

The validity of examinations can be determined in a number of ways. Usually, an examination must relate to a published curriculum. The new College
curriculum (Royal College of Psychiatrists, 2001) was developed following wide consultation within the College. This presupposes that any examination that follows this curriculum derives its legitimacy from it, that is, it validly examines what the body of the College has determined as requisite to the proper practice of psychiatry in the UK. There are ancillary methods for ensuring validity. A ‘blueprint’ for the preparation of the examination must be developed. A blueprint is simply a grid or map developed from the curriculum that allows questions to be commissioned or the examination itself to be audited. Such a blueprint exists for the College’s OSCE. It is also possible to test for validity by carrying out a survey of clinicians and trainees to assess how far the questions reflect clinical realities in psychiatry. The College has every intention of carrying out such a survey when appropriate. Predictive validity, as opposed to content validity, can be tested by exploring how far performance on the OSCE predicts performance on other aspects of the examination or on career progression. Aspects of the predictive validity of the OSCE will be investigated in the future.

Objective structured clinical examinations are recognised as highly reliable examinations. Data collected from the College’s pilot OSCE, yet to be reported, confirm this. The $\kappa$ score for the examination as a whole was about 0.8. The second pilot OSCE took place in April 2002, and the statistic of the OSCEs will shortly be reported.

### What is the added value of OSCEs over the current method of examination?

Objective structured clinical examinations have all the advantages discussed above. They allow a wide area of skills to be tested and reduce the impact of any one examiner on the overall outcome for the candidate. They also allow the College to test areas of practice currently unexamined. These include: the ability to communicate diagnosis and treatments to patients and their relatives; physical examinations; interpretation of results; and communicating complex judgements to other clinicians, including nurses, physicians and senior psychiatrists.

### What skills can be adequately tested by binary checklists?

The marking schedule is not binary. The schedule is a 5-point scale as used in other College clinical examinations. OSCEs are used in the General Medical Council Professional Language and Assessment Board (PLAB) examinations and have been introduced by the Royal College of Physicians, among others. There is no reason to think that there is any problem with marking OSCEs. OSCEs are objectively marked. This means that the weighting of particular objectives within each OSCE station is determined before the examination and the examiner’s task is to award marks for each objective as listed on the mark sheet. Whether the candidate passes or not is determined by his or her performance on these objectives and by the relative weighting of these objectives.

### How can key psychiatric skills such as empathy and building rapport with patients be assessed?

All the OSCE stations in the College examinations will have communication skills as an objective. In practice it is not difficult to identify the candidate who is unable to establish rapport and empathise with the patient. Inability to use all the well-known interviewing techniques such as open-to-closed cone questions, summary and reflective statements will indicate to examiners that this is a poor candidate. In the College’s pilot OSCE held in April 2002, the reliability of the simulated patients’ assessment of the candidates’ communication skills was tested. In North America, simulated patients’ opinions are taken into account in determining the candidate’s final mark. There is no intention to do this in the College examinations. However, we will have data, following the next pilot examination, to explore this issue in more detail.

### Conclusion

The changes to the College examination, of which OSCEs are only an example, demonstrate the College’s commitment to continue to improve its examinations in line with best current evidence. In many areas, the Royal College of Psychiatrists’ approach is in advance of other medical Royal Colleges. The aim is to have the fairest, most valid and reliable examination that is possible.

### References

