

# **Assessment of competence of trainees in psychiatry**

**Created by Brian Hodges, MD, MEd, FRCPC  
Associate Professor and Vice-Chair (Education)  
Department of Psychiatry and the Wilson Centre for Research in Education  
University of Toronto**

## **Quick Index**

### **Overview**

**A very brief history of assessment in psychiatry**

### **Blueprinting**

### **Three dimensional evaluation:**

#### **Dimension 1: Objectives, context and philosophy**

**Longitudinal vs. cross-sectional**

**Formative vs. summative**

**Normative vs. criterion-based**

**Holistic vs. atomistic**

#### **Dimension 2: Psychometric considerations**

**Reliability**

**Validity**

**Feasibility**

#### **Dimension 3: Assessment Methods, Tools and Instruments**

##### **Written Tests**

**Multiple choice questions**

**Short answer question/short essay questions**

**Essays**

##### **Longitudinal Assessments**

**Ward ratings**

**Interprofessional Assessments**

**Peer Assessments**

**Self Assessments**

##### **Performance-based Assessments**

**Objective Structured Clinical Examinations**

**Oral Examinations**

## Overview

The assessment of clinical competence in psychiatry is a priority for educators at undergraduate, postgraduate and certification levels. Recently, there have been tremendous changes in the technology of performance-based assessment, particularly with the advent of methods such as the objective structured clinical exam (OSCE) and other standardised patient-based assessment methods. And while the importance of assessment is widely recognised, selecting appropriate evaluation instruments can be confusing. While few teachers have the time to become experts in the burgeoning field of assessment, knowledge of some basic principles of assessment can make the whole process much more rewarding for teachers and satisfying for students.

A rich and diverse literature is available to assist educators in planning, maintaining and improving evaluation methods in psychiatry. This web site is not intended to reproduce that literature, but rather to provide a starting place for educators interested in assessment. The following information is intended to illustrate the terrain in assessment and is organised using a three-dimensional approach. Educators designing assessments will find it helpful to consider each dimension in the planning and development process. The first dimension concerns important philosophical issues related to student assessment. The second dimension addresses the psychometric qualities of measures and methods chosen and the third dimension consists of specific formats of assessment. The first two sections are 'filters' through which all of the methods can be viewed. By considering general principles first, it will be easier to select those methods that reflect the desired goals of an assessment program. Each section includes references that provide more information to assist in developing an assessment program.

## A very brief history of assessment in psychiatry

An evolution of assessment practices over the last century has affected all health professions. These changes are summarized in Table 1. Following the World Wars, efforts to reliably categorize individuals led to the widespread adoption of multiple choice examinations in many fields, including psychiatry. At mid-century, however, questions about the relationship of knowledge to actual clinical practice led to a new or in some cases renewed interest in clinical oral examinations. Using an oral exam format students were observed interviewing a patient and then asked questions to test their synthesis of knowledge in a clinical context. Observation of performance was thought to more validly reflect actual skills in practice. Toward the latter part of the 20<sup>th</sup> century, research on assessment raised concerns about the reliability of clinical orals and illustrated that the use of only one case led to wild variations in student scores that were not always reflective of the abilities observed over long periods of time. These doubts, combined with the emerging technology of standardised patients, led to a rise in the use of objective structured clinical examinations (OSCE) and other formats that allowed multiple observations of clinical scenarios. Multi-station simulations also allowed observation of skills in circumstances that rarely present with ‘real’ patients. Now as we enter the new millennium, assessment researchers have begun to explore more advanced issues of validity, questioning how examinations affect student learning and development in areas such as expert problem solving, cultural competence and self-directed learning.

<b>Era</b>	<b>Dominant conceptualization of competence to be assessed</b>	<b>Measure</b>
19 <sup>th</sup> Century	The good citizen	Personal assessment by apprenticeship supervisor
Early 20 <sup>th</sup> Century	Extensive knowledge base	Multiple choice examinations(MCQs), anatomical and pathological identification
Mid 20 <sup>th</sup> Century	Ability to synthesize knowledge in context, make diagnoses and treatment plans	In addition to MCQs, written ‘patient management problems’ (PMPs) and oral examinations
Late 20 <sup>th</sup> Century	Performances of physical examination and communication skills	Instead of oral examinations or PMPs, standardized patient-based OSCEs
21 <sup>st</sup> Century	Broader concepts of competence including cultural competence, interprofessional skills, legal and ethical considerations	Multi-dimensional simulations including other health professionals, cultural and other communication challenges

## Blueprinting

Before selecting evaluation tools, it is necessary to create an evaluation “blueprint”. The blueprint is a matrix that outlines parameters for the assessment in three major areas: domains to be assessed (knowledge, skills or attitudes), content (mood disorders, psychotic disorders, psychogeriatrics, etc.) and assessment instruments (written test, OSCE, ward rating). The blueprint matrix ensures that there is balance across the assessment formats and between different administrations of similar exams. In order to achieve the blueprint objectives, multiple assessment methods are usually required. Selection of different assessment instruments allows one to balance the strengths and weaknesses of each method. Strengths and weaknesses of assessment instruments include the psychometric properties, the acceptability to faculty members and students, and finally the feasibility. For example, the use of global ward ratings that have very low reliability, high face validity and low cost, might be balanced by the use of an OSCE with much higher reliability but higher cost. The sections below address three dimensions of a complete assessment scheme: objectives, psychometrics and assessment tools.

## Three dimensional evaluation

### Dimension 1: Objectives, context and philosophy

Prior to creating an assessment program and choosing actual assessment tools, educators must address several fundamental philosophical issues. What is the purpose of the assessment? How will the results be used? What effects will the assessment have on student learning? Four prominent dichotomies listed below should be considered carefully.

### Longitudinal vs. cross-sectional

Is the assessment a ‘one-shot’, cross-sectional biopsy of knowledge, skills and attitudes, or is will it take place over an extended period of time? Examples of cross-sectional assessments include a pre-course multiple-choice assessment of knowledge, an end of course Objective Structured Clinical Examination, or a national licensure examination. Examples of longitudinal assessments include ratings of interprofessional skills over a six-week clerkship, weekly logs of patients seen in the emergency department and final in-training evaluations by a clerkship or residency supervisor.

	longitudinal	cross-sectional
strengths	<input type="checkbox"/> multiple points of observation can lead to greater reliability	<input type="checkbox"/> easier to define specific objectives and structure the test content
limitations	<input type="checkbox"/> development of a relationship between student and assessor can bias rating, particularly	<input type="checkbox"/> more intrusive and artificial if removed from clinical context

	where there are problems <input type="checkbox"/> busy teachers may not observe very much, and give ratings of ambiguous qualities such as personality	
--	---	--

### Formative vs. summative

Summative assessments are conducted for the purpose of coming to summary conclusions about a student's competence and result in assignment of a mark or grade. Course-end exams or year-end comprehensives are common examples. Formative assessments by contrast are given in order to change, improve or shape the knowledge, skills or attitudes of a student through feedback about their performance. While an examination can have both functions, general any summative examination that occurs at the end of a clinical experience has little function in further learning.

	formative	summative
strengths	<input type="checkbox"/> can guide student to further learning	<input type="checkbox"/> can provide a clear, diagnostic statement about a student's competence
limitations	<input type="checkbox"/> may not be taken as seriously by non-self directed learners	<input type="checkbox"/> usually occurs too late to have any meaningful impact on learning

### Normative vs. criterion-based

Simply put, test makers must decide if they wish to use the results of their assessment to compare students to each other or to predetermined standards for performance. Normative systems use numerical systems to classify performance (e.g. bottom 10% fail) where as criterion based systems tie success to actual outcomes. While in principle criterion-based testing is preferred by many, it is rarely operationalised properly. Setting a random pass mark, such as 60% is not criterion-based testing. A true criterion-based test describes exactly what is expected of the competent student on every element. For example, 'the competent student will ask about suicidal ideation in a depressed patient' or 'the competence student will ask about at least 10 or 12 symptoms of depression', etc.

	normative	criterion based
strengths	<input type="checkbox"/> prevents fluctuations in the pass rate	<input type="checkbox"/> test criteria can easily related to learning objectives
limitations	<input type="checkbox"/> homogenizes strong or	<input type="checkbox"/> very hard to define the

	weak cohorts of students <input type="checkbox"/> does not emphasize specific competencies	criteria <input type="checkbox"/> may result in wildly variable pass rates
--	---	---

### Holistic vs. atomistic

Scoring systems used in assessment can be categorised on a continuum ranging from the most holistic (or “global”) to the most analytic. Hunter and colleagues defined 5 broad categories along this continuum ranging from most to least holistic: general impression scoring> holistic scoring>primary trait scoring>analytic scoring>atomistic scoring. Adapting this classification to assessment in psychiatry, OSCE checklists are “atomistic” because they involve counting elements of performance. Primary trait scoring breaks into component parts, the performance or construct of interest by asking raters to provide a score on a number of specific attributes or traits. Raters are guided by a unique scoring rubric for each trait, but these rubrics may be applied to a number of different performances (i.e. the same anchored ratings are used in multiple OSCE stations). Holistic scoring, which utilises as system of descriptors and numerical anchors to increase rigor and consistency is used in OSCEs, written exams with model answers, etc. General impression scoring includes intraining evaluations that involve making an overall pass- fail rating or assigning a grade without reference to any defined behavioural anchors.

	holistic spectrum	atomistic spectrum
strengths	<input type="checkbox"/> more integrated <input type="checkbox"/> probably better captures the characteristics of expertise and other global qualities	<input type="checkbox"/> more likely to be reproducible across time and across raters <input type="checkbox"/> easily converted to student feedback
limitations	<input type="checkbox"/> can be highly subjective <input type="checkbox"/> hard to explain to students <input type="checkbox"/> not useful for specific feedback	<input type="checkbox"/> emphasizes minutiae of human behaviour and can lose broader elements of the “art” of practice and expertise

## **Dimension 2: Psychometric considerations**

Once you have grappled with the preceding four assessment principles and located your self and your group on each of these dimensions, you can turn your attention to the psychometric properties of various assessment tools. The following sections examine important properties of the test instruments you will select and balance.

### **Reliability**

Reliability (precision) is the degree to which an instrument yields the same score on different occasions or with different observers. For example, when 2 teachers watch a clerk interview a patient and independently assign very different marks the test can be said to have low inter-rater reliability. Much has been written about the reliability of various forms of assessment. In general, reliability will be greater as the number of observations increases. This can be accomplished by using multiple situations/questions or multiple raters. The number of test items required to achieve sufficient reliability varies with the nature of the test. While acceptable reliability may require over 150 multiple choice questions, perhaps 10–15 OSCE stations will produce a reliable performance-based assessment. Reliability cannot be predetermined and should be assessed following each examination.

### **Validity**

Validity is the degree to which a test measures what it is intended to. If for example, at the end of a psychiatry rotation, a clerk who was very punctual, friendly, out-going and motivated is given high marks in the category "interviewing skills" (when he or she had not been observed performing interviews) the rating could be said to have low validity.

There are several types of validity. Construct validity is the degree to which a test can be shown to be measuring a coherent theoretical domain. Content validity is the degree to which a test has good coverage of a domain. Predictive validity is the degree to which results on a test predict results on a future test or other future event. Concurrent validity concerns the degree to which test results correlate with another "gold standard" known to measure the same domain. And finally, face validity is the degree to which a test appears to measure what it is supposed to.

Validity is harder to assess than reliability and each of the above forms is examined in a different way. Educators should ask themselves, how are we going to know if the assessment is testing what we think it is? A means of answering this question might involve comparing test scores across different types of assessments, comparing test results with actual performance, or comparing the ability of students at higher levels of competence to obtain higher scores on the same test. If professionals with higher levels of expertise do not obtain higher scores on a test then the validity of that test is in question.

It could be that the measures are not sensitive to increasing levels of expertise or may indicate that other constructs - such as "test taking ability" or "sunny personality" are

confounding measurement of the desired construct. We have published an example of a validity study that may assist exam teams interested in pursuing this aspect of OSCE development. (See references below)

### **Feasibility**

A crucial, but often neglected aspect of planning evaluations is consideration of feasibility of each of the methods chosen. How much is the test going to cost? How many hours of faculty time will be involved? Are support staff needed? Are standardized patients involved? All of these can have a major impact whether a given method is feasible or not. While the cost of evaluation methods is beyond the scope of this discussion, a rough estimate of the relative cost is shown in the table comparing methods below.

Issues of reliability, validity and feasibility must be balanced. Sometimes they are in competition. For example, the number of OSCE stations required to achieve acceptable reliability might not be economically feasible. At this stage, you may wish to create an overall assessment plan that lists the methods you are considering, along with their psychometric properties and costs. Below is a sample assessment plan.

### **Assessment Program for a Department of Psychiatry 6-week clerkship.**

Evaluation Method	Primary domain for which test has validity	Reliability	Cost (time and people)
multiple choice test: 150 items	knowledge	high	low
2 case reports marked by same examiner, anchored rating scale	knowledge, skills	low - moderate	low
2 observed interviews, by different faculty members, global rating	skills	low-moderate	moderate
10 station OSCE	skills	mod - high	high
global evaluation ward-rating	Knowledge, skills and attitudes	very low - low	low

In general common problems to avoid are:

- ❑ Using written exams to assess any domain other than knowledge. They are not valid in assessing skills and attitudes
- ❑ Relying heavily on in-training rating scales to assess knowledge and skills. They are highly unreliable and generally measure personality variables rather than skills
- ❑ Using an oral exam with one patient or one examiner. A single observation has low reliability based on use of only clinical domain. A test which is not reliable cannot be valid

### **Dimension 3: Assessment Methods, Tools and Instruments**

#### **Index:**

#### **Written Tests**

There are many formats of written tests ranging from the highly structured (multiple choice tests) to the highly open-ended (essays). While written tests can theoretically assess a variety of domains, they are generally best used to assess knowledge. The construction of good written tests is difficult. Below is a brief review of some of the most common forms of written test.

#### **Multiple choice tests (MCQs)**

The reliability of MCQs is usually very high due to multiple observations. The problem is writing good questions. A helpful guide is produced by the National Board of Medical Examiners in the United States (see references below). It is a challenge to produce questions that test more than factual recall. Security is also an issue. Generally where question banking is used, it is necessary to have several thousand items before any are reused. Also, the proportion of test items that are recycled should be controlled, and not exceed a third on each subsequent test iteration. Multiple choice tests can sample a wide range of content, but they tend to be predictive only of other written tests. They are general not very predictive of clinical performance, and in some cases are negatively correlated.

#### **Short answer questions (SAQs)**

Short answer questions utilise a short ‘stem’ that introduces a problem or clinical case, followed by a question or series of questions requiring answers of a few words or sentences. The brevity of questions means that many can be used in the same test – often 8-10 per hour. The brevity of responses means that SAQs can be scored with more consistency than formats such as essays, which require extensive interpretation. To test basic factual recall, MCQs are probably more efficient.

#### **Essays**

Essays involve multidimensional questions and demand complex, narrative responses. Scoring essays is difficult because of the degree of interpretation that is required to compare the content and prose of different student responses. Essays do tap into students’ abilities to create coherent arguments and to write clearly, two skills that are often greatly neglected among medical trainees. However, because of their length, essays generate many fewer scores or ‘data-points’ than do MCQs or SAQs. Fewer judgments result in lower reliability. All of these factors have discouraged the use of essays in high stakes examinations where MCQs and SAQs continue to be the standard.

## Longitudinal Assessments

### Global In-Training Ratings

In-training assessments are notoriously unreliable but universally employed and may be unavoidable. Nevertheless, it is important to be cognizant of the strengths and limitations of using global in-training ratings as a major part of an assessment program. A typical In-Training Assessment might look like this:

**Student:** Eric Burton      **Grade:** A+  
**Evaluation Narrative:** Eric is a very reliable, responsible and organized clerk. His knowledge base is excellent and his written notes are outstanding. He gets along very well with the multi-disciplinary team.

However, we often find out later that although Eric is a very poor interviewer, his supervisor doesn't know this because he/she never actually observed Eric's performance. Researchers have tried to determine what domains of competence in-training ratings assess. Unfortunately they often capture a subjective personal impression that is predictive of very little. Means of improving the reliability of in-training ratings have been described. In general, the most helpful approach is to use measures that are as behaviourally descriptive as possible. As well, evaluators should insist on narrative comments. It has been shown that these often contain really useful information even when an assigned grade does not.

The use of global ratings for more structured assessments such as standardized patient interviews is a different matter. Several authors have advanced arguments for an increased use of global ratings over behavioural checklists for assessments in which there is an observation of student performance. At least within structured assessments such as OSCEs, global ratings appear to have psychometric properties (including inter-station reliability or internal consistency, concurrent validity and construct validity) that are at least as good as, and often better than, checklists. (See references below) Second, there is a growing literature that suggests that clinicians with higher levels of expertise do not solve problems in clinical settings using checklists. Thus, the use of binary (yes/no) checklists tends to neglect higher components of clinical competence such as empathy and the organisation of knowledge in favour of linear accumulation of facts. Thus there are several reasons to utilise some form of global rating in addition to binary content checklists for structured performance-based assessments.

### Interprofessional assessment

While interprofessional communication skills are now deemed to be essential in the practice of contemporary health care, health professional learners receive little formal training in this area. Despite the fact that interprofessional communication skills are increasingly considered to be a core competence for physicians, there are few valid

measures to assess these skills. While the assessment of interprofessional skills is a new and growing area, it is not clear if this is best done in actual practice settings (with global performance evaluations) or in simulated situations (such as OSCEs). Much research and development needs to be done in this area.

### **Peer Assessments**

Peers assessment is appealing in theory, but difficult in practice. While very valuable information about performance can be obtained from peers, and indeed it can be argued that no one knows a student's strengths and weaknesses better than his or her peers do, the actual use of peer assessment must be approached cautiously. First, it is necessary to establish an environment of trust. Good peer assessment depends on honest and helpful information, but this will only be forthcoming if students feel that they can provide comments that will be handled with respect. Peer assessment can be done face-to-face, as occurs in problem-based learning groups, or anonymously. Peer assessment is more successful if multiple comments can be gathered and synthesized and is more effective if students receive some training regarding how to provide useful, behaviourally anchored feedback.

### **Self-Assessments**

Although self-assessment is essential for competent clinical practice and self-regulation, it is rarely incorporated into health professional training. Nevertheless, emphasising self-assessment skills can help students develop the ability for self-reflection that is essential for continuing education. While there are very few examples in the literature of self-assessments used for summative purposes, there are several papers that provide helpful information for use of self-assessment formatively. Below is a brief synthesis of this literature.

1. While an end of rotation exam might provide interesting data, a self-assessment that will influence further learning should occur early in a clinical rotation.
2. Measures to be used must have adequate reliability and validity both for the expert AND the student. This requires orientation to scales and practice using them prior to self-evaluation.
3. The criteria on scales must be specific. For example, on a 5 point Likert scale measuring empathy, the student must understand what each point means. Anchoring the scale with defined behavioural criteria helps. Perhaps more useful is a 'relative ranking' system in which students rank their own strengths and weaknesses, rather than comparing themselves to peers. Relative ranking is a complex but effective method of self-assessment (References are provided below).
4. It is important the students have an awareness of what the performance standard is. If they are to compare themselves to peers they must have *seen* other peers perform before they can use this standard. Too often, students are not exposed to the standard that they are compared to.

## **Performance-Based Assessments**

### **Objective Structured Clinical Examinations (OSCEs)**

The Objective Structure Clinical Exam (OSCE) was first described by Dr. Ronald Harden in the 1970s (see references below). The OSCE is a timed examination in which students move from one station to the next and demonstrate some combination of history taking, physical examination, counselling, or other aspect of patient management. At each station, candidates' performances are rated on checklists and global rating scales. As a new evaluation tool that allowed examiners to observe students performing in many different clinical situations, the OSCE was a major improvement over oral examinations in which only one clinical encounter was observed. As well the OSCE incorporated the technology of standardized patients first described by Barrows and Abrahamson in 1964. The use of standardized patients allowed the nature of problems and the level of difficulty to be standardized for all students. This combination of multiple observations and standardization of content and difficulty made the OSCE a very popular evaluation tool. Further, extensive research demonstrated that OSCEs could have excellent psychometric properties. As a result, OSCEs are now extensively used in medical schools throughout the world for the assessment of medical students, clinical clerks and residents and for licensure and certification. As well, OSCEs are used extensively for the assessment of the competence of other health professionals including chiropractors, nurses, nurse practitioners, pharmacists and physiotherapists.

While OSCEs have been used in many medical disciplines since the 1970s, psychiatric educators were initially slow to adopt this method of evaluation. Since 1990 however, psychiatry OSCEs have been gaining popularity. In the reference section below, a series of papers are listed that examine the development and psychometric properties of psychiatry OSCEs including reliability, and validity, and most recently a manual for creating and improving psychiatry OSCEs.

### **Oral Examinations**

An oral examination is usually an unstructured interaction between a student and one or two examiners. Often oral exams follow an observed or unobserved clinical interview. At one time they were promoted for evaluating "depth, breadth, problem solving and capacity under stress". However, a large body of literature criticized oral exams beginning in the 1960s. Problems with oral examinations largely stem from their unreliability. Examiner variation such as the presence of so called "hawks and doves" and the "halo effect" trouble oral exams. The National Board of Medical Examiners in the United States studied their oral exams for 10,000 US medical students over 3 years and found a mean correlation between 2 examiners in single patient encounters of less than 0.25. Thus oral examinations were discontinued for licensure in the US in 1963.

The oral examination has a validity problem by extension. A test that is not reliable cannot be valid. Several studies have shown that orals do not test knowledge synthesis well, and that commonly fragmented recall of isolated facts is tested. Maguire found that

examiners asked an average of one question every 40 seconds, of which 90% were recall thereby making the oral exam a verbally administered knowledge test. In terms of expense and efficiency, a method such as a paper and pencil test would be much more efficient and practical for such a purpose.

There are advantages of oral examinations. They can, for example, capture hard to measure domains such as of interpersonal skills, “thinking on one’s feet” and responding to unexpected situations and some authors have given emphasis to their role as a ritual or rite of passage. However, in psychiatry specifically, researcher have decried the “luck of the draw” element that plays a significant role in outcome oral exams, particularly for borderline candidates. Leichner suggested that “since it is difficult to improve the reliability of individual raters due to time and cost effectiveness issues, increasing the number of different tasks and evaluators may be a feasible means of improving the validity and reliability of oral examinations.”

If oral examinations are to be used, attention should be given to the following:

1. Use oral examinations to measure candidate characteristics not assessable by other means.
2. Ensure the oral examination is fairly and uniformly applied to all candidates. The content and process should be defined and relevant, and the results should be reproducible.
3. Be aware of common sources of error: hawk/dove, halo, central tendency, contrast error
4. Provide specific training for examiners in questioning techniques, role-play and videotaping
5. Ensure that individual oral examiners’ assessments of candidates are recorded prior to any discussion with others.
6. Avoid fine discrimination such as grades on a 100 point scales, which are likely impossibly unreliable. Use global categories (e.g. pass/condition/fail )
7. Monitor examiner performance using experienced examiners, one-way mirrors, videos, roving observers, etc.

## References

### General

Barrows HS and Tamblyn RM. (1980) *Evaluation of problem-based learning and clinical reasoning*. Springer, New York pp. 110-155

Brown G, Bull J and Pendlebury M. (1997) *Assessing student learning in higher education*. Routledge, London.

Epstein R and Hundert E. (2002) Defining and assessing professional competence. *JAMA* 287(2): 226-244.

Neufeld VR and Norman GR. (1985) *Assessing clinical competence*. Springer, New York.

The Royal College of Physicians and Surgeons of Canada. (1993) *Report on the evaluation system for specialist certification*, RCPSC, Ottawa.

Rowntree D. (1977) *Assessing students: How shall we know them?* Kogan Page, London.

Shore JS and Scheiber SC. (1994) *Certification, recertification and lifetime learning in psychiatry*. American Psychiatric Press, Washington.

### Multiple-choice examinations

Case S and Swanson DB. (1998) Constructing written test questions for the basic and clinical sciences. National Board of Medical Examiners, Philadelphia

### Standardized Patients

Barrows HS. *Simulated patients (programmed patients): development and use of a new technique in medical education*. Thomas, Springfield, IL, 1971.

Colliver JA, Swartz MH, Robbs RS, Cohen DS. (1999) Relationship Between Clinical Competence and Interpersonal and Communication Skills in Standardized-Patient Assessment. *Acad Med* 74:271-274.

Hanson M, Tiberius R, Hodges B, McKay S, McNaughton N, Dickens, S and Regehr G. (2002) Adolescent standardized patients: Methods of selection and assessment of benefits and risks. *Teaching and Learning in Medicine* 14(2), 104-113.

King A, Perkowski-Roger L and Pohl S. Planning standardized patient programs: Case development, patient training and costs. *Teaching and Learning in Medicine* 1994;6(1)

McNaughton N, Tiberius R, Hodges B. (1999) Effects of portraying psychologically and emotionally complex standardized patient roles. *Teaching and Learning in Medicine* 11(3), 135-141

Regehr G, Freeman R, Robb A, Missiha N, Heisey R. (1999) OSCE performance evaluations made by standardized patients: comparing checklist and global rating scores. *Acad Med* 1999 Oct;74(10 Suppl):S135-7

Stillman P, Ruggill JS, Rutala PJ and Saber DL. Patient instructors as teachers and evaluators. *Journal of Medical Education* 1980;55;186.

van der Vleuten CPM and Swanson D (1990) Assessment of clinical skills with standardized patients: State of the art. *Teaching and Learning in Medicine*, 2(2), 58-76

### **OSCEs**

Cohen R et al (1989) A comprehensive assessment of graduates of foreign medical schools. *Annals RCPSC*; 21:505-509

Cox K. No Oscar for OSCE. (1990) *Medical Education* 1990; 24, 540-545.

Cusimano MD, Cohen R, Tucker W, Murnaghan J, Kodama R, Reznick R (1994) A comparative analysis of the costs of administration of an OSCE. *Acad Med* 69(7):571-576

Educating Future Physicians for Ontario: *How to run an OSCE*. (1994) (video and manual) University of Toronto, University of Ottawa, Queen's University, McMaster University and University of Western Ontario

Famuyiwa OO, Zachariah MP and Ilechukwu STC: (1991) The objective structured clinical exam in psychiatry. *Med Educ*, 25: 45-50

Hanson M, Hodges B, McNaughton N, Regehr G. (1998) The integration of child psychiatry into a psychiatry clerkship OSCE. *Canadian Journal of Psychiatry* 43: 614-618

Harden RM and Gleeson FA (1979) Assessment of clinical competence using an observed structured clinical examination. *Med Educ* 13:41-47

Hodges B, Turnbull J, Cohen R, Bienenstock D, Norman G. (1996) Evaluating communication skills in the OSCE format: Reliability and generalizability. *Medical Education* 30: 38-43

Hodges B, Lofchy J. (1997) Examining psychiatry clinical clerks with a mini-OSCE. *Academic Psychiatry* 21(4), 219-225

Hodges B, Regehr G, Hanson M, McNaughton N. (1997) Evaluating psychiatric clinical clerks with an objective structured clinical examination. *Academic Medicine* 72(8): 715-721

Hodges B, Regehr G, Hanson M, and McNaughton N. (1998) The objective structured clinical exam in psychiatry: A validation study. *Academic Medicine* 73(8): 74-76

Hodges B, Hanson M, McNaughton N, Regehr G. (2002) Creating, maintaining and improving a psychiatry OSCE: A guide for faculty. 60-page manuscript published as a special edition of *Academic Psychiatry*. In press for vol 26(3).

Loschen EL (1993) Using the objective structured clinical examination in a psychiatry residency. *Acad Psychiatry* 17:2.

Rothman AI, Cohen R. (1995) Understanding the Objective Structures Clinical Examination (OSCE): Issues and Options. *Annals RCPSC*.

### **Oral Examinations**

Abrahamson S, (1985) The oral examination: The case for and against, in *Evaluating the Skills of Medical Specialists*, American Board of Medical Specialties, Chicago, pp121-124.

Finberg L and Lloyd JS, Suggested guidelines for ideal oral examinations, in *Specialty Board Certification*, JS Lloyd (Ed), American Board of Medical Specialties, Chicago, 1985, pp125-131.

Jayawickramarajah PT (1985) Oral examinations in medical education. *Med Educ*, 19:290-293.

Leichner P, Sisler GC, Harper D (1984) A study of the reliability of the clinical oral examination in psychiatry. *Can J Psychiatry*, 29(5):394-397.

Leichner P, Sisler GC, Harper D (1986) The clinical oral examination in psychiatry: The patient variable, *Annals RCPSC*, 19:283-284.

McGuire CH (1966) The oral examination as a measure of professional competence. *J Med Educ*, 41:267-274.

Muzzin LJ and Hart L (1985) Oral Examinations in *Assessing Clinical Competence*, edited by Neufeld VR and Norman GR. Springer Publishing Company, London.

## **Scoring and Psychometric Issues**

Charlin B, Tardif J and Boshuizen PA. (2000) Scripts and medical diagnostic knowledge: Theory and applications for clinical reasoning instruction and research *Academic Medicine* 72(2), 182-190.

Herold McIlroy J, Hodges B, McNaughton N, Regehr G. (2002) The effect of candidates' perception of evaluation methods on reliability of checklist and global rating scores in an objective structured clinical examination. *Academic Medicine* (In press)

Hodges B, Regehr G, McNaughton N, Tiberius R, Hanson M. (1999) OSCE checklists do not capture increasing levels of expertise. *Academic Medicine* 74(10):64-69

Hodges B, McNaughton N, Regehr G, Hanson M, Tiberius R. (2002) Improving OSCE measure to capture the characteristics of expertise. *Medical Education* (in press)

Hunter DM, Jones RM and Randhawa BS. (1996) The use of holistic versus analytic scoring for large-scale assessment of writing. *The Canadian Journal of Program Evaluation* 11 (2): 61-85.

Regehr G, MacRae H, Reznick R, Szalay D. (1998) Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. *Academic Medicine* 73(9):993-997.

Regehr G, Freeman R, Hodges B, Russell L. (1999) Assessing the generalizability of OSCE measures across content domains. *Academic Medicine*

Reznick R, Regehr G, Yee G, Rothman A, Blackmore D, Dauphinee D. (1998) Process rating forms versus task specific checklists in an OSCE for medical licensure. *Academic Medicine* 73(10):S97-S99.

Schmidt HG, Norman GR, Boshuizen E. (1990) A cognitive perspective on medical expertise: Theory and implications. *Academic Medicine* 65, 611-621.

## **Interprofessional Learning**

Leaviss J. (2000) Exploring the Perceived Effect of an Undergraduate Multiprofessional Educational Intervention. *Medical Education* 34:483-486.

Parsell G, Bligh J. (1998) Interprofessional learning. *Postgrad Med J* 74:89-95.

Yarborough M, Jones T, Cyr TA, Phillips S, Stelzner D. (2000) Interprofessional Education in Ethics at an Academic Health Sciences Centre. *Acad Med* 75:793-800.

## **Psychiatry Resident Assessment**

Hodges B, Hanson M, McNaughton N, Regehr G. (1999) What do psychiatry residents think of an Objective Structured Clinical Examination? *Academic Psychiatry*, 23(4), 1-7.

Shanfield SB, Terrell CD and Littlefield JH. (1997) Using narratives to evaluate psychiatry residents' competence. *Academic Psychiatry* 21(2)

Sierles FS, Daghestani A, Weiner C, de Vito R, Fichtner CG and Garfield DAS.(2001) Psychometric properties of ABPN-style oral examinations jointly administered by two psychiatry residency programs. *Academic Psychiatry* 25(4)

Woodman C and Schultz SK. (1999) Faculty assessment of residents and the psychiatry resident in-training examination: Is there a correlation? *Academic Psychiatry* 23(3)

## **Peers and Self-Assessment**

Arnold L, Willoughby TL, Calkins EV: (1985) Self-evaluation in undergraduate medical education: A longitudinal perspective. *J Med Educ* 60:21.

Cavanagh, G and Styles, K (1988) Focusing on process: Some principles and practices for self and peer evaluation *Indirections* vol 13(2).

Furhman B, Weissburg M (1987) Self-evaluation In *Evaluating Clinical Competence in the Health Professions* M Morgan, D Irby, eds. 139-150, Mosby, St Louis, Missouri.

Gordon MJ (1991) A review of the validity and accuracy of self-assessments in health professional training *Academic Medicine* 66(12):762-769

Hays RB, (1990) Self-evaluation of videotaped consultations. *Teaching and Learning in Medicine* 2(4) 232-236

Henbest RJ, Fehrsen GS. (1985) Preliminary study at the Medical university of Southern Africa on student self-assessment as a means of evaluation. *Journal of Medical Education* 60, 66-67.

Hodges B, Regehr G, Martin D. (2001) Difficulties in recognizing one's own incompetence: Doctors who are unskilled and unaware of it. *Academic Medicine*, October 2001, 76(10 Suppl):S87-9.

Kennell JH, Tempio CR, Wile MA: (1973) Self-evaluation by first year medical students in a clinical sciences programme. *Br J Med Educ* 7(4):230

Klevans DR, Smutz WD, Shuman SB, Bershad C (1992) Self-assessment: Helping professionals discover what they do not know. *New Directions for Adult and Continuing Education* 55:17.

Lichtenstein S and Fischhoff B (1977) Do those who know more also know more about how much they know? *Organisational Behaviour and Human Performance* 20:159-183.

Mahoney MJ, Moore BS, Wade TC et al: (1973) Effects of continuing and intermittent self-monitoring on academic behaviour. *J Consult Clin Psychol* 41:65.

Regehr G, Hodges B, Tiberius R, Lofchy J. (1996) Measuring self-assessment skills: An innovative relative ranking model. *Academic Medicine* Vol. 71, No. 10, PP 52-54.  
October supplement