Validity: on the meaningful interpretation of assessment data

Steven M Downing

Context All assessments in medical education require evidence of validity to be interpreted meaningfully. In contemporary usage, all validity is construct validity, which requires multiple sources of evidence; construct validity is the whole of validity, but has multiple facets. Five sources – content, response process, internal structure, relationship to other variables and consequences – are noted by the *Standards for Educational and Psychological Testing* as fruitful areas to seek validity evidence.

Purpose The purpose of this article is to discuss construct validity in the context of medical education and to summarize, through example, some typical sources of validity evidence for a written and a performance examination.

Summary Assessments are not valid or invalid; rather, the scores or outcomes of assessments have more or less evidence to support (or refute) a specific interpretation (such as passing or failing a course). Validity is approached as hypothesis and uses theory, logic and the scientific method to collect and assemble data to support or fail to support the proposed score interpretations, at a given point in time. Data and logic are assembled into arguments – pro and con – for some specific interpretation of assessment data. Examples of types of validity evidence, data and information from each source are discussed in the context of a high-stakes written and performance examination in medical education.

Conclusion All assessments require evidence of the reasonableness of the proposed interpretation, as test data in education have little or no intrinsic meaning. The constructs purported to be measured by our assessments are important to students, faculty, administrators, patients and society and require solid scientific evidence of their meaning.

Keywords Education, Medical, Undergraduate/*standards, Educational measurement, Reproducibility of results.

Medical Education 2003;37:830–837

Introduction

The purpose of this paper is to discuss validity in the context of assessment in medical education and to present examples of the five types of validity evidence typically sought to support or refute the valid interpretations of assessment data.[1] This essay builds on and expands the older and more traditional view of test validity expressed in the first article in this series[2] and extends the validity discussion into state-of-the-art 21st century educational measurement.

Validity refers to the evidence presented to support or refute the meaning or interpretation assigned to assessment results. All assessments require validity evidence and nearly all topics in assessment involve validity in some way. Validity is the *sine qua non* of assessment, as without evidence of validity, assessments in medical education have little or no intrinsic meaning.

Validity is always approached as hypothesis, such that the desired interpretative meaning associated with assessment data is first hypothesized and then data are collected and assembled to support or refute the validity hypothesis. In this conceptualization, assessment data are more or less valid for some very specific purpose, meaning or interpretation, at a given point in time and only for some well-defined population. The assessment itself is never said to be 'valid' or 'invalid' rather one speaks of the scientifically sound evidence presented to either support or refute the proposed interpretation of assessment scores, at a particular time period in which the validity evidence was collected.

In its contemporary conceptualization,[1,3–14] validity is a unitary concept, which looks to multiple sources of

Department of Medical Education (MC 591), College of Medicine, University of Illinois at Chicago, Chicago, Illinois, USA

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

Key learning points

Validity is a unitary concept, with construct validity as the whole of validity.

Assessments are not valid or invalid, rather assessment scores have more (or less) validity evidence to support the proposed interpretations.

Validity requires multiple sources of evidence to support or refute meaningful score interpretation.

Validity is always approached as hypothesis.

Validation research uses theory, data and logic to argue for or against specific score interpretations.

evidence. These evidentiary sources are typically logically suggested by the desired types of interpretation or meaning associated with measures. All validity is construct validity in this current framework, described most eloquently by Messick[8] and embodied in the current *Standards of Educational and Psychological Measurement.*[1] In the past, validity was defined as three separate types: content, criterion and construct, with criterion-related validity usually subdivided into concurrent and predictive depending on the timing of the collection of the criterion data.[2,15]

Why is construct validity now considered the sole type of validity? The complex answer is found in the philosophy of science[8] from which, it is posited, there are many complex webs of inter-related inference associated with sampling content in order to make meaningful and reasonable inferences to a domain or larger population of interest. The more straightforward answer is: Nearly all assessments in the social sciences, including medical education, deal with *constructs* – intangible collections of abstract concepts and principles which are inferred from behavior and explained by educational or psychological theory. *Educational achievement* is a construct, usually inferred from performance on assessments such as written tests over some well-defined domain of knowledge, oral examinations over specific problems or cases in medicine, or highly structured standardized patient examinations of history-taking or communication skills.

Educational *ability* or *aptitude* is another example of a familiar construct – a construct that may be even more intangible and abstract than *achievement* because there is less agreement about its meaning among educators and psychologists.[16] Tests that purport to measure educational ability, such as the Medical College Admissions Test (MCAT), which is relied on heavily in North America for selecting prospective students for medical school admission, must present scientifically sound evidence, from multiple sources, to support the reasonableness of using MCAT test scores as one important selection criterion for admitting students to medical school. An important source of validity evidence for an examination such as the MCAT is likely to be the predictive relationship between test scores and medical school achievement.

Validity requires an evidentiary chain which clearly links the interpretation of the assessment scores or data to a network of theory, hypotheses and logic which are presented to support or refute the reasonableness of the desired interpretations. Validity is never assumed and is an ongoing process of hypothesis generation, data collection and testing, critical evaluation and logical inference. The validity argument[11,12] relates theory, predicted relationships and empirical evidence in ways to suggest which particular interpretative meanings are reasonable and which are not reasonable for a specific assessment use or application.

In order to meaningfully interpret scores, some assessments, such as achievement tests of cognitive knowledge, may require fairly straightforward content-related evidence of the adequacy of the content tested (in relationship to instructional objectives), statistical evidence of score reproducibility and item statistical quality and evidence to support the defensibility of passing scores or grades. Other types of assessments, such as complex performance examinations, may require both evidence related to content and considerable empirical data demonstrating the statistical relationship between the performance examination and other measures of medical ability, the generalizability of the sampled cases to the population of skills, the reproducibility of the score scales, the adequacy of the standardized patient training and so on.

Some typical sources of validity evidence, depending on the purpose of the assessment and the desired interpretation are: evidence of the content representativeness of the test materials, the reproducibility and generalizability of the scores, the statistical characteristics of the assessment questions or performance prompts, the statistical relationship between and among other measures of the same (or different but related) constructs or traits, evidence of the impact of assessment scores on students and the consistency of pass–fail decisions made from the assessment scores.

The higher the stakes associated with assessments, the greater the requirement for validity evidence from multiple sources, collected on an ongoing basis and continually re-evaluated.[17] The ongoing documentation of validity evidence for a very high-stakes testing programme, such as a licensure or medical specialty certification examination, may require the allocation of many resources and the contributions of many different professionals with a variety of skills – content specialists, psychometricians and statisticians, test editors and administrators.

In the next section, five major sources of validity evidence are discussed in the contexts of example assessments in medical education.

Sources of evidence for construct validity

According to the *Standards*: 'Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests'[1] (p. 9). The current *Standards*[1] fully embrace this unitary view of validity, following closely on Messick's work[8,9] that considers all validity as construct validity, which is defined as an investigative process through which constructs are carefully defined, data and evidence are gathered and assembled to form an argument either supporting or refuting some very specific interpretation of assessment scores.[11,12] Historically, the methods of validation and the types of evidence associated with construct validity have their foundations on much earlier work by Cronbach,[3–5] Cronbach and Meehl[6] and Messick.[7] The earliest unitary conceptualization of validity as construct validity dates to 1957 in a paper by Loevinger.[18] Kane[11–13] places validity into the context of an interpretive argument, which must be established for each assessment; Kane's work has provided a useful framework for validity and validation research.

The *Standards*

The *Standards*[1] discuss five distinct sources of validity evidence (Table 1): content, responses, internal structure, relationship to other variables and consequences. Each source of validity evidence (Table 1) is associated with some examples of the types of data that might be collected to support or refute specific assessment interpretations (validity). Some types of assessment demand a stronger emphasis on one or more sources of evidence as opposed to other sources and not all sources of data or evidence are required for all assessments. For example, a written, objectively scored test covering several weeks of instruction in microbiology, might emphasize content-related evidence, together with some evidence of response quality, internal structure and consequences, but very likely would not seek much or any evidence concerning relationship to other variables. On the other hand, a high-stakes

Table 1 Some sources of validity evidence for proposed score interpretations and examples of some types of evidence

Content	Response process	Internal structure	Relationship to other variables	Consequences
• Examination blueprint • Representativeness of test blueprint to achievement domain • Test specifications • Match of item content to test specifications • Representativeness of items to domain • Logical/empirical relationship of content tested to achievement domain • Quality of test questions • Item writer qualifications • Sensitivity review	• Student format familiarity • Quality control of electronic scanning/scoring • Key validation of preliminary scores • Accuracy in combining different formats scores • Quality control/accuracy of final scores/marks/grades • Subscore/subscale analyses: • Accuracy of applying pass-fail decision rules to scores • Quality control of score reporting to students/faculty • Understandable/accurate descriptions/interpretations of scores for students	• Item analysis data: 1. Item difficulty/discrimination 2. Item/test characteristic curves (ICCs/TCCs) 3. Inter-item correlations 4. Item-total correlations • Score scale reliability • Standard errors of measurement (SEM) • Generalizability • Dimensionality • Item factor analysis • Differential Item Functioning (DIF) • Psychometric model	• Correlation with other relevant variables • Convergent correlations - internal/external: 1. Similar tests • Divergent correlations-internal/external 1. Dissimilar measures • Test-criterion correlations • Generalizability of evidence	• Impact of test scores/results on students/society • Consequences on learners/future learning • Positive consequences outweigh unintended negative consequences? • Reasonableness of method of establishing pass-fail (cut) score • Pass-fail consequences: 1. P/F Decision reliability-Classification accuracy 2. Conditional standard error of measurement at pass score (CSEM) • False positives/negatives • Instructional/learner consequences

summative Objective Structured Clinical Examination (OSCE), using standardized patients to portray and rate student performance on an examination that must be passed in order to proceed in the curriculum, might require all of these sources of evidence and many of the data examples noted in Table 1, to support or refute the proposed interpretation of the scores.

Sources of validity evidence for example assessments

Each of the five sources of validity evidence will now be considered, in the context of a written assessment of cognitive knowledge or achievement and a performance examination in medical education. Both example assessments are high-stakes, in that the consequences of passing or failing are very important to students, faculty and, ultimately, patients. The written assessment is a summative comprehensive examination in the basic sciences – a test consisting of 250 multiple-choice questions (MCQs) covering all the pre-clinical instruction in the basic sciences – and a test that must be passed in order to proceed into clinical training. The performance examination is a standardized patient (SP) examination, administered to medical students toward the end of their clinical training, after having completed all of their required clerkship rotations. The purpose of the SP examination is to comprehensively assess graduating medical students' ability to take a history and do a focused physical examination in an ambulatory primary care setting. The SP examination consists of 10 20-minute SP cases, presented by a lay, trained standardized patient who simulates the patient's presenting problem and rates the student's performance at the conclusion of the examination. The SP examination must be passed in order to graduate medical school.

Documentation of these five sources of validity evidence consists of the systematic collection and presentation of information and data to present a convincing argument that it is reasonable and defensible to interpret the assessment scores in accordance with the purpose of the measurement. The scores have little or no intrinsic meaning; thus the evidence presented must convince the skeptic that the assessment scores can reasonably be interpreted in the proposed manner.

Content evidence

For the written assessment, documentation of validity evidence related to the content tested is the most essential. The outline and plan for the test, described by a detailed test blueprint or test specifications, clearly relates the content tested by the 250 MCQs to the domain of the basic sciences as described by the course learning objectives. The test blueprint is sufficiently detailed to describe subcategories and subclassifications of content and specifies precisely the proportion of test questions in each category and the cognitive level of those questions. The blueprint documentation shows a direct linkage of the questions on the test to the instructional objectives. Independent content experts can evaluate the reasonableness of the test blueprint with respect to the course objectives and the cognitive levels tested. The logical relationship between the content tested by the 250 MCQs and the major instructional objectives and teaching/learning activities of the course should be obvious and demonstrable, especially with respect to the proportionate weighting of test content to the actual emphasis of the basic science courses taught. Further, if most learning objectives were at the application or problem-solving level, most test questions should also be directed to these cognitive levels.

The quality of the test questions is a source of content-related validity evidence. Do the MCQs adhere to the best evidence-based principles of effective item-writing.[19] Are the item-writers qualified as content experts in the disciplines? Are there sufficient numbers of questions to adequately sample the large content domain? Have the test questions been edited for clarity, removing all ambiguities and other common item flaws? Have the test questions been reviewed for cultural sensitivity?

For the SP performance examination, some of the same content issues must be documented and presented as validity evidence. For example, each of the 10 SP cases fits into a detailed content blueprint of ambulatory primary care history and physical examination skills. There is evidence of faculty content–expert agreement that these specific 10 cases are representative of primary care ambulatory cases. Ideally, the content of the 10 clinical cases is related to population demographic data and population data on disease incidence in primary care ambulatory settings. Evidence is documented that expert clinical faculty have created, reviewed and revised the SP cases together with the checklists and ratings scales used by the SPs, while other expert clinicians have reviewed and critically critiqued the SP cases. Exacting specifications detail all the essential clinical information to be portrayed by the SP. Evidence that SP cases have been competently edited and that detailed SP training guidelines and criteria have been prepared, reviewed by faculty experts and implemented by experienced SP trainers are all important sources of content-related validity evidence.

There is documentation that during the time of SP administration, the SP portrayals are monitored closely to ensure that all students experience nearly the same case. Data are presented to show that a different SP, trained on the same case, rates student case performance about the same. Many basic quality-control issues concerning performance examinations contribute to the content-related validity evidence for the assessment.[20]

Response process

As a source of validity evidence, response process may seem a bit strange or inappropriate. *Response process* is defined here as evidence of data integrity such that all sources of error associated with the test administration are controlled or eliminated to the maximum extent possible. Response process has to do with aspects of assessment such as ensuring the accuracy of all responses to assessment prompts, the quality control of all data flowing from assessments, the appropriateness of the methods used to combine various types of assessment scores into one composite score and the usefulness and the accuracy of the score reports provided to examinees. (Assessment data quality-control issues could also be discussed as content evidence.)

For evidence of response process for the written comprehensive examination, documentation of all practice materials and written information about the test and instructions to students is important. Documentation of all quality-control procedures used to ensure the absolute accuracy of test scores is also an important source of evidence: the final key validation after a preliminary scoring – to ensure the accuracy of the scoring key and eliminate from final scoring any poorly performing test items; a rationale for any combining rules, such as the combining into one final composite score of MCQ, multiple true–false and short-essay question scores.

Other sources of evidence may include documentation and the rationale for the type of scores reported, the method chosen to report scores and the explanations and interpretive materials provided to explain fully the score report and its meaning, together with any materials discussing the proper use and any common misuses of the assessment score data.

For the SP performance examination, many of the same response process sources may be presented as validity evidence. For a performance examination, documentation demonstrating the accuracy of the SP rating is needed and the results of an SP accuracy study is a particularly important source of response process evidence. Basic quality control of the large amounts of data from an SP performance examination is important to document, together with information on score calculation and reporting methods, their rationale and, particularly, the explanatory materials discussing an appropriate interpretation of the performance-assessment scores (and their limitations).

Documentation of the rationale for using global versus checklist rating scores, for example, may be an important source of response evidence for the SP examination. Or, the empirical evidence and logical rationale for combining a global rating-scale score with checklist item scores to form a composite score may be one very important source of response evidence.

Internal structure

Internal structure, as a source of validity evidence, relates to the statistical or psychometric characteristics of the examination questions or performance prompts, the scale properties – such as reproducibility and generalizability, and the psychometric model used to score and scale the assessment. For instance, scores on test items or sets of items intended to measure the same variable, construct, or content area should be more highly correlated than scores on items intended to measure a different variable, construct, or content area.

Many of the statistical analyses needed to support or refute evidence of the test's internal structure are often carried out as routine quality-control procedures. Analyses such as item analyses – which computes the difficulty (or easiness) of each test question (or performance prompt), the discrimination of each question (a statistical index indicating how well the question separates the high scoring from the low scoring examinees) and a detailed count of the number or proportion of examinees who responded to each option of the test question, are completed. Summary statistics are usually computed, showing the overall difficulty (or easiness) of the total test scale, the average discrimination and the internal consistency reliability of the test.

Reliability is an important aspect of an assessment's validity evidence. Reliability refers to the reproducibility of the scores on the assessment; high score reliability indicates that if the test were to be repeated over time, examinees would receive about the same scores on retesting as they received the first time. Unless assessment scores are reliable and reproducible (as in an experiment) it is nearly impossible to interpret the meaning of those scores – thus, validity evidence is lacking.

There are many different types of reliability, appropriate to various uses of assessment scores. In both example assessments described above, in which the stakes are high and a passing score has been established, the reproducibility of the pass–fail decision is a very important source of validity evidence. That is, analogous to score reliability, if the ultimate outcome of the assessment (passing or failing) can not be reproduced at some high level of certainty, the meaningful interpretation of the test scores is questionable and validity evidence is compromised.

For performance examinations, such as the SP example, a very specialized type of reliability, derived from generalizability theory (GT)[21,22] is an essential component of the internal structure aspect of validity evidence. GT is concerned with how well the specific samples of behaviour (SP cases) can be generalized to the population or universe of behaviours. GT is also a useful tool for estimating the various sources of contributed error in the SP exam, such as error due to the SP raters, error due to the cases (case specificity), and error associated with examinees. As rater error and case specificity are major threats to meaningful interpretation of SP scores, GT analyses are important sources of validity evidence for most performance assessments such as OSCEs, SP exams and clinical performance examinations.

For some assessment applications, in which sophisticated statistical measurement models like Item Response Theory (IRT) models[23,24] the measurement model itself is evidence of the internal structure aspect of construct validity. In IRT applications, which might be used for tests such as the comprehensive written examination example, the factor structure, item-intercorrelation structure and other internal structural characteristics all contribute to validity evidence.

Issues of bias and fairness also pertain to internal test structure and are important sources of validity evidence. All assessments, presented to heterogeneous groups of examinees, have the potential of validity threats from statistical bias. Bias analyses, such as differential item functioning (DIF)[25,26] analyses and the sensitivity review of item and performance prompts are sources of internal structure validity evidence. Documentation of the absence of statistical test bias permits the desired score interpretation and therefore adds to the validity evidence of the assessment.

Relationship to other variables

This familiar source of validity evidence is statistical and correlational. The correlation or relationship of assessment scores to a criterion measure's scores is a typical design for a 'validity study', in which some newer (or simpler or shorter) measure is 'validated' against an existing, older measure with well known characteristics.

This source of validity evidence embodies all the richness and complexity of the contemporary theory of validity in that the relationship to other variables aspect seeks both confirmatory and counter-confirmatory evidence. For example, it may be important to collect correlational validity evidence which shows a strong positive correlation with some other measure of the same achievement or ability and evidence indicating no correlation (or a strong negative correlation) with some other assessment that is hypothesized to be a measure of some completely different achievement or ability.

The concept of convergence and divergence of validity evidence is best exemplified in the classic research design first described by Campbell and Fiske.[27] In this 'multitrait multimethod' design, different measures of the same trait (achievement, ability, performance) are correlated with different measures of the same trait. The resulting pattern of correlation coefficients may show the convergence and divergence of the different assessment methods on measures of the same and different abilities or proficiencies.

In the written comprehensive examination example, it may be important to document the correlation of total and subscale scores with achievement examinations administered during the basic science courses. One could hypothesize that a subscale score for biochemistry on the comprehensive examination would correlate more highly with biochemistry course test scores than with behavioural science course scores. Additionally, the correlation of the written examination scores with the SP final examination may show a low (or no) correlation, indicating that these assessment methods measure some unique achievement, while the correlation of the SP scores with other performance examination scores during the students' clinical training may be high and positive.

As with all research, issues of the generalizability of the results of these studies and the limitations of data interpretation pertain. Interpretation of correlation coefficients, as validity coefficients, may be limited due to the design of the study, systematic bias introduced by missing data from either the test or the criterion or both and statistical issues such as restriction of the range of scores (lack of variance).

Consequences

This aspect of validity evidence may be the most controversial, although it is solidly embodied in the current *Standards*.[1] The consequential aspect of validity refers to the impact on examinees from the assessment scores, decisions and outcomes, and the impact of assessments on teaching and learning. The consequences of assessments on examinees, faculty, patients and society can be great and these consequences can be positive or negative, intended or unintended.

High-stakes examinations abound in North America, especially in medicine and medical education. Extremely high-stakes assessments are often mandated as the final, summative hurdle in professional education. For example, the United States Medical Licensure Examination (USMLE) sequence, sponsored by the National Board of Medical Examiners (NBME), consists of three separate examinations (Steps 1, 2 and 3) which must be passed in order to be licensed as a physician. The consequences of failing any of these examinations is enormous, in that medical education is interrupted in a costly manner or the examinee is not permitted to enter graduate medical education or practice medicine. Likewise, most medical specialty boards in the USA mandate passing a high-stakes certification examination in the specialty or subspecialty, after meeting all eligibility requirements of postgraduate training. The consequences of passing or failing these types of examinations are great, as false positives (passing candidates who should fail) may do harm to patients through the lack of a physician's specialized knowledge or skill and false negatives (failing candidates who should pass) may unjustly harm individual candidates who have invested a great deal of time and resources in graduate medical education.

Thus, consequential validity is one very important aspect of the construct validity argument. Evidence related to consequences of testing and its outcomes is presented to suggest that no harm comes directly from the assessment or, at the very least, more good than harm arises from the assessment. Much of this evidence is more subjective than other sources.

In both example assessments, sources of consequential validity may relate to issues such as passing rates (the proportion who pass), the subjectively judged appropriateness of these passing rates, data comparing the passing rates of each of these examinations to other comprehensive examinations such as the USMLE Step 1 and so on. Evaluations of false positive and false negative outcomes relate to the consequences of these two high-stakes examinations.

The passing score (or grade levels) and the process used to determine the cut scores, the statistical properties of the passing scores, and so on all relate to the consequential aspects of validity.[28] Documentation of the method used to establish a pass–fail score is key consequential evidence, as is the rationale for the selection of a particular passing score method. The psychometric characteristics of the passing score judgements and the qualification and number of expert judges – all may be important to document and present as evidence of consequential validity.

Other psychometric quality indicators concerning the passing score and its consequences (for both example assessments) include a formal, statistical estimation of the pass–fail decision reliability or classification accuracy[29] and some estimation of the standard error of measurement at the cut score.[30]

Equally important consequences of assessment methods on instruction and learning have been discussed by Newble and Jaeger.[31] The methods and strategies selected to evaluate students can have a profound impact on what is taught, how and exactly what students learn, how this learning is used and retained (or not) and how students view and value the educational process.

Threats to validity

The next essay in this series will discuss the many threats to the meaningful interpretation of assessment scores and suggest methods to control these validity threats.

Conclusion

This paper has reviewed 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 collected in order to assign meaningful interpretation to assessment scores or outcomes, which were designed for a specific purpose and at one specific point in time. Validity always refers to score interpretations and never to the assessment itself. The process of validation is closely aligned with the scientific method of theory development, hypothesis generation, data collection for the purpose of hypothesis testing and forming conclusions concerning the accuracy of the desired score interpretations. Validity refers to the impartial, scientific collection of data, from multiple sources, to provide more or less support for the validity hypothesis and relates to logical arguments, based on theory and data, which are formed to assign meaningful interpretations to assessment data.

This paper discussed five typical sources of validity evidence – content, response process, internal structure, relationship to other variables and consequences – in the context of two example assessments in medical education.

Acknowledgements

The author wishes to thank Michael T Kane, PhD, for his critical review of this manuscript.

Funding

There was no external funding for this project.

References

1 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.

2 Crossley J, Humphris G, Jolly B. Assessing health professionals. *Med Educ* 2002;**36**:800–4.

3 Cronbach LJ. Test validation. In: *Educational Measurement*, 2nd edn. Ed: Thorndike RL. Washington, DC: American Council on Education 1971:443–507.

4 Cronbach LJ. Five perspectives on validity argument. In: *Test Validity*. Eds: Wainer H, Braun H. Hillsdale, NJ: Lawrence Erlbaum 1988:3–17.

5 Cronbach LJ. Construct validation after 30 years. In: *Intelligence: Measurement, Theory, and Public Policy*. Ed: Linn RE. Urbana, IL: University of Illinois Press 1989:147–71.

6 Cronbach LJ, Meehl PE. Construct validity in psychological tests. *Psychol Bull* 1955;**52**:281–302.

7 Messick S. The psychology of educational measurement. *J Educ Measure* 1984;**21**:215–37.

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

9 Messick S. Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *Am Psychologist* 1995;**50**:741–9.

10 Messick S. Standards of validity and the validity of standards in performance assessment. *Educ Measure Issues Prac* 1995;**14**:5–8.

11 Kane MT. An argument-based approach to validation. *Psychol Bull* 1992;**112**:527–35.

12 Kane MT. Validating interpretive arguments for licensure and certification examinations. *Evaluation Health Professions* 1994;**17**:133–59.

13 Kane MT. Current concerns in validity theory. *J Educ Measure* 2001;**38**:319–42.

14 Kane MT, Crooks TJ, Cohen AS. Validating measures of performance. *Educ Measure Issues Prac* 1999;**18**:5–17.

15 Cureton EE. Validity. In: *Educational Measurement*. Ed: Lingquist EF. Washington, DC: American Council on Education 1951:621–94.

16 Lohman DF. Teaching and testing to develop fluid abilities. *Educational Reser* 1993;**22**:12–23.

17 Linn RL. Validation of the uses and interpretations of results of state assessment and accountability systems. In: *Large-Scale Assessment Programs for All Students: Development, Implementation, and Analysis*. Eds: Tindal G, Haladyna T. Mahwah, NJ: Lawrence Erlbaum 2002.

18 Loevinger J. Objective tests as instruments of psychological theory. *Psychol Reports, Monograph* 1957;**3** (Suppl.) 635–94.

19 Haladyna TM, Downing SM, Rodriguez MC. A review of multiple-choice item-writing guidelines for classroom assessment. *Appl Measure Educ* 2002;**15**:309–34.

20 Boulet JR, McKinley DW, Whelan GP, Hambelton RK. Quality assurance methods for performance-based assessments. *Adv Health Sci Educ* 2003;**8**:27–47.

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

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

23 Van der Linden WJ, Hambleton RK. Item response theory. Brief history, common models, and extensions. In: *Handbook of Modern Item Response Theory*. Eds: van der Linden WJ, Hambleton RK. New York: Springer-Verlag 1997:1–28.

24 Downing SM. Item response theory: Applications of modern test theory in medical education. *Med Educ* 2003;**37**:1–7.

25 Holland PW, Wainer H, eds. *Differential Item Functioning*. Mahwah, NJ: Lawrence Erlbaum 1993.

26 Penfield RD, Lam RCM. Assessing differential item functioning in performance assessment: review and recommendations. *Educ Measure Issues Prac* 2000;**19**:5–15.

27 Campbell DT, Fiske DW. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psych Bull* 1959;**56**:81–105.

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

29 Subkoviak MJ. A practitioner's guide to computation and interpretation of reliability indices for mastery tests. *J Educ Measure* 1988;**25**:47–55.

30 Angoff WH. Scales, norms, and equivalent scores. In: *Educational Measurement*, 2nd edn. Ed: Thorndike RL. Washington, DC: American Council on Education 1971:508–600.

31 Newble DI, Jaeger K. The effects of assessment and examinations on the learning of medical students. *Med Educ* 1983;**17**:165–71.

Received 29 *May* 2003; *accepted for publication* 3 *June* 2003