STANDARDIZED PATIENT

Dr Mohammad Arif bin Kamarudin

Assistant Head of SP Unit,

Department Of Medical Education,

Faculty Of Medicine,

Universiti Kebangsaan Malaysia

Malaysian Journal of Medicine and Health Sciences (ISSN 1675-8544); Vol. 11 (2) June 2015: 29-34

Training standardized patients for undergraduate Psychiatry examinations: experience of a Malaysian university

Suzaily Wahab*,[1] Rosdinom Razali[1], Ahmad Khaldun Ismail,[2]
Mohammad Arif Kamarudin,[3] Noorlaili Mohd Tohit,[4] Ruth Packiavathy Rajen Durai,[5]
Nabishah Mohamad,[3] Harlina Halizah Siraj[3]

[1]Department of Psychiatry, University Kebangsaan Malaysia Medical Centre, Cheras, Kuala Lumpur, Malaysia
[2]Department of Emergency Medicine, University Kebangsaan Malaysia Medical Centre, Cheras, Kuala Lumpur, Malaysia
[3]Department of Medical Education, University Kebangsaan Malaysia Medical Centre, Cheras, Kuala Lumpur, Malaysia
[4]Department of Family Medicine, University Kebangsaan Malaysia Medical Centre, Cheras, Kuala Lumpur, Malaysia
[5]Department of Nursing, University Kebangsaan Malaysia Medical Centre, Cheras, Kuala Lumpur, Malaysia

ABSTRACT

Simulated/ standardized patients (SPs) have become one of the significant components in today's medical education and students' assessment. Some differences exist in the training method of SPs for psychiatry examinations compared to other medical disciplines. This brief report highlights the challenges encountered in the training process and methods to overcome those challenges. A well-structured, intensive training remains as one of the most important factors in ensuring standardization of SPs for psychiatric examinations.

Keywords: standardized patients, examination, training, psychiatry

SP IN PSYCHIATRY TRAINING

- students found the experience of engaging in a mental health simulation with standardized patients a positive experience. (Louise,2014)
- allows students to practice their communication skills and improving their confidence level in conducting mental status examination and suicide risk assessment by reducing anxiety. (Yong-Shian, 2016)
- the advantage to nursing students was the ability to improve their interviewing skills (bipolar disorder, anxiety, and schizophrenia) in a safe educational environment before encountering these patients in a clinical experience. (Doolen, 2013)
- SP's quality of role playing was evaluated as the poorest while playing the psychiatric disorder "depression/suicidal tendencies." (Monika, 2018)

THREE PARTS

- Part 1- What is SP
- Part 2- How to write SP script
- Part 3- How to train SP

HISTORY

- Dr. Howard S. Barrows, M.D., a neurologist and medical educator, created the first standardized patient in 1963
- Barrows experienced difficulties when he tried to find patients with specific findings for Psychiatry and Neurology board examinations and realized that some findings could be SIMULATED.

SIMULATED PATIENT

- a lay person (a normal person) who has been instructed carefully to be an actual patient in terms of presenting the signs and symptoms (Barrows 1963)

Simulated patient to standardized patient

STANDARDIZED PATIENT

- People who are trained to simulate a patient's illness in a standardized way (Wallace P)
- SPs are trained to provide a standardized response with little variation between encounters.

EXAM/OSCE SETTING

WHY STANDARDIZATION IS IMPORTANT?

EXAMPLE 1

- LO: Students should be able to demonstrate complete history taking of patients with suspected pneumonia

Question : Why are you in the ward?

Patient A

“I have fever”

Patient B

“I have fever, cough and difficulty breathing for one week”

Patient C

“I am currently being treated for pneumonia”

EXAMPLE 2

Student A: What medication are you on?

Patient A: I take two types of medicine

Student A: Can you tell me the name?

Patient A: I don't remember the name.

Student A: Can you tell me the shape of the medicine?

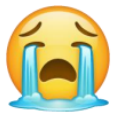
Patient A: Welll….hmmm…one is round in shape and the other is… oval in shape I think…sorry… I'm not sure"

Student A: Do you remember how frequent you take the medicine?

Patient A: I take them only when I remember to take them.

Student A: No.. I mean how frequent you should take them. As instructed by your doctor.

Patient A: That I don't remember.

Student A: 

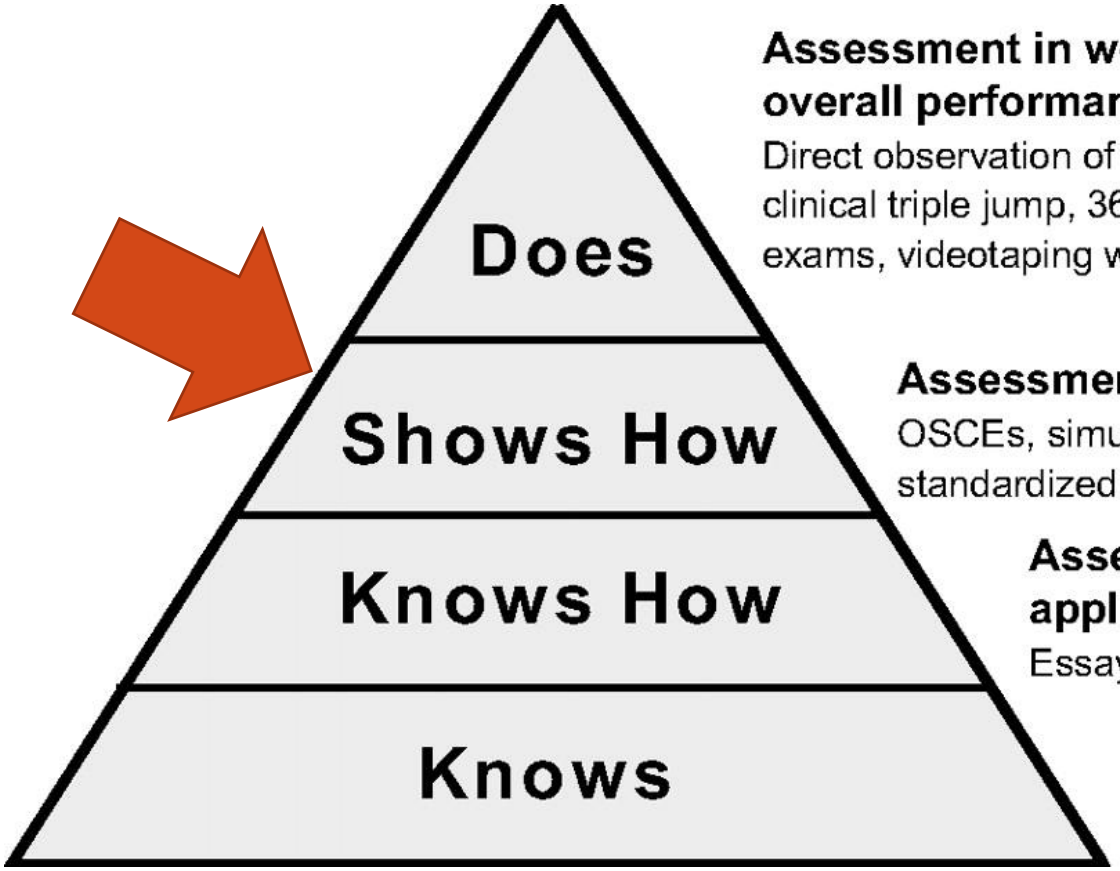
EXAMPLE 2

Student B: What medications are you on?

Patient B: "I take amlodipine and valsartan, both taken once daily"

Student B: 

WHEN TO USE



Assessment in work environment; focus on overall performance, not components

Direct observation of learner performance, portfolios, clinical triple jump, 360 assessment, clinical competency exams, videotaping with follow-up review

Assessment in controlled situations

OSCEs, simulations, lab practicals, standardized patients

Assess capacity for clinical-context application

Essays, triple jump, case-based MCQs

Test factual recognition

Context-free MCQs, reports written by students, oral exams

Teaching-Learning & Assessment

- Clinical skills (history taking, physical examination)
- Procedural skills (+/- hybrid)

- Communication skill
- Trauma moulage
- +/- Feedback

ADVANTAGES

- SPs are being increasingly regarded as alternatives to provide medical students early experiences in clinical skills. (Colliver JA)
- SP encounters can be arranged at any time and in any setting, unlike encounters with real patients whose presence in hospital or general practice is difficult to control
- SPs can provide a reliable learning experience for students, offer valuable feedback and could be used to assess clinical skills acquisition by students (Van der Vleuten)
- They may serve as a transition to the real patient and provide students with an opportunity to improve their history taking and physical examination skills.

- The SP can be manipulated for educational purposes in a manner which may be difficult with real patients.
- SPs are increasingly being used instead of real patients during the Objective Structured Clinical Examination (OSCE) as they provide a consistent clinical scenario and may help reduce variability between students' experiences. (Adamo G)
- Some centers are also using SPs as examiners and they evaluate students using a checklist. (McLaughlin et al)

DISADVANTAGES

- Time-consuming
- Financial factor
- Not all signs can be simulated

PART 2

SP SCRIPT WRITING

WHAT TO DO?

- Assessment blueprint (what to assess)
- Design scenario/script and checklist
- Vetting of the scenario/script
- Book SP
- Train SP
- Exam (revisit with SP before exam start)
- Debriefing & Feedback

THINGS TO CONSIDER

- Level of examinees
- Type of exam (Long case, OSCE)
- Focus of exam (history taking, physical/systemic examination skills, communication skills)
- Duration of the exam
- *availability of the SP

IMPORTANT POINTS

-Use layman term (no medical jargon)
-The script is designed for the SP
-Treat the script as confidential (don't allow SP to take the script home)
    - SP can make notes
    - remind SP to ensure confidentiality
-Do not announce the use of SP to examinees
-Beware of *unexpected* questions from students
-Decide when to give-in?
-Cross check the script with the examiner checklist
-Most important thing - STANDARDIZATION

. . . IN PSYCHIATRY

- the highest challenge for the SPs was to understand the psychopathology of psychiatry patients and being able to play the roles well.
- psychiatric diagnosis mostly relies on the history and mental state examinations which include detail observation of appearance, behaviour and emotion.

PSYCHIATRY CASES USING SPS (UKMMC EXPERIENCE)

- SCHIZOPHRENIA
- BIPOLAR MOOD DISORDER
- ANXIETY DISORDERS
- SUBSTANCE USE DISORDERS

*only experienced SPs (those who have been actively involved in the SP programme for at least a year and showed good performance) to act the role of psychiatric patients

SP SCRIPT TEMPLATE

Standardized Patient Script

Acute Myocardial Infarction

Patient's Detail		
SP Name Mustafa	**Past Medical History** Hypertension for 15 years	
Age/Race/Gender 52/Malay/Male	**Medications** Anti-hypertensive agent	**Allergies** Nil
Presenting Complaints (Verbatim) "I have chest pain"		

Presentation
Affect Grimacing in pain. Leaning forward. Right hand on left chest. Talking intermittently.

Prop List
Clothes Office attire. A pack of cigarette in front pocket of shirt.

Appearance
Tidy.

Setup
Location of Interview Emergency Department
Physical Characteristics Obese

Past History

History of Presenting Illness
Chest pain for one hour prior to presentation.
Pain started when sitting, attending a meeting at work.
Left sided, radiated to neck and jaw, feeling of tightness and heaviness, getting worse (from 5/10, now 7/10). No relieving or aggravating factors.
Associated with sweating, numbness of left arm and difficulty breathing.

Past medical history:
Hypertension for 15 years and on an anti-hypertensive agent but do not remember the name.
White in colour, round in shape, took once a day. Often forget to take them because was busy with work.
Visiting multiple clinics (general practitioners), BP reading was not well-controlled.
Unaware if having other diseases because never underwent full medical check-up or screening.

Social history:
Smoke 20 cigarettes every day since secondary school.
Does not consume alcoholic beverages.
Does not abuse drugs.
No history of sexual promiscuity.
Working as a manager in private sector. Stressful working environment. Handling a new project.
Does not exercise regularly.
Loves to eat fried foods.

Past surgical history:
Nil.

Hospitalizations:
Nil.

Allergies:
Nil.

Family History

Parents

Father died at the age of 53 years old due to heart problem.
Father had history of hypertension for many years but not sure if he had any other medical conditions.
Father was a smoker too, heavy but cannot be certain about the number of cigarettes he smoked every day.
Mother has diabetes mellitus and currently staying in Penang. Not sure if mother has any other medical conditions.

Siblings

First child in the family with three younger brothers and one younger sister.
Not sure whether siblings have any chronic diseases.

Own
Married with two wives and seven school going children.

Check-list for the SP:

	Question	Answer
1	Why do you come today?	"I have chest pain"
2	For how long?	"One hour"
3	Where is the pain? Can you show me the location?	"On my left chest"
4	Can you describe the pain?	"I feel heaviness and tightness around my left chest"
5	Is the pain improving or worsening?	"Worsening"
6	Was there anything that made the pain worse?	"I'm not sure doctor"
7	Did the pain travel to other part of your body?	"Yes...the pain radiated to my jaw and neck"
8	Did you vomit?	"No"
9	Did you sweat?	"Yes"


HOW TO TRAIN YOUR SP?

WHY?

- to familiarize SP with the script/case
- to get feedback from SP and other trainers about the script
- STANDARDIZATION
- *feedback

- Must be done with the presence of all SPs involved in the exam
- Usually done 3-7 days before the exam
- Usually take 2-3 hours depending on the difficulty of the case and number of SPs
- If the case/scenario requires special props e.g walking stick, wheelchair etc, ensure that these will be made available to all SPs during training session

SESSION 1

- General orientation
- to familiarize the SPs with the case
- SPs' role in the case is defined
- SPs are provided with information about the purpose of the exam/assessment
- logistics of the examination or assessment where and when this case will be used
- The SPs' rights and responsibilities are discussed and clarified.

USE OF VIDEO

- to show an example of the mental state to help SPs understand their role better.
- also help in introducing psychopathology such as flight of ideas, loosening of association and psychomotor retardation to the SPs

THE STEPS

- Divide SPs into smaller groups
- Distribute the script to SPs
- All SPs to read the script aloud (allow SP to stop to ask for clarification)
- The trainer clarifies the patient's personality, manner, attitude and how SP should portray it (body language or gestures or verbal responses)
- Trainers can also take the SPs' perceptions and ideas and incorporate them
- Allow SPs time to digest and memorize the details of the script (10-20 mins)

*treat the script/case material as confidential

SESSION 2

- Role-play
- SPs are trained on physical examination manoeuvres (if any)
- Select one SP and role-play. First with lecturer/trainer as student.
- The trainer will role-play students who are average, above average and below average
- As each SP role-plays the case, the other SPs observe and comment on the performance
- repeat role-play with feedback with all SPs until the performance of the SP matches that of the case and standardization is achieved

*review checklist

SESSION 3*

- Feedback training
- giving feedback from the patient's perspective
- specific feedback in the form of behaviors that the student exhibited and how the patient felt as a result of the behaviors
- sandwich technique with a positive comment followed by a constructive suggestion and closing with another positive comment.

Give positive feedback

Provide constructive criticism

Give positive feedback

STUCK IN THE ROLE (TROUBLE LETTING GO)

- Emotionally distressed/depressed
- Tired
- Imagining that they have symptoms or the illness that they acted out recently.
- Playing the role over and over again in their heads after the scenario has finished

SP DEBRIEFING (LEAVING THE ROLE BEHIND)

- get together afterwards to let off some steam. Take some time to laugh and joke around!
- go and visit a friend to take mind off the job.
- try listening to some soft music or turn the television on.
- read a good book.
- engage in physical activity.

PRECAUTIONS

- SP should think sensibly and realistically when accepting a demanding role.
- Avoid roles that could be too close to tough situations that they had experienced in the past.
- If they took on a tough role in the past and had an issue, put a plan in place the next time they take on such a role.
- If they have done a role more than once and cannot let go no matter how hard they try to forget about it, it might be time to ditch the role for something new or different.
- the maximum time allocated for each SP to be interviewed by students was only one hour before being replaced by another SP.

POST-EXAM

- SP's performance feedback form

Universiti Kebangsaan Malaysia
The National University of Malaysia

CONFIDENTIAL

ASSESSMENT OF STANDARDIZED PATIENTS (SP) BY EXAMINER

Dear examiner,

As part of our continuous effort to monitor and improve our SP programme, we would like to have your feedback on our SP's performance.

Real name of the SP : .....

Scenario : .....

Name of examiner : .....

Date : .....

Please rate (√) the following items according to the rating scale below

1 Complete Disagreement | 2 Disagree | 3 Unsure | 4 Agree | 5 Complete Agreement

AUTHENTICITY OF THE SP	RATING
1. SP able to withhold information as required	1 2 3 4 5
2. SP stays in his/her role all the time (consistency)	1 2 3 4 5
3. SP challenging / testing the students	1 2 3 4 5
4. SP simulates physical complaints realistically	1 2 3 4 5
5. SP appearance fits the role (dressing, mood, affect, physical charateristic)	1 2 3 4 5
6. SP able to maintain no threatening approach / not over acting	1 2 3 4 5
7. SP able to provide information as required	1 2 3 4 5

Your comments / suggestions regarding SP :

.....

.....

.....

Please return this form to Department of Medical Education.

REFERENCES

- Adamo G. Simulated and standardized patients in OSCEs: achievements and challenges 1992-2003. Med Teach. 2003;25:262–70.
- Barrows HS. An overview of the uses of standardized patients for teaching and evaluating clinical skills. AAMC. Acad Med 1993; 68(6): 443-51. doi:

  10.1097/00001888-199306000-00002
- Colliver JA, Williams RG. Technical issues: test application. Acad Med. 1993;68:454–60.
- Collins J, Harden R. The use of real patients, simulated patients and simulators in clinical examinations 2004. Association for Medical Education in Europe (AMEE) Guide No 13. Available from: http://78.158.56.101/archive/MEDEV/static/uploads/resources/amee_summaries/Guide13summaryMay04.pdf
- McLaughlin K, Gregor L, Jones A, Coderre S. Can standardized patients replace physicians as OSCE examiners? BMC Med Educ. 2006;6:12.
- Wallace P. Following the threads of an innovation: the history of standardized patients in medical education. Caduceus 1997; 13(2):5–28.

Examiner Rating

A=Excellent/B=Very Good/C=Clear Pass/D=Borderline/E=Clear Fail ABCDE

Clear Fail:

- Disorganized approach, no evidence of planning – tends to random actions, process and questions
- Unable to synthesize findings, or reach a diagnosis/plan

Borderline

- Able to commence station, but often uncertain, and struggles to proceed to completion
- Some organisation of approach, but 'formulaic' with no flexibility (e.g. 'lists' of questions for patients) and no evidence of reasoning/discrimination

Clear Pass

- Systematic overall approach to station/task
- Demonstrates sufficient organization to permit completion of task with some evidence of flexibility of approach
- Able to summarize (e.g. present history/explain) and manage additional questioning with evidence of reasoning

Very Good Pass

- Clearly professional approach to station. Good levels of organization with clear evidence of flexibility
- Clearly able to synthesize findings, or reach a diagnosis/plan
- Clear evidence of planning, ability to summarize and manage questioning

Excellent

- Overall superior approach – excellent organizational skills, and fluent management of task in hand
- Flexible, adaptive approach to changing circumstances within a station – e.g. reacting to patients, emergency situations
- High levels of professionalism and clinical reasoning – applies knowledge critically when questioned.

FORM B - Domain-based, rating scales

A = Very Good B = Good C = Acceptable D = Poor E = Very Poor

History-taking/ Information gathering station

1. General approach to patient **A B C D E**

Appropriate introduction (full name & role)

Checks patient's/ relative's name

Explains what interview/task will be about & checks consent

Start with an open question & listens without interruption

2. Information gathering: clinical content **A B C D E**

As appropriate to the station

3. Information gathering: clinical communication **A B C D E**

Questioning skills: (appropriate blend of open and closed questions, clarity, avoids or explains jargon)

Listens actively: (attentive, pick up cues, responds to answers, does not repeat questions)

Organised: (systematic, summarises, signposts change in focus of questions)

Closure: (e.g.explains next steps, thanks patient)

4. Findings **A B C D E**

Accurate summary of history

5. Diagnosis **A B C D E**

Plausible differential

6. Rapport and Professionalism **A B C D E**

Shows interest, respect and concern for pt

Appropriate non verbal communication

(eye contact, appropriate use of touch, maintains comfortable distance from pt)

Professional behaviour:

(e.g. attitude, maintains dignity and privacy)

CLINICAL SKILLS (Physical examination)

1. General approach to patient **A B C D E**

Introduction and orientation

(Name and role; purpose of the examination; explains what examination will involve; consent)

2. Clinical skills/physical examination **A B C D E**

Important features

Appropriate/ acceptable examination method

Performs examination/ procedure in fluent and organised manner

3. Findings **A B C D E**

Clear and accurate explanation of findings

Clear and accurate summary

4. Diagnosis **A B C D E**

Plausible differential diagnosis

5. Rapport and professionalism **A B C D E**

Gives clear instructions to patient through examination

Treats patient courteously and maintains dignity throughout

Leaves patient comfortable

6. Data Interpretation **A B C D E**

Accurate interpretation

Diagnosis

7. Management **A B C D E**

As appropriate e.g. investigations, treatment, admission, referral

CLINICAL SKILLS (Procedures)

1. General approach to patient **A B C D E**

Introduction and orientation

(Name and role; purpose of the procedure; explains what procedure will involve; consent)

2.Clinical Skills: Procedure **A B C D E**

Specific items for the performance of the task

Appropriate/ acceptable method

Performs procedure in fluent and organised manner

3. Rapport and professionalism **A B C D E**

Gives clear instructions to patient through examination

Treats patient courteously and maintains dignity throughout

Leaves patient comfortable

4. SP to mark **A B C D E**

I felt that the students showed I felt that the students showed respect and treated me with dignity

Explanation / Information-giving / Negotiation

1. General approach to patient **A B C D E**

Appropriate introduction (full name & role)

Checks patient's/ relative's name

Explains what interview/task will be about & checks consent

Start with an open question & listens without interruption

2. Explanation/information giving/negotiation: clinical content **A B C D E**

As appropriate to the task

3. Explanation/information giving/negotiation: communication **A B C D E**

Explaining skills (chunks information, clear & given at patient's level of understanding, well paced, some dialogue with pt)

Checks understanding of main points

Negotiating plan

Closure (reiterates next steps, thanks patient)

4. Rapport and Professionalism **A B C D E**

Shows interest, respect and concern for pt

Appropriate non verbal communication (eye contact, appropriate use of touch, maintains comfortable distance from pt)

Professional behaviour (e.g. attitude, maintains privacy)

OSCE domain rating scale scoring with rubrics					
	Scoring				
Domain	**A = Very Good**	**B = Good**	**C = Acceptable**	**D = Poor**	**E = Very Poor**
1. Approach to patient	Full name, role, full explanation purpose / welcoming, courteous, establishes rapport and puts patient at ease quickly	Full name and role / full and clear explanation of purpose	Full name and role / attempts to explains purpose interaction	Incomplete name / role, fails to adequately explain purpose	Fails to identify self / role or purpose of interaction / patient uncomfortable
2. Information gathering/ history taking: clinical content	Full comprehensive history including addressing patient concerns / fluent and clearly reasoned questioning / adapts to patient's answers when required	Most points of history elicited including addressing patient concerns / no major omissions / well structured approach to history	Main points of history elicited including some recognition of patient concerns / no major omissions / reasonably structured approach	Some attempt at history but with significant omissions / little apparent structure to history	Failure to elicit relevant history / major omissions throughout /disorganised with no apparent logic or order
3. Information-gathering/ history taking: communication	Completely clear questions / Avoids or explains jargon / listens actively / builds in structure using appropriate signposts and accurate summary / fluent	Completely clear questions / Avoids or explains jargon / demonstrates some active listening / generally well structured using appropriate signposts and accurate summary / reasonably fluent	Most questions clear / avoids or explains jargon / some attempt to build in structure	Many questions unclear / Some use of or failure to explain jargon / often does not listen to answers	Totally unclear questions / repeatedly uses or does not explain jargon or uses leading or multiple questions / does not listen to answers

4. Clinical skills/ physical examination					
5. Explanation / information giving/ Negotiation: clinical content	All points covered clearly and correctly using appropriate language / demonstrates in depth understanding of topic / acknowledges clinical uncertainties where appropriate	Most points covered clearly and are factually correct / demonstrates good understanding of topic	Main points covered clearly and are factually correct / demonstrates reasonable understanding of topic	Delivers some important incorrect clinical information / demonstrates only partial understanding of topic / some lack of clarity	Delivers incorrect clinical information / demonstrates little or no understanding of topic / obviously confuses patient
6. Explaining/ information-giving/Negotiation: communication	Well paced and encourages dialogue throughout / information given in manageable amounts / thorough check of patient's understanding / demonstrates response to all patient's cues	Mostly well paced with dialogue generally encouraged / information given in manageable amounts / checks patient understanding / mostly picks up patient cues	Some dialogue encouraged / some attempt to manage information load / checks patient understanding / some recognition of patient's cues	Minimal dialogue encouraged / little attempt to manage information load or check understanding / fails to pick up patient's cues	"Talks at" the patient / no dialogue / completely fails to pick up patient cues / fails to check patient's understanding

7. Clinical skills/ procedure					
8. Findings	Reports all clinical findings correctly	Reports most clinical findings correctly	Reports the most important findings correctly	Reports some correct but also incorrect findings	Fails to elicit or report any correct findings
9. Data interpret-tation					

10. Diagnosis	Suggests all correct diagnostic possibilities	Suggests most correct diagnoses	Suggests most important diagnosis	Suggests less relevant diagnoses	Suggests incorrect diagnoses
11. Manage-ment					
12. Rapport and Professiona lism	Complete interest, respect, empathy and concern for patient / maintains patient dignity throughout / entirely appropriate verbal or non-verbal communication / inspires confidence and trust	Demonstrates interest, respect, empathy and concern for patient / maintains patient dignity throughout / appropriate verbal or non-verbal communication	Engages with patient and demonstrates interest, empathy and concern / maintains dignity / appropriate verbal or non-verbal communication	Limited engagement with patient / some inappropriate verbal or non-verbal communication / little empathy demonstrated	Failure to engage with patient / oblivious to patient's physical or emotional needs / inappropriate verbal or non-verbal communication / No empathy demonstrated
13. SP to mark (rubric will depend on station task)					

Validity threats: overcoming interference with proposed interpretations of assessment data

Steven M Downing[1] & Thomas M Haladyna[2]

CONTEXT Factors that interfere with the ability to interpret assessment scores or ratings in the proposed manner threaten validity. To be interpreted in a meaningful manner, all assessments in medical education require sound, scientific evidence of validity.

PURPOSE The purpose of this essay is to discuss 2 major threats to validity: construct under-representation (CU) and construct-irrelevant variance (CIV). Examples of each type of threat for written, performance and clinical performance examinations are provided.

DISCUSSION The CU threat to validity refers to undersampling the content domain. Using too few items, cases or clinical performance observations to adequately generalise to the domain represents CU. Variables that systematically (rather than randomly) interfere with the ability to meaningfully interpret scores or ratings represent CIV. Issues such as flawed test items written at inappropriate reading levels or statistically biased questions represent CIV in written tests. For performance examinations, such as standardised patient examinations, flawed cases or cases that are too difficult for student ability contribute CIV to the assessment. For clinical performance data, systematic rater error, such as halo or central tendency error, represents CIV. The term *face validity* is rejected as representative of any type of legitimate validity evidence, although the fact that the appearance of the assessment may be an important characteristic other than validity is acknowledged.

CONCLUSIONS There are multiple threats to validity in all types of assessment in medical education. Methods to eliminate or control validity threats are suggested.

KEYWORDS education, medical, undergraduate/ *standards; educational measurement/*standards; clinical competence/ standards; reproducibility of results.

*Medical Education 2004; **38**: 327–333*
doi:10.1046/j.1365-2923.2004.01777.x

INTRODUCTION

The purpose of this paper is to call attention to threats to validity in the context of assessment in medical education and to suggest potential remedies for these threats. Validity refers to the degree of meaningfulness for any interpretation of a test score. In a previous paper in this series[1] validity was discussed and sources of validity evidence based on the *Standards for Educational and Psychological Testing*[2] were exemplified for typical assessments in medical education. This essay addresses some of the variables or issues that tend to interfere with the meaningfulness of interpretation of assessment scores and, thereby, reduce the validity of interpretation and the subsequent usefulness of these assessments.

THREATS TO VALIDITY

There may be at least as many threats to validity as there are sources of validity evidence. Any factors that interfere with the meaningful interpretation of assessment data are a threat to validity. Messick[3]

[1]University of Illinois at Chicago, College of Medicine, Department of Medical Education, Chicago, Illinois, USA
[2]College of Education, Arizona State University West, Phoenix, Arizona, USA

Correspondence: Steven M Downing PhD, Associate Professor of Medical Education, University of Illinois at Chicago, College of Medicine, Department of Medical Education (MC 591), 808 South Wood Street, Chicago, Illinois 60612-7309, USA. Tel: 00 1 312 996 6428; Fax: 00 1 312 413 2048; E-mail: sdowning@uic.edu

Key learning points

There may be as many threats to validity as there are sources of validity evidence.

Any factors that interfere with the proposed interpretation of assessment scores or ratings threaten validity.

For all assessments, construct under-representation (CU) is a major threat to validity.

Construct-irrelevant variance (CIV) is systematic measurement error that reduces the ability to accurately interpret scores or ratings.

Face validity is not any type of true, scientific evidence of validity.

noted 2 major sources of validity threats: construct under-representation (CU) and construct-irrelevant variance (CIV). Construct under-representation refers to the undersampling or biased sampling of the content domain by the assessment instrument. Construct-irrelevant variance refers to systematic error (rather than random error) introduced into the assessment data by variables unrelated to the construct being measured. Both CU and CIV threaten validity evidence by reducing the ability to reasonably interpret assessment results in the proposed manner.

Table 1 lists examples of some typical threats to validity for written assessments, performance examinations, such as objective structured clinical examinations (OSCEs) or standardised patient (SP) examinations, and clinical performance ratings. These threats to validity are organised by CU and CIV, following Messick's model.[3]

Written examinations

In a written examination, such as an objective test in a basic science course, CU is exemplified in an examination that is too short to adequately sample the domain being tested. Other examples of CU are: test item content that does not match the examination

Table 1 Threats to validity of assessments

Written test	Performance examination	Ratings of clinical performance
Construct under-representation (CU)		
Too few items to sample domain adequately	Too few cases/OSCEs for generalisability	Too few observations of clinical behaviour
Biased/unrepresentative sample of domain	Unstandardised patient raters	Too few independent raters
Mismatch of sample to domain	Unrepresentative cases	Incomplete observations
Low score reliability	Low reliability of ratings	Low reliability of ratings/ low generalisability
Construct-irrelevant variance (CIV)		
Flawed item formats	Flawed cases/checklists/ rating scales	Inappropriate rating items
Biased items (DIF)	DIF for SP cases/rater bias	Rater bias
Reading level of items inappropriate	SP use of inappropriate jargon	Systematic rater error: halo, severity, leniency, central tendency
Items too easy/too hard/ non-discriminating	Case difficulty inappropriate (too easy/too hard)	Inadequate sample of student behaviours
Cheating/insecure items	Bluffing of SPs	Bluffing of raters
Indefensible passing score methods	Indefensible passing score methods	Indefensible passing score methods
Teaching to the test	Poorly trained SPs	Poorly trained raters

specifications well, so that some content areas are oversampled while others are undersampled; use of many items that test only low level cognitive behaviour, such as recall or recognition of facts, while the instructional objectives require higher level cognitive behaviour, such as application or problem solving; and, use of items that test trivial content that is unrelated to future learning.[4]

The remedies for these CU threats to validity are straightforward, although not always easily achievable. Written tests of achievement should be composed of test items that adequately sample the achievement domain tested. Tests must have sufficient numbers of items in order to sample adequately (generally, at least 30 items) and, if the instructional objectives require higher-order learning, these items should be written to test higher cognitive levels; items should test important information, not trivia. Construct-irrelevant variance may be introduced into written examination scores from many sources. Construct-irrelevant variance represents systematic 'noise' in the measurement data, often associated with the scores of some but not all examinees. This CIV 'noise' represents the unintended measurement of some construct that is off-target, not associated with the primary construct of interest, and therefore interferes with the validity evidence for assessment data. For example, flawed or poorly crafted item formats, which make it more difficult for some students to give a correct answer, introduce CIV into the measurement,[5] as does the use of many test items that are too difficult or too easy for student achievement levels and items that do not discriminate high-achieving from low-achieving students. Construct-irrelevant variance is also introduced by including statistically biased items[6] on which some subgroup of students under- or over-performs compared to their expected performance, or by including test items which offend some students by their use of culturally insensitive language. If some students have prior access to test items and other students do not have such access, this type of test insecurity CIV makes score interpretation difficult or impossible and seriously reduces the validity evidence for the assessment. Likewise, other types of test irregularities, such as cheating, introduce CIV and compromise the ability to interpret scores meaningfully. A related CIV issue is 'teaching to the test', such that the instructor uses actual test items for teaching, thus creating misleading or incorrect inferences about the meaning of scores.

If the reading level of achievement test items is inappropriate for students, reading ability becomes a CIV variable which is unrelated to the construct measured, thereby introducing CIV.[7] This reading level issue may be particularly important for students taking tests written in a language that is non-native to them. By using complex sentence structures and challenging vocabulary and jargon, we run the risk of underestimating the medical knowledge of any student whose first language is not English. While guessing is not a major issue on long, well crafted multiple-choice test items with at least 3 options,[8] random guessing of correct answers on multiple-choice items can introduce CIV, as the student's luck, cleverness or propensity to guess is not directly related to the achievement construct being measured.[9]

A final example of CIV for written tests concerns the use of indefensible passing scores.[10] All passing score determination methods, whether relative or absolute, are arbitrary. These methods and their results should not be capricious, however. If passing scores or grade levels are determined in a manner such that they lack reproducibility or produce cut-off scores that are so unrealistic that unacceptably high (or low) numbers of students fail, this introduces systematic CIV error into the final outcome of the assessment.

What are the solutions to these types of CIV problems? On written achievement tests, items should be well crafted and follow the basic evidence-based principles of effective item writing.[11,12] The item format itself should not be an impediment to student assessment. The reading ability of students should not be a major factor in the assessment of the achievement construct. Most items should be targeted in difficulty to student achievement levels. All items which are empirically shown to be biased or which use language that might offend some cultural, racial or ethnic group should be eliminated from the test. Test items must be maintained securely and tests should be administered in proctored, controlled environments so that any potential cheating is minimised or eliminated. Instructors should not teach directly to the content of the test. Instead, teaching should be targeted at the content domain of which the test is a small sample. Finally, passing scores (or grading standards) should be established in a defensible manner, which is fair to all students.

Performance examinations

Objective structured clinical examinations or SP examinations increase the fidelity of the assessment and are intended to measure performance, rather than knowledge or skills.[13] Performance assessments

are closer in proximity to the actual criterion performance of interest, but these types of assessment also involve constructs, because they sample performance behaviour in a standardised, ideal environment. They are simulations of the real world, but are not the real world. The performance of students, rated by trained SPs in a controlled environment on a finite number of selected cases requiring maximum performance, is not *actual* performance in the real world; rather, inferences must be made from performance ratings to the domain of performance, with a specific interpretation or meaning attributed to the checklist or rating scale data. Validity evidence must be documented to support or refute the proposed meaning associated with these performance-type constructs.

There are many potential CU and CIV threats to validity for performance assessments. Table 1 presents some examples of validity threats. Many threats are the same as noted for written tests. One major CU threat arises from using too few performance cases to adequately sample or generalise to the domain. The case specificity of performance cases is well documented.[14,15] Approximately 12 SP encounters, lasting 20 minutes each, may be required to achieve even minimal generalisability to support inferences to the domain.[16] Lack of sufficient generalisability represents a CU threat to validity. If performance cases are unrepresentative of the performance domain of interest, CU threatens validity. For example, in an SP examination of patient communication skills, if the medical content of the cases is atypical and unrepresentative of the domain, it may be impossible for students to demonstrate their patient communication skills adequately.

Many SP examinations use trained lay SPs to portray actual patient medical problems and to rate student performance after the encounter, using standardised checklists or rating scales. The quality of the SP portrayal is extremely important, as is the quality of the SPs' training in the appropriate use of checklists and rating scales. If the SPs are not sufficiently well trained to consistently portray the patient in a standardised manner, different students effectively encounter different patients and slightly different patient problems. The construct of interest is therefore misrepresented, because all students do not encounter the same patient problem or stimulus.

Remedies for CU in SP examinations include the use of large numbers of representative cases, using well trained SP raters. Standardised patient monitoring, during multiday performance examinations, is critical, so that any slippage in the standard portrayal can be corrected during the time of the examination.

For a performance examination, such as an OSCE or SP examination, there are many potential CIV threats. Construct-irrelevant variance on an SP examination concerns issues such as systematic rater error that is uncorrected statistically, such that student scores are systematically higher or lower than they should be. Standardised patient cases that are flawed or of inappropriate difficulty for students and checklist or rating scale items that are ambiguous may introduce CIV. Statistical bias for 1 or more subgroups of students, which is undetected and uncorrected, may systematically raise or lower SP scores, unfairly advantaging some students and penalising others. Racial or ethnic rater bias on the part of the SP rater creates CIV and makes score interpretation difficult or impossible.

It is possible for students to bluff SPs, particularly on non-medical aspects of SP cases, making ratings higher for some students than they actually should be. Establishing passing scores for SP examinations is challenging; if these cut-off scores are indefensibly established, the consequential aspect of validity will be reduced and CIV will be introduced to the assessment, making the evaluation of student performance difficult or impossible.

The remedies for CU in performance examinations are obvious, but may be difficult and costly to implement. Low reliability is a major threat to validity,[3] thus using sufficient numbers of reliable and representative cases to adequately generalise to the proposed domain is critical. Generalisability must be estimated for most performance-type examinations, using generalisability theory.[17,18] For high-stakes performance examinations, generalisability coefficients should be at least 0·80; the phi-coefficient is the appropriate estimate of generalisability for criterion-referenced performance examinations (which have absolute, rather than relative passing scores).[16] Standardised patients should be well trained in their patient roles and their portrayals monitored throughout the time period of the examination to ensure standardisation. To control or eliminate CIV in performance examinations, checklists and rating scales must be well developed, critiqued, edited and tried out and be sufficiently accurate to provide reproducible data when completed by SPs who are well trained in their use. Methods to detect statistical bias in performance examination ratings should be implemented for high-stakes examinations.[19] Performance cases should be

pretested with a representative group of students prior to their final use, testing the appropriateness of case difficulty and all other aspects of the case presentation. Standardised patient training is critical, in order to eliminate sources of CIV introduced by variables such as SP rater bias and student success at bluffing the SP. If passing scores or grades are assigned to the performance examination results, these scores must be established in a defensible, systematic, reproducible and fair manner.

Ratings of clinical performance

In medical education, ratings of student clinical performance in clerkships or preceptorships (on the wards) are often a major assessment modality. This method depends primarily on faculty observations of student clinical performance behaviour in a naturalistic setting. Clinical performance ratings are unstandardised, often unsystematic, and are frequently carried out by faculty members who are not well trained in their use. Thus, there are many threats to validity of clinical performance ratings by the very nature of the manner in which they are typically obtained.

The CU threat is exemplified by too few observations of the target or rated behaviour by the faculty raters (Table 1). Williams *et al.*[20] suggest that 7–11 independent ratings of clinical performance are required to produce sufficiently generalisable data to be useful and interpretable. The use of too few independent observations and ratings of clinical performance is a major CU threat to validity.

Construct-irrelevant variance is introduced into clinical ratings in many ways. The major CIV threat is due to systematic rater error. Raters are the major source of measurement error for these types of observational assessments, but CIV is associated with systematic rater error, such as rater severity or leniency errors, central tendency error (rating in the centre of the rating scale) and restriction of range (failure to use all the points on the rating scale). The halo rater effect occurs when the rater ignores the traits to be rated and treats all traits as if they were one. Thus, ratings tend to be repetitious and inflate estimates of reliability.

Although better training may help to reduce some undesirable rater effects, another way to combat rater severity or leniency error is to estimate the extent of severity (or leniency) and adjust the final ratings to eliminate the unfairness that results from harsh or lenient raters. Computer software is available to estimate these rater error effects and adjust final ratings accordingly. While this is a potentially effective method to reduce or eliminate CIV due to rater severity or leniency, other rater error effects, such as central tendency errors, restriction in the use of the rating scale, and idiosyncratic rater error remain difficult to detect and correct.[7]

Rating scales are frequently used for clinical performance ratings. If the items are inappropriately written, such that raters are confused by the wording or misled to rate a different student characteristic from that which was intended, CIV may be introduced. Unless raters are well trained in the proper use of the observational rating scale and trained to use highly similar standards, CIV may be introduced into the data, making the proposed interpretation of ratings difficult and less meaningful. Students may also attempt to bluff the raters and intentionally try to mislead the observer into 1 or more of the systematic CIV rater errors noted.

As with other types of assessment, the methods used to establish passing scores or grades may be a source of CIV. Additionally, methods of combining clinical performance observational data with other types of assessment data, such as written test scores and SP performance examination scores may be a source of CIV. If the procedures used to combine different types of assessment data into 1 composite score are inappropriate, CIV may be introduced such that the proposed interpretation of the final score is incorrect or diminished in meaning.[21]

Remedies for the CU and CIV threats to validity of clinical performance data are suggested by the specific issues noted. For CU, many independent ratings of behaviour are needed, by well trained raters who are qualified to make the required evaluative judgements and are motivated to fulfil these responsibilities. The mean rating, over several independent raters, may tend to reduce the CIV due to systematic rater error, but will not entirely eliminate it, as in the case of a student who luckily draws 2 or more lenient raters or is unlucky in being rated by 2 or more harsh raters.

Passing score determination may be more difficult for observational clinical performance examinations, but is an essential component of the assessment and a potential source of CIV error. The method and procedures used to establish defensible, reproducible and fair passing scores or grades for clinical performance examinations are as important as for other assessment methods and similar procedures may be used.[10,22]

What about face validity?

The term *face validity*, despite its popularity in some medical educators' usage and vocabulary, has been derided by educational measurement professionals since at least the 1940s. *Face validity* can have many different meanings. The most pernicious meaning, according to Mosier, is: '…the validity of the test is best determined by using common sense in discovering that the test measures component abilities which exist both in the test situation and on the job.'[23](p 194) Clearly, this meaning of *face validity* has no place in the literature or vocabulary of medical educators. Thus, reliance on this type of face validity as a major source of validity evidence for assessments is a major threat to validity.

Face validity, in the meaning above, is not endorsed by any contemporary educational measurement researchers.[24] Face validity is not a legitimate source of validity evidence and can never substitute for any of the many evidentiary sources of validity.[2]

However, as the term *face validity* is sometimes used in medical education, can it have any legitimate meaning? If by *face validity* one means that the assessment has superficial qualities that make it appear to measure the intended construct (e.g. the SP case looks like it assesses history taking skills), this may represent an essential characteristic of the assessment, but it is not validity. This SP characteristic has to do with acceptance of the assessment by students and faculty or is important for administrators and even the public, but it is not validity. (The avoidance of this type of *face invalidity* was endorsed by Messick.[3]) The appearance of validity is not validity; appearance is not scientific evidence, derived from hypothesis and theory, supported or unsupported, more or less, by empirical data and formed into logical arguments.

Alternative terms for *face validity* might be considered. For example, if an objective test looks like it measures the achievement construct of interest, one might consider this some type of value-added and important (even essential) trait of the assessment that is required for the overall success of the assessment programme, its acceptance and its utility, but this clearly is not sufficient scientific evidence of validity. The appearance of validity may be necessary, but it is not sufficient evidence of validity. The congruence between the superficial look and feel of the assessment and solid validity evidence might be referred to as *congruent* or *sociopolitical meaningfulness*, but it is clearly not a primary type of validity evidence and can not, in any way, substitute for any of the 5 suggested primary sources of validity evidence.[2]

CONCLUSIONS

This paper has summarised 2 common, general threats to validity in the context of the contemporary meaning of validity – a unitary concept with multiple facets, which considers construct validity as the whole of validity. Validity evidence refers to the data and information documented in order to assign meaningful interpretations to assessment scores or outcomes. Validity always refers to the meaningfulness of an interpretation of a test score or a rating and never to the assessment itself.

Construct under-representation threats relate primarily to undersampling or biased sampling of the content domain or the selection or creation of assessment items or performance prompts that do not match the construct definition. Construct-irrelevant variance threats introduce systematic, rather than random, measurement error, and reduce the ability to interpret assessment outcomes in the proposed manner. *Face validity* is rejected as any legitimate source of validity evidence and reliance on face validity as an important source of validity evidence is suggested to be a threat to validity.

CONTRIBUTORS

SMD is an Associate Professor in the Department of Medical Education, College of Medicine, at the University of Illinois at Chicago. TMH is a Professor of Educational Psychology at the University of Arizona West in Phoenix.

FUNDING

There was no external funding for this project.

REFERENCES

1 Downing SM. Validity: On the meaningful interpretation of assessment data. *Med Educ* 2003;**37**:1–8.

2 American Educational Research Association, American Psychological Association, National Council on Measurement in Education. *Standards for Educational and*

Psychological Testing. Washington, DC: American Educational Research Association 1999.

3 Messick S. Validity. In: Linn RL, ed. *Educational Measurement*. 3rd edn. New York: American Council on Education, Macmillan 1989;13–104.

4 Downing SM. Threats to the validity of locally developed multiple-choice tests in medical education: construct-irrelevant variance and construct underrepresentation. *Adv Health Sci Educ* 2002;**7**:235–41.

5 Downing SM. Construct-irrelevant variance and flawed test questions: do multiple-choice item writing principles make any difference? *Acad Med* 2002;**77** (10):103–4.

6 Holland PW, Wainer H, eds. *Differential Item Functioning*. Mahwah, New Jersey: Lawrence Erlbaum 1993.

7 Haladyna TM, Downing SM. Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues Pract* 2003; (in press).

8 Haladyna TM, Downing SM. How many options is enough for a multiple-choice test item? *Educ Psychol Measurement* 1993;**53**:999–1010.

9 Downing SM. Guessing on selected-response examinations. *Med Educ* 2003;**37**:1–2.

10 Norcini JJ. Setting standards on educational tests. *Med Educ* 2003;**37**:464–9.

11 Case SM, Swanson DB. *Constructing Written Test Questions for the Basic and Clinical Sciences*. 2nd edn. Philadelphia: National Board of Medical Examiners 1998.

12 Haladyna TM, Downing SM, Rodriguez MC. A review of multiple-choice item-writing guidelines. *Appl Measurement Educ* 2002;**15** (3):309–33.

13 Miller GE. The assessment of clinical skills/competence/performance. *Acad Med (Suppl)* 1990;**65**:63–7.

14 Elstein AS, Shulman LS, Sprafka SA. *Medical Problem Solving: An Analysis of Clinical Reasoning*. Cambridge, Massachusetts: Harvard University Press 1978.

15 Norman GR, Tugwell P, Feightner JW, Muzzin LJ, Jacoby LL. Knowledge and problem solving. *Med Educ* 1985;**19**:344–56.

16 van der Vleuten CPM, Swanson DB. Assessment of clinical skills with standardised patients: state of the art. *Teach Learn Med* 1990;**2**:58–76.

17 Brennan RL. *Generalizability Theory*. New York: Springer Verlag 2001.

18 Crossley J, Davies H, Humphris G, Jolly B. Generalisability: a key to unlock professional assessment. *Med Educ* 2002;**36**:972–8.

19 De Champlain AF, Floreck LM. Assessing potential bias in a large scale standardised patient examination: an application of common DIF methods for polytomous items. [Paper prepared for the 2002 Annual Meeting of the American Educational Research Association, New Orleans.] 2002.

20 Williams RG, Klamen DA, McGaghie WC. Cognitive, social and environmental sources of bias in clinical performance ratings. *Teach Learn Med* 2003;**15**: 270–92.

21 Norcini J, Guille R. Combining tests and setting standards. In: Norman GR, van der Vleuten CPM, Newble DI, eds. *International Handbook of Research in Medical Education*. Dordrecht: Kluwer Academic Publishers 2002;811–34.

22 Norcini JJ, Shea JA. The credibility and comparability of standards. *Appl Measurement Educ* 1997;**10**:39–59.

23 Mosier CI. A critical examination of the concepts of face validity. *Educ Psychol Measurement* 1947;**7**:191–205.

24 Downing SM. Test validity evidence: what about face validity? CLEAR Exam Review 1996;31–3.

Received 6 June 2003; editorial comments to authors 21 July 2003; accepted for publication 20 August 2003

Revisiting 'Assessment of clinical competence using an objective structured clinical examination (OSCE)'

Ronald M Harden

Editor's note: As part of our 50th volume celebrations, Medical Education is looking back at its most impactful articles, as defined by citation count. The most cited articles from each 5-year interval were identified and the original authors of one of them (or other knowledgeable scholars if the original authors could not be found) were asked to comment on the state of the field at the time of publication, the impact of the article, and what we have learned since then. The article illustrated in Figure 1 was one of the most cited articles in our journal in the 1977–1981 period. To see the other top-cited articles from Volumes 1–50, please view the interactive PDF by visiting www.mededuc.com.

The Association for the Study of Medical Education (ASME) publication 'Assessment of clinical competence using an objective structured clinical examination (OSCE)'[1] provided the first complete description of the use of the OSCE to assess a student's clinical competence (Fig. 1). This current paper describes the background to the introduction of the OSCE and how it became the reference standard for performance assessment.

In the late 1960s and early 1970s, as a senior lecturer in medicine in Glasgow and later in Dundee, I was responsible for student assessment. Three things struck me. The first was that the assessment of a student's clinical skills was regarded as important and a student could not graduate without passing the clinical examination. Secondly, there were major deficiencies in the clinical examination represented by the impact of the luck of the draw on both the type of patient seen by the student in the long case and the two examiners assigned to assess the student's competence. In the traditional clinical examination, the marks awarded by one examiner often varied considerably from those awarded by another examiner observing the same performance.[2] John Stokes, an experienced examiner, described the clinical examination as the 'sacred cow of British medicine' and as a 'half-hour disaster session'.[3] Thirdly, it seemed to me that it should be possible to construct an examination that reliably assessed the range of competencies expected of the student, in which what was to be assessed at each station would be defined clearly in advance and reflected in a checklist and rating scale to be completed by the examiner. I was aware of the work of Barrows and Abrahamson[4] and others on the use of simulated patients and felt that in some areas, such as the assessment of communication skills, a simulated patient could replace a real patient in the examination. In other situations, such as those concerning a patient with a hernia or goitre, the student should be assessed with a real patient.

In the traditional clinical examination, marks awarded by one examiner often varied considerably from those awarded by another observing the same performance

In Dundee I found a culture that encouraged innovation in medical education. I recall a conversation in the hospital car park with Alfred Cuschieri, who had been appointed professor of

Medical Education 2016: 50: 376–379,
doi: 10.1111/medu.12801

Dundee, UK

Correspondence: Ronald M Harden, Association for Medical Education in Europe, 12 Airlie Place, Dundee DD1 4HJ, UK. Tel: 00 44 1382 381953; E-mail: r.m.harden@dundee.ac.uk

Medical Education, 1979, **13**, 41–54

Assessment of clinical competence using an objective structured clinical examination (OSCE)

R. M. HARDEN AND F. A. GLEESON

Ninewells Hospital and Medical School, Dundee, Scotland

1. Introduction

Assessment of students is a matter of continuing concern for medical teachers. Numerous attempts have been made to improve the reliability and validity of written examinations, and recent ASME booklets have described multiple choice questions of the one from five type (Lennox, 1974) and the modified essay question (Knox, 1975). The clinical examination is regarded by many examiners as of key importance in the assessment of a student's competence to practice medicine and the cornerstone in qualifying examinations. While deficiencies in the conventional or traditional clinical examination have been clearly identified (Stokes, 1974; Wilson *et al.*, 1969), few attempts have been made to improve the assessment of a student's clinical skills. Indeed, in the U.S.A. the tendency has been to move away from examinations at the bedside and towards patient management problems (Hubbard, 1971; Newble, 1976).

This booklet describes a procedure—the objective structured clinical examination (OSCE)—designed to assess clinical competence at the bedside, and suggests that there are many advantages if this approach is incorporated into examinations aimed at testing clinical skills. Guidelines for the organization of such an examination are described. A preliminary report describing the OSCE has been published (Harden *et al.*, 1975).

2. Criteria for a procedure to assess clinical competence

The ideal examination should fulfil three criteria.

Correspondence: Professor Ronald M. Harden, Centre for Medical Education, Ninewells Hospital and Medical School, Dundee, Scotland.

(a) *Is it valid?* Does it measure what it is supposed to measure? Is there evidence for what the examiners think they have seen? Can the examiners generalize from what they have seen?
(b) *Is it reliable?* Is the examination an objective assessment? Are the results accurate and consistent? Would other assessors agree with the examiner's interpretation of the student's behaviour? As Rowntree (1977) notes: 'There is an assumption rampant in talk of academic standards, that all qualified assessors feel, understand and judge in much the same way when confronted with the work of a particular student. It is presumed that they would notice and value the same skills and qualities and would broadly agree in their assessments. Abundant evidence attests to the falsity of such assumptions'.
(c) *Is it practical?* Can the requirements for staff and accommodation be met? Can it cope with sufficient numbers of students?

Validity

A valid clinical examination should assess the components of clinical competence, including the ability to: (a) obtain from the patient a detailed and relevant history; (b) carry out a physical examination of the patient; (c) identify the patient's problems from the information obtained and reach a differential diagnosis; (d) identify the appropriate investigations; (e) interpret the results of the investigations; and (f) recommend and undertake appropriate management, including patient education.

Many of these abilities are to a greater or lesser extent ignored in standard procedures used to assess clinical competence. Frequently in the examination attention is paid to the detection of abnormal physical findings while no attempt is made to watch the student taking a history from a patient. 'We are all anxious', wrote Matthews

Figure 1 Title page from 'Assessment of clinical competence using an objective structured clinical examination (OSCE)'[1]

surgery. He made clear his determination to change the format of the final examination in surgery. Working with his two senior lecturers, Paul Priece and Robert Wood, and with Fergus Gleeson, who had come across from Ireland to work with me in the field of medical education, we planned to introduce the OSCE as the final examination in surgery at Dundee. The faculty board agreed that we could run a pilot final surgery OSCE alongside the traditional final surgical clinical examination. The result was a success and the following year the board agreed to replace the traditional surgery final clinical examination with an OSCE.[5] This was possible because, in the UK, final examinations are the responsibility not of a national body, but of each school independently.

It should be possible to construct an examination that reliably assessed the range of competencies expected of the student

This early development of the OSCE has been described in more detail.[6–9] A preliminary report that described the OSCE concept was published in the *British Medical Journal*[10] and a more complete description was published in *Medical Education* as an ASME medical education booklet.[1]

The OSCE became widely adopted as an examination tool with which to assess students' clinical competence. Teachers became aware of the approach through the published papers, and external examiners from other schools who participated in the Dundee OSCE spread the initiative to their schools, as did Dundee staff when they transferred to other schools. Ian Hart, a senior physician in Ottawa, Ontario, Canada, with whom I had previously collaborated in the area of thyroid research, spent a sabbatical in Dundee and rapidly became a convert to the OSCE. Together we organised the first Ottawa Conference in 1985, which aimed to share across the Atlantic approaches to the assessment of clinical competence, including the OSCE.

The OSCE is now used in countries around the world to assess clinical competence in a range of disciplines, in different health care professions and in the different phases of education. It has also been used outside medicine, for example in the police force.[8] More than 1600 papers on the OSCE have been published, including about 400 since 2011 (almost one new paper every 3 days!).

Why has the OSCE been widely adopted as the recommended approach for the assessment of clinical competence and become the reference standard for performance assessment?[11] Schneider[12] identified four characteristics that lead to the adoption of an innovation. The first characteristic is perceived significance. Ideas that are adopted, Schneider argues,[12] stand out not necessarily because they *are* but, rather, because they *seem* to be significant. The OSCE was perceived by teachers as addressing an important problem: the assessment of a learner's clinical competence. The second characteristic is philosophical compatibility: teachers must view the innovation as appropriate for use. Clinical teachers and examiners easily identified the OSCE with their own thinking. Schneider's[12] third characteristic refers to occupational realism: ideas must be practical and easy to put into immediate use. This is certainly true of the OSCE. The fourth characteristic is transportability: the approach must be easily explainable to a busy colleague and adaptable for use in different situations. The OSCE has proved to be user-friendly and can be adapted for use in different contexts.[8] Its characteristics of perceived significance, philosophical compatibility, occupational realism and transportability have facilitated the wide adoption of the OSCE as a tool to assess clinical competence.

The OSCE is now used in countries around the world to assess clinical competence in a range of disciplines

Reflecting on my experience with the OSCE over the last 35 years, I find I have learned eight lessons.

Firstly, as demonstrated with our initial implementation of the OSCE, obtaining agreement for a pilot test is a quick way of introducing a new assessment approach.

Secondly, having powerful champions is vital. The support of senior professors within the school of medicine was important and facilitated the introduction of the OSCE.

Thirdly, there are major advantages if a medical school has the freedom to innovate in assessment. The medical school in Dundee had the authority to design its own assessment procedures and was not dependent on the agreement of a national examination body.

Fourthly, flexibility and the ability to adapt a method to local contexts are key to the success of an innovation.

Fifthly, scarce resources and the presence of large numbers of students need not stand in the way of innovation. I have yet to see an example of a situation in which the cost or the number of students to be assessed prevents the adoption of the approach. The only limitation is the imagination of the developer.

My sixth lesson refers to the discovery that here are 'good' OSCEs and 'not so good' OSCEs. Reliability and validity are related to how the OSCE is implemented. The OSCE is really a POSCE (*potentially* objective structured clinical examination).

There are 'good' OSCEs and 'not so good' OSCEs. Reliability and validity are related to how the OSCE is implemented.

My seventh lesson reflects the knowledge that the clinical teacher is in a good position to advance medical education. I was a senior lecturer in medicine when I started the work on the OSCE. Although there is a place in medical education for large-scale research projects, more attention needs to be paid to the teacher as an action researcher.

Although there is a place in medical education for large-scale research projects, more attention needs to be paid to the teacher as an action researcher.

Finally, an approach to education will continue to evolve with time. Although the basic principles of the OSCE are as true today as when they were first described in the paper published in *Medical Education* in 1979,[1] we can see developments in the use of technology to support the OSCE, such as in the use of simulators and in new automated marking schemes using, for example, iPads. Further, more serious attention is now paid to standard setting and to the assessment of competencies, such as teamwork skills and error management, and to patient safety, none of which featured much on the agenda in 1979.

REFERENCES

1 Harden RM, Gleeson FA. Assessment of clinical competence using an objective structured clinical examination (OSCE). *Med Educ* 1979;**13**(1):39–54.
2 Wilson GM, Lever R, Harden RM, Robertson JIS, MacRitchie J. Examination of clinical examiners. *Lancet* 1969;**293**:37–40.
3 Stokes J. *The Clinical Examination – Assessment of Clinical Skills. Medical Education Booklet 2.* Dundee: Association for the Study of Medical Education 1974.
4 Barrows HS, Abrahamson S. The programmed patient: a technique for appraising student performance in clinical neurology. *J Med Educ* 1964;**39**:802–5.
5 Cuschieri A, Gleeson FA, Harden RM, Wood RA. A new approach to a final examination in surgery. *Ann R Coll Surg Engl* 1979;**61**:400–5.
6 Hodges B. OSCE! Variations on a theme by Harden. *Med Educ* 2003;**37**(12):1134–40.
7 Hodges B. *The Objective Structured Clinical Examination: A Socio-History.* Berlin: Lambert Academic Publishing 2009.
8 Harden RM, Lilley P, Patricio M. *The Definitive Guide to the OSCE: The Objective Structured Clinical Examination as a Performance Assessment.* Edinburgh: Elsevier 2015.
9 Centre for Medical Education Dundee. Interview with Professor Ronald Harden about the OSCE. *https://vimeo.com/67224904. [Accessed 30 December 2015.]*
10 Harden RM, Stevenson M, Downie WW, Wilson GM. Assessment of clinical competence using an objective structured clinical examination. *Br Med J* 1975;**1**:447–51.
11 Humphrey-Murto S, Touchie C, Smee S. Objective structured clinical examinations. In: Walsh K, ed. *Oxford Textbook of Medical Education.* Oxford: Oxford University Press 2013;524–536.
12 Schneider J. Closing the gap… between university and schoolhouse. *Phi Delta Kappan* 2014;**96**:30–5.

Standardized Patient Script

Title:

Patient's Detail		
SP Name	**Past Medical History**	
Age/Race/Gender	**Medications**	**Allergies**
Presenting Complaints (Verbatim)		

Presentation
Affect
Appearance

Prop List
Clothes

Setup
Location of Interview
Physical Characteristics

Past History

History of Presenting Illness

Past medical history:

Social history:

Past surgical history:

Hospitalizations:

Allergies:

Family History

Ibu dan Bapa

Adik-beradik

Keluarga

Check-list for the SP:

	Question	Answer
1		
2		
3		
4		
5		
6		
7		
8		
9		

OBJECTIVE STRUCTURED CLINICAL EXAMINATIONS
UNIVERSITI ……
EXAMINATION……
SESI AKADEMIK ……

STATION INFORMATION

1) Station code name : P3

2) Duration : 10 minutes

3) Station requirements : Patient

4) Date of vetting at department : 27-7-2021

5) Date of vetting at faculty : 29-7-2021

6) Author :

7) Corrected author :

INSTRUCTIONS TO CANDIDATE:

Scenario:

Task:

1. .
2. :

Duration:

You have ? minutes to complete the task.

- You are expected to complete your history taking by ? minutes
-

INSTRUCTIONS EXAMINER:

Objectives of the test:

1)
2) **..**
3)

Task:

1. Observe the candidate interviewing the patient. Allocate ? minutes for the candidate to complete this exercise. **Do not interrupt** or prompt the candidate during this examination.
2. If the candidate has not completed the examination in this time, you may interrupt and proceed with the discussion.
3.

INSTRUCTIONS TO PATIENT

Scenario:

You will be interviewed by the student for the purpose of reaching a diagnosis.

1. Do not volunteer information unless asked.
2. Do not disclose the name of diagnosis/medication.
3. Do not guide the students but be cooperative and assist them accordingly.
4. Each student will take ? minutes for each interview.

Arahan kepada pesakit:

Senario:

Anda akan ditemu-bual oleh pelajar dengan tujuan mencapai diagnosis.

1. Jangan beri maklumat yang tidak ditanya.
2. Jangan beritahu nama penyakit/ubat.
3. Jangan beri panduan kepada pelajar tetapi bekerjasama dan membantu mereka dengan sewajarnya.
4. Setiap pelajar akan mengambil masa ? minit untuk temubual.

II. Global Rating on candidate's overall performance *(Please Circle)*

Do you have concerns regarding this candidate's ethical and/or professional behavior?

□ Yes (please specify) □ No

_____

Examiner's name

Signature :

Date :