



SPECIAL ARTICLE

Guidelines for Assessing Clinical Competence

David Newble

*University of Adelaide
Adelaide, Australia
(Editor)*

Dale Dauphinee

*McGill University
Montréal, Canada*

Morag Macdonald

Edinburgh, Scotland

Helen Mulholland

*University of Dundee
Dundee, Scotland*

Beth Dawson

*Southern Illinois University
Springfield, Illinois, USA
(Editor)*

Gordon Page

*University of British Columbia
Vancouver, Canada*

David Swanson

*National Board of Medical Examiners
Philadelphia, Pennsylvania, USA*

Alex Thomson

*University of Tasmania
Hobart, Australia*

Cees van der Vleuten

*University of Limburg
Maastricht, The Netherlands*

Editors' Note: The authors of this article provide literature-based guidelines for anyone wanting to follow the strategies outlined in Section Three to "formally assess the clinical skills of medical students.") They also emphasize the need to define objectives for assessing clinical competence, as outlined in the strategies in Section One: "Define objectives for student education and assessment."

Medical organizations responsible for assessing the clinical competence of large numbers of examinees have traditionally used written, oral, and observation-based examination methods. The results from these examinations form the basis for major professional decisions regarding promotion or privileges of medical students or physicians. In this article, we present a set of guidelines that examining bodies should follow in developing and implementing assessment procedures that are a valid reflection of examinees' current level of competence and of their ability to perform satisfactorily at the next stage of training or practice. The guidelines are based on our collective experiences as well as the growing literature on assessment of clinical competence. The discussion covers issues that have not been fully addressed in previously published reviews, including identifying the competencies to be tested, selecting appropriate and realistic test methods, dealing with test administration and scoring, and setting standards for the desired level of performance.

We gratefully acknowledge the substantive contributions the following individuals made to the topics discussed in this article: Howard Barrows, Laeora Berkson, Georges Bordage, Janet Grant, Brian Joly, Geoff Norman, Emil Petrusa, and Reed Williams.

For a more extensive discussion of the issues raised in this article, see the chapter published in *The Certification and Recertification of Doctors: Issues in the Assessment of Clinical Competence*. Cambridge University Press, 1994.

Correspondence may be sent to David Newble, Associate Professor of Medicine, Department of Medicine, The Queen Elizabeth Hospital, Woodville, South Australia 5011 Australia.

Medical schools, medical licensing authorities, and specialty certification bodies are responsible for developing procedures to assess the clinical competence of large numbers of examinees. Historically, these organizations have used written, oral, and observation-based examination formats. The performance of examinees, as measured by these procedures, provides the basis for decisions regarding promotion to the next stage of training, entry into medical practice, or determination of hospital privileges. The decision process assumes, sometimes with little foundation, that the scores derived from the assessment procedures are a valid reflection of examinees' current level of competence and of their readiness to perform satisfactorily at the next stage of training or practice.

In this article, we present a set of guidelines that examining bodies should follow, or be cognizant of, in developing and implementing assessment procedures. The perspective offered reflects our experiences as well as a literature base that has become considerably more extensive in recent years. These guidelines do not constitute a comprehensive review of the literature on the assessment of clinical competence; rather, they discuss some key issues that have not been fully addressed in reviews published elsewhere.¹⁻⁴ Throughout, the discussion assumes the assessment is intended for a group of medical students, but the comments apply equally well to residents or practicing physicians.

The guidelines also assume that the purpose of the assessment has been determined. They are divided into the following sections: defining what is to be tested; selecting test methods; addressing issues of test administration and scoring; and setting standards for performance.

Defining What Is to Be Tested

The most important phase in developing an assessment procedure is defining what is to be tested. This definition guides test development and forms the basis for evaluating the content validity of the assessment procedure.⁵⁻⁷ It determines whether the procedure provides a representative sample of the competencies expected of the examinee (e.g., the medical student). If an assessment procedure does not possess content validity, all of its other attributes (e.g., face validity, internal consistency, reliability, or reproducibility) are of little consequence.

The approach to determining what is to be tested can be divided into three steps. The first two steps define the range of competencies (exit or terminal objectives) that students must know or be able to do at the end of the course of study. The third step identifies the sample of competencies to be tested in the assessment procedure.

Step 1: Identify the Clinical Desired Level of Resolution

Clinical problems can be identified by simply asking medical faculty to list the problems relevant to their area of specialization. If identified in this manner, the problems should be reviewed by generalists and other specialists to ensure their applicability. A second approach is to observe and analyze the problems and tasks to be faced by the graduates at the next stage of training, in most situations, some form of internship. It is particularly important to identify the problems a beginning intern is expected to handle with relative independence.

The problems may be stated as presenting complaints, such as headache, cough, or diarrhea, or they may be phrased in the form of conditions, such as diabetes, chronic bronchitis, or stroke. It may be desirable to include a range of expected practical skills, such as intravenous drip insertion, catheterization, or suturing. In compiling the list of problems, other considerations include the frequency of occurrence, the importance of early detection, the severity, and the need for immediate intervention at the time of patient contact.

Step 2: For Each Problem, Define the Clinical Tasks at Which Students Are Expected To Be Competent

The term *clinical task* refers to actions that are specific to particular clinical problems (e.g., characterizing chest pain as suggestive of impending myocardial infarction, ordering laboratory studies to evaluate thyroid function). Clinical tasks comprise the range of knowledge or skills students should know or have. For patient-related problems, clinical tasks consist of all aspects of patient workup and management, including eliciting and interpreting history (medical and psychosocial), physical examination, and laboratory data; selecting and administering treatment; arranging follow-up; talking to relatives; and so on. For public health or community-related problems, clinical tasks comprise of all aspects of problem identification, as well as the actions (e.g., health education, inspection of health standards) that should be taken to deal with the problem.

Research on assessing clinical competence has repeatedly shown that satisfactory performance of the clinical task on one problem does not provide a basis for accurate prediction of the ability to perform a similar task on a different problem.⁸⁻¹¹ For example, a student who demonstrates competence in eliciting an appropriate history from a patient with acute chest pain may not demonstrate a similar degree of competence with a patient with an acute abdomen. This surprising but consistent characteristic of clinical competence assessment has important consequences for the sampling of problems and the length of tests.

A difficult aspect of clinical task definition is specifying the expected level of performance. At a given stage of training, some clinical tasks may be relatively trivial, whereas others may exceed reasonable expectations. For the problem of diabetes, for example, it may be reasonable to expect graduating students to manage a patient with controlled, adult-onset Type II diabetes, but it may not be reasonable to expect these students to exhibit the same degree of proficiency with an adolescent patient with unstable Type I diabetes.

Step 3: Prepare a Blueprint to Guide the Selection of Problems to Be Included in the Assessment Procedure

Test blueprints can be very simple and unidimensional, such as assigning percentage weights to disciplines covered on an examination (although this approach is not advised because disciplinary content coverage is often poorly coordinated). Blueprints can also be complex, multidimensional "grids" in which the dimensions reflect patient problems, organ systems, clinical setting, patient age, or activities composing the diagnostic/therapeutic process.

We advocate strongly the blueprint approach to ensure that the clinical problems (and their embedded clinical tasks) are representative and constitute an adequate sample of problems students should be capable of resolving. Such blueprints then guide the selection of problems from the range of problems previously defined. This practice also ensures that subsequent test forms constructed from the same blueprint will assess reasonably parallel content.

Blueprints should not be used to predetermine the distribution of clinical tasks to be tested; clinical tasks should be defined in relation to the nature of the clinical problems themselves. However, the distribution of clinical tasks should be checked retrospectively to ensure a reasonable balance. An appropriate distribution of other characteristics, such as patient age and sex, or the problem prevalence, severity, and treatability, should be considered when cases are being prepared to represent the selected problems.

This approach to blueprinting assumes that clinical tasks should be tested in the context of specific, relevant clinical problems and not in isolation. Although there may be subcategories of needed skills related to a given clinical task, these are interrelated within any specific problem and should not be tested in isolation except, perhaps, at early stages of clinical training. The assessment procedure should be designed to test only those tasks that are most critical to the successful resolution of the problem and that examinees will be expected to perform relatively independently in their next clinical setting.

These steps, although time consuming, provide a practical yet comprehensive approach to the identifica-

tion of the content to be tested with the assessment procedure. The subsequent use of elegant testing methods, computerized scoring systems, and/or sophisticated analytic procedures cannot compensate for a failure to define adequately the content to be tested.

Selecting Test Methods

These guidelines cannot detail the various test methods available for use in the assessment of clinical competence; this has been done elsewhere.^{1,2,11} Rather, we wish to emphasize an approach to the selection of methods based on the content to be tested. Three steps should be taken in selecting testing methods to implement the approach we advocate.

Step 1. Select Test Methods That Are Most Appropriate to the Clinical Tasks Being Assessed

Recent efforts have been made to develop forms of assessment that are more valid, efficient, and reliable than traditional forms of clinical assessment. These methods include paper-and-pencil and computer-based patient management problems (PMPs), written key-issue simulations,¹² the objective structured clinical examination (OSCE) or multiple-station examinations, and standardized patients (SPs). Each of these approaches has its limitations, one of the most frustrating being the prolonged testing time required to achieve an acceptable level of reliability.^{8,9,13} This limitation, common to all methods for assessing aspects of clinical competence, occurs because of the inherent nature of dealing with clinical problems in which performance on one problem does not predict performance on another. This phenomenon has been referred to as case specificity, content specificity, and problem specificity.¹⁴⁻¹⁶

When the purpose of the test is limited to determining if a student can identify the appropriate actions to take in a specific situation (e.g., ordering diagnostic studies), a low-fidelity method, such as a paper-and-pencil test, will suffice. On the other hand, history-taking or counseling tasks that require effective interaction with the patient are likely to require relatively higher fidelity methods, such as real or SP cases.

Assessing discrete, well-structured skills, such as the ability to perform a knee examination or auscultate the lungs, may be accomplished using relatively short stations (e.g., 5 min in length) as contained on the typical OSCE. Longer stations are needed to assess more complex skills, the ability to integrate information, such as from the history and physical examination, or to determine whether students will use the skill with an appropriate clinical problem.¹⁷ Depending on the goals of the assessment, it can be appropriate to combine paper-and-

pencil tasks with SP cases, such as interpreting an x-ray after the examination of a patient's chest. Research is needed on the effect of station complexity on the reliability of examinations.

Step 2. Let the Clinical Task Dictate the Method by Which It Is Tested

Letting the purpose determine the method may seem so obvious that it need not be mentioned. Unfortunately, selection of assessment methods commonly occurs first and subsequently determines to a large extent the tasks that can be tested. For example, an exclusive reliance on a multiple-choice question examination will result in a restricted and low-fidelity assessment of many clinical tasks. On the other hand, clerkship ratings provided by faculty may appear to have high fidelity but may fail to assess clinical tasks involving history taking and physical examination simply because students are rarely directly observed. Because no assessment method is optimal in assessing all tasks, a comprehensive assessment procedure will include more than one testing method.

Step 3. Recognize the Practical Constraints on Selecting Optimal Examination Methods

It is not usually possible to achieve an ideal match between the tasks to be posed and the method of assessment. Constraints include the amount of time available for the test; the resources available for constructing and conducting the examination (e.g., money, examiners, organizers, patients, facilities); the measurement characteristics of the available test methods; and the acceptability of the ideal approach to faculty, examinees, and the profession.

Addressing Issues of Test Administration and Scoring

Administrative problems, such as the need to test large numbers of students at different sites or times and resource limitations, have often led to the adoption of multiple-choice test methods as a major component of many assessment procedures. Although the goals of broad sampling, efficiency, and reliability are easily achieved using multiple-choice tests, there are significant disadvantages in relying too heavily on this format. Multiple-choice tests tend to assess behaviors in isolation rather than as an integrated whole, and they are unlikely to test the full range of competencies of prime interest to the examiners. Additionally, overreliance on multiple-choice tests ignores the principle that assessment will exert a strong influence on learning.^{18,19} Clearly, it is desirable to match the methods we use to the aspects of competence we aim to assess, even

though this requires use of a wider range of methods than is generally the case.

Six steps are needed to ensure proper administration and scoring of a test.

Step 1: Decide on the Level of Efficiency Needed in the Particular Testing Environment

Various test-administration strategies are being explored to minimize some of the difficulties associated with long examinations. A most promising approach is based on variable-length testing procedures. The *multiple-hurdles* approach is used by many postgraduate colleges and some specialty boards. An effective and efficient strategy is to use as the first hurdle a simple method with a high failure rate, such as a fairly difficult multiple-choice test. The second, resource-intensive hurdle, such as a clinical or oral examination, is then reached only by a subset of the examinee population. To be psychometrically sound, this design requires that the second hurdle be free from any shortcomings discussed in the previous section.

Another approach is *adaptive testing*, in which each component or item on the examination is selected on the basis of the student's performance on the previous component(s) or item(s), with the criteria for selection designed to maximize the amount of information obtained. Although it is theoretically sound, the complexity of this procedure means that it is practical only with computer-based tests; therefore, it appears to be a limited option for the testing formats used in the assessment of clinical competence.

Sequential testing is a step toward adaptive testing but is less complex. The test length needed to make a reproducible decision about the competence of a student decreases as the difference increases between the student's level of competence and the pass/fail cutoff score. Therefore, a relatively short (and thus relatively unreliable) examination may quite accurately identify students well above or well below the cutting score. These students can safely be excused from further testing, and only those for whom doubt persists need continue with a further sequence of the examination procedure. Sequential testing does not reduce the number of hours of testing time required for examinees near the cutting score. However, it reduces the total resources required, particularly if a substantial portion of the examination involves nonwritten simulations. For example, an OSCE requiring three long test rotations to accommodate all students might be reduced to one long circuit for all students with additional second or third rotations for students about whom doubt still exists. An important assumption of sequential testing is that each rotation consists of a random selection from the overall content to be tested (i.e., the rotations are parallel). Recent examples of sequential

testing procedures are now available in the context of SP examinations.^{20,21}

Step 2: Decide How the Student's Performance Is To Be Recorded or Captured

There are a number of technical issues to be addressed in the scoring of complex clinical simulations.^{22,23} The observers may be faculty examiners or other trained observers. On simulations using SPs, students may be observed and scored by the SPs themselves, although extensive training may be required.¹³ Fine distinctions among specific behaviors may not be possible; the complexity of the simulation affects a SP's ability to remember the actions and behaviors of the student. Observation by clinical faculty examiners increases the costs of administering the examination, although some of the cost is offset by reduced time for training. Faculty examiners may be able to assess more complex skills than other observers, and an added benefit is the feedback they receive regarding the effectiveness of their teaching and the strengths and weaknesses of the curriculum. Observers other than clinical faculty can also be used; the amount of training required depends on their level of expertise. In simulations involving SPs, an advantage of using trained observers other than the SP is their ability to record the student's behavior as it occurs.

Both checklists and rating scales are commonly used to record performance on tests in which a student's behavior is observed.²⁴ Checklists have demonstrated higher agreement among observers than rating scales, but the decision regarding which to use should be dictated by the skills to be tested.¹³ Although research is limited, when completed by trained observers, longer checklists tend to be more reliable than shorter ones.²⁵ When completed by SPs, accuracy begins to decrease if more than 15–20 items have to be recorded.²⁶

As with all examinations using observers, there is evidence for systematic rater bias (i.e., some observers are "hawks" and some are "doves"). Rater bias is not a problem if all students are scored by the same rater, and a relative standard (see next section) is used. If not, this bias can be minimized by training observers to attain greater consistency or, with a large number of cases, by randomly assigning students to raters.²⁶

Step 3: Determine a Method To Assign Scores to the Cases and/or Elements Within Cases

In discussing specific scoring topics, we use the term *case* to refer to a conceptual unit on the examination,

whether an item or set of items on a written examination, a PMP, or a station on an OSCE. Scores on cases may be calculated as a percentage of all possible points assigned to the elements in the case or as a simple score (0–1 or pass/fail) for the entire case. The latter approach produces a total test score that is less reproducible because binary or dichotomous scoring leads to a loss of measurement information. On the other hand, 0–1 scores may be more meaningful as they more directly reflect acceptable performance on the cases in the examination. There has been no reported research on how to combine elements within cases or how to combine cases themselves when tested by different testing formats; the entire issue of how to form case scores deserves further exploration.²⁷

Weighting of items or elements in a case does not appear to have a major effect on the reproducibility of scores.²⁸ Therefore, we advocate that weighting be viewed in the context of validity (i.e., the weighting of elements in cases should be guided by their importance). Weighting of content across cases should be accomplished through the design of the test blueprint rather than by weighting scores on some cases more heavily in calculation of a total score (i.e., more important tasks should be assessed more frequently).

The scores on cases can be combined so that each case receives a single score or so that certain elements can be combined across cases to produce a score in a well-defined skill, such as history taking or laboratory use. The approach selected should be guided by the purpose and design of the examination. The latter approach requires rigorous definition and sampling of the skills to be tested. In addition, elements purported to test a specific skill should be highly correlated across cases.

The use of scoring keys prepared before administration of the examination is a prerequisite for efficient and timely scoring of any examination, but they are especially crucial with many of the complex formats discussed in this article. Sufficient sampling of expert opinion during the preparation of the examination is necessary to ensure that the key is valid. After the examination has been administered, the key should be reviewed in the light of examinee performance.

Step 4: Take Appropriate Steps To Ensure That the Test Provides an Unbiased Measure of Performance

Bias is a situation in which the test score has meanings or implications for a subgroup of test takers that are different from the meanings or implications for other test takers of the same ability.²⁹ Although bias may be present in written tests, there are additional risks for bias in examinations involving observations of behavior.³⁰ These risks increase as the behavior being observed becomes more difficult to define and is scored more subjectively.

Little research has been reported on bias and its impact on scores and on pass/fail decisions when observers are used to rate performance. In particular, bias that might be related to factors such as sex and ethnicity, from the perspective of both raters and examinees, has not been systematically investigated, though reports are now appearing in the literature.³¹ Guidelines are needed for how to avoid, monitor, detect, and adjust for bias.

Step 5: Evaluate the Need for Equating Scores Across Different Examinations

Equating is the process of ensuring that scores on two different tests, developed according to the same blueprint, are interchangeable.³² Equating is necessary to ensure comparability of scores and pass/fail decisions across multiple test forms and testing occasions. For written tests, a common equating method is to embed the same items (the link or anchor items) in the different forms of the examination (say 30% of the total items). It is important that the anchor items fully represent the content blueprint for the entire examination. To equate two test forms, the performance on the anchor items is compared across the forms to estimate the relative ability of the group taking each test. The relationship between performance on the anchor items and other items in a form is then used to place scores on the same scale.

Information is scarce on appropriate methods of equating except for those based on written tests.³³ Equating would seem to be a particularly difficult problem for some of the testing formats used in clinical competence assessment; probably each case should be treated as an "item." With SPs, it may not be logistically feasible to have the same person portraying the case from one time or site to another, posing problems of equivalence. Additionally, the traditional guideline of 30% overlap in link or anchor items may not be sufficiently large for effective equating of examinations using SPs. At the same time, reuse of a larger number of cases may pose security problems. Limited research has been performed in this area, and more is clearly needed.³⁴

Step 6: Review the Procedure To Ensure That Trivialization Has Not Occurred

Trivialization is a phenomenon to which all forms of testing are susceptible. Multiple-choice questions have frequently been criticized for their tendency to focus on assessing knowledge of isolated facts, thus trivializing what is being tested. Performance-based tests are not immune from similar criticisms. Trivialization in checklist development and in the selection of scoring elements is a danger.³⁵ Unfortunately, it is easy to have a reliable finished test on which students have been scored on a set of clearly defined criteria that do not reflect the student's real ability or overall performance.

The criteria may reflect only those behaviors or skills that are easy to measure, ignoring components of clinical competence that most would agree are more critical and valid.^{22,23}

Setting Standards for Performance

Standard setting is the process of determining the score needed to pass an examination. The true nature of standard setting is often not fully appreciated by examining bodies (i.e., all standard-setting procedures are, to a degree, arbitrary but need not be capricious). Two steps are involved in setting standards and communicating the results.

Step 1. Determine the Type of Standard Desired and an Appropriate Standard-Setting Method.

Standards are commonly classified as either relative or absolute.³⁶ Relative standards depend on the performance of the examinees taking the examination and are the most frequently used. For example, a test in which the lowest scoring 20% of students fail, or one in which any student scoring less than one standard deviation below the mean fails, has a relative standard.

Absolute standards can be based on an analysis of the content of the examination or the choice of an arbitrary percentage correct for passing; they are independent of the performance of the examinees. We recommend that, in principle, absolute standards should be used because a student's passing of the examination should not depend on the performance of other students taking the examination.

Two content-based methods for setting absolute standards on written tests are the Angoff method and the Ebel method. Both require that a group of experts review the items or elements on the examination very carefully and estimate how often a borderline examinee will perform satisfactorily on the item. The Hofstee method is a compromise between absolute and relative standards. There is limited experience in using these methods on performance-based tests.³⁷ However, analogues to these methods are currently a major area of research.

Step 2: Develop Procedures for Effectively Communicating the Results of the Test

Regardless of how the scores are formed and standards are set, the test-development process must ensure that all reported scores are both reliable and valid. Scores should not be reported on subtests and/or cases unless they reach an accepted level of reliability. In general, scores on individual cases do not meet these criteria and should not be reported. When possible, standard errors of measurement should be reported.

Because the major purpose of reporting scores is to provide effective communication regarding performance, many of the guidelines for reporting research data apply.³⁸ Graphs are useful to provide information about size of standard errors, especially if there are subtests. Graphs and tables summarizing the performance of all examinees are also appropriate for tests using a relative standard. In all situations, adequate description of the meaning of the scores and the correct interpretations and inferences should be provided.

Summary

The purpose of these guidelines is to provide a rationale and some suggestions for developing tests that assess clinical competence. Most of the illustrations have focused on the assessment of skills in patient care, but many of the points we have made are relevant to testing a student's skill in dealing with public health or community-related issues as well.

The cornerstone of developing a test of clinical competence is careful and thorough planning—identifying the problems, specifying the clinical tasks pertinent to the problems, and preparing a blueprint for the test. Problems that are important should be heavily represented in the blueprint and form a large component of the examination. Once these initial planning steps are completed, the test methods that provide the most real-to-life testing of the problems should be selected for the assessment procedure. Some clinical tasks, such as history taking and communication with the patient, are best assessed using an OSCE format with SPs. There is considerable experience in the effective administration of this examination format, especially in a single site, such as a medical school. Many of the problems related to wide-scale, multisite testing are currently being addressed by testing agencies in both the United States and Canada. Throughout our discussion, a recurring theme has been the need for more research, especially in the general areas of scoring and standard setting.

References

1. Neufeld VR, Norman GR (Eds.). *Assessing clinical competence*. New York: Springer, 1985.
2. Wakeford RE (Ed.). *Directions in clinical assessment*. Proceedings of the First Cambridge Conference. Cambridge, England: Cambridge University School of Medicine, 1985.
3. Newble DI. Assessing clinical competence at the undergraduate level. ASME Medical Education Booklet No. 25. *Medical Education* 1992;6:504-11.
4. Kane MT. The assessment of professional competence. *Evaluation in the Health Professions* 1992;15:163-82.
5. American Education Research Association, American Psychological Association and National Council on Measurement in Education. *Standards for educational and psychological testing*. Washington, DC: American Psychological Association, 1985.
6. Ebel RL. The practical validation of tests of ability. *Educational Measurement: Issues and Practice* 1983;2:7-10.
7. Kane MT. The validity of licensure examinations. *American Psychologist* 1982;37:911-18.
8. Swanson DB. A measurement framework for performance-based tests. In IR Hart, RM Harden (Eds.), *Further developments in assessing clinical competence* (pp. 13-45). Montreal: Can-Heal, 1987.
9. Newble DI, Swanson DB. Psychometric characteristics of the objective structured clinical examination. *Medical Education* 1988;22:325-34.
10. Colliver JA, Verhulst SJ, Williams RG, Norcini JJ. Reliability of performance on standardized patient cases: A comparison of consistency measures based on generalizability theory. *Teaching and Learning in Medicine* 1989;1:31-7.
11. Van der Vleuten CPM, Newble DI. Methods of assessment in certification. In DI Newble, B Jolly, RE Wakeford (Eds.), *The certification and recertification of doctors* (pp.105-25). Cambridge, England: Cambridge University Press, 1994.
12. Bordage G, Page G. An alternative approach to PMPs: The "key features" concept. In IR Hart, RM Harden (Eds.), *Further developments in assessing clinical competence* (pp. 59-72). Montreal: Can-Heal, 1987.
13. Van der Vleuten CPM, Swanson DB. Assessment of clinical skills with standardized patients: State of the art. *Teaching and Learning in Medicine* 1990;2:58-76.
14. Barrows HS, Neufeld VR, Feightner JW, Norman GR. *An analysis of the clinical methods of medical students and physicians*. Toronto: Final Report to Ontario Ministry of Health, 1978.
15. Elstein A, Shulman L, Sprafka S. *Medical problem solving*. Cambridge: Harvard University Press, 1978.
16. Norman GR. Problem-solving skills, solving problems and problem-based learning. *Medical Education* 1988;22:279-86.
17. Barrows HS. On overview of the uses of standardized patients for teaching and evaluating clinical skills. *Academic Medicine* 1993;68:443-53.
18. Frederiksen N. The real test bias. *American Psychologist* 1984;37:911-18.
19. Newble DI, Jaeger K. The effect of assessments and examinations on the learning of medical students. *Medical Education* 1983;17:165-71.
20. Swanson DB, Norcini JJ. Factors influencing the reproducibility of tests using standardized patients. *Teaching and Learning in Medicine* 1989;1:158-66.
21. Colliver JA, Vu NV, Barrows HS. Screening test length and pass-fail cut-offs for sequential testings with a standardized-patient based examination: A receiver operating characteristic (ROC) analysis. *Academic Medicine* 1992;67:592-5.
22. Van der Vleuten CPM, Norman GR, de Graaff E. Pitfalls in the pursuit of objectivity: Issues of reliability. *Medical Education* 1991;25:110-18.
23. Norman GR, van der Vleuten CPM, de Graaff E. Pitfalls in the pursuit of objectivity: Issues of validity, efficiency and acceptability. *Medical Education* 1991;25:119-26.
24. Stillman PL. Technical Issues: Logistics. *Academic Medicine* 1993;68:464-70.
25. Tamblyn RM. Use of standardized patients in the assessment of clinical competence (Doctoral dissertation, McGill University, 1991). *Dissertation Abstracts International*, 51, 4149.
26. Tamblyn RM, Klass DJ, Schnable GK, Kopelow ML. Sources of unreliability and bias in standardized-patient rating. *Teaching and Learning in Medicine* 1991;3:74-85.
27. Swanson DB, Norcini JJ, Grosso LJ. Assessment of clinical competence: Written and computer-based simulations. *Assessment and Evaluation in Higher Education* 1987;12:220-46.
28. Norcini JJ, Swanson DB, Webster GD, Grosso LJ. A comparison of several methods for scoring patient management problems. In *Proceedings of the 22nd Annual Conference on Research in Medical Education* (pp. 41-6). Washington, DC: Association of American Medical Colleges, 1983.
29. Cole NN, Moss PA. Bias in test use. In RL Linn (Ed.), *Educational measurement* (3rd ed., pp. 201-20). New York: Macmillan, 1989.

30. Dawson-Saunders B. Examining bias on OSCE examinations. In RM Harden, IR Hart, H Mulholland (Eds.), *Approaches to the assessment of clinical competence* (pp. 409-14). Dundee: Centre for Medical Education, 1992.
31. Rutala PJ, Witzke DB, Fulginiti JV. The influence of student and standardized patient gender on scoring in an objective structured clinical examination. *Academic Medicine* 1991;66:S28-S30.
32. Peterson NS, Kolen MJ, Hoover HD. Scaling, norming and equating. In RL Linn (Ed.), *Educational Measurement* (3rd ed., pp. 221-62). New York: Macmillan, 1989.
33. Skaggs G, Lissitz RW. IRT test equating: Relevant issues and a review of recent research. *Review of Educational Research* 1986;56:495-529.
34. Rothman AI, Cohen R, Dawson-Saunders B, Poldre PP, Ross J. Testing the equivalence of multiple-station tests of clinical competence. *Academic Medicine* 1992;67:40-1.
35. Norman GR. Summary of the conference. In W Bender, RJ Heimstra, AJJA Scherpier, RP Zwierstra (Eds.), *Teaching and assessing clinical competence* (pp. 599-609). Groningen: Boekwerk, 1989.
36. Livingston SA, Zieky MJ. *Passing scores*. Princeton: Educational Testing Service, 1982.
37. Meskauskas JA. Setting standards for credentialing examinations. *Evaluation in the Health Professions* 1986;9:187-203.
38. Wainer H. Understanding graphs and tables. *Educational Researcher* 1992;21:14-23.

Received 24 July 1992

Final revision received 23 July 1993