Conceptualising and classifying validity evidence for simulation

Pamela B Andreatta & Larry D Gruppen

CONTEXT The term 'validity' is used pervasively in medical education, especially as it relates to curriculum, assessment, measurement and instrumentation. Exactly what is meant by the term 'validity' in the medical education literature is not always clearly defined.

OBJECTIVES This study attempts to clarify, conceptualise and classify how validity fits within the context of assessment and to provide a framework for medical educators to determine the type and degree of validity evidence required for their specific assessment and evaluation needs.

METHODS We apply a structure for considering validity, and its association with validation, in medical education. We build this discussion around the use of simulation in medical training because of its rapid growth as a foundation for numeric measurement of performance in the development of clinical skills and reasoning. We explain why validity is inextricably tied to the assessment process in both simulation-based medical training and traditional medical education.

RESULTS This logical framework structures the type and degree of validity evidence for various assessment and evaluation needs. We also provide an example for medical educators to reference and follow in collecting and reviewing their own needs for validity evidence in all aspects of medical education.

CONCLUSIONS Assessment is integral to measurement and decision making in medical education. The implications of assessment results are variably dependent on the inferences and decisions made from them. As such, validity evidence is critical, but is also flexibly tied to those decisions and not all assessments require the same degree of validity rigor. The framework described herein reinforces a model for medical educators to use in developing their assessment and evaluation needs and associated requirements for validity evidence.

*Medical Education 2009: **43**: 1028–1035*
doi:10.1111/j.1365-2923.2009.03454.x

Department of Medical Education, University of Michigan Medical School, Ann Arbor, Michigan, USA

Correspondence: Pamela B Andreatta, University of Michigan Medical School, 1500 East Medical Center Drive, G1105 Towsley Center, Ann Arbor, Michigan 48109-5201, USA. Tel: 00 1 734 615 0273; Fax: 00 1 734 936 1641; E-mail: pandreat@umich.edu

INTRODUCTION

Governing bodies such as the Accreditation Council for Continuing Medical Education, Residency Review Committees and Continuing Medical Education all want evidence of clinical proficiency and, consequently, there is much discussion about assessment in medical education. Assessment provides evidence that a learner has acquired knowledge and skills within a field of instruction. Progression along the continuum of knowledge and skills may require more complex, rigorous or differing types of assessment evidence. The inevitable issues of validity and reliability arise when we consider how to design both curricula and metrics for performance measures, especially in simulation-based contexts in which the transfer of knowledge and skills to applied clinical contexts is expected.[1–3] In this article, we describe a framework for conceptualising 'validity' and for determining the appropriate level of validity evidence required for assessment in simulation-based training programmes.

VALIDITY AS AN EVIDENCE-BASED ARGUMENT: MAKING THE CASE

There has been a significant transformation in how psychometricians and measurement experts conceptualise validity.[4–9] Validity is fundamentally about decisions made from the interpretation of scores derived from assessment methods, such as simulator metrics or a performance rating scale. It is a function of what you are trying to measure (the construct), how you are trying to measure it (the context and the tool), and how you are using those data to make decisions.

The following example may help clarify this. Suppose a residency programme director decides to assess (judge) the clinical skills of a group of internal medicine interns by having them perform a cerebral angiography using an endovascular simulator with built-in assessment measures derived for senior-level neurosurgery fellows. Such an assessment may demonstrate excellent reliability (consistency of measurement) and indeed address aspects of clinical skills. However, making decisions on the basis of the scores from such assessments would not be considered a valid judgement of the clinical skills of internal medicine interns or even of their ability to perform cerebral angiography because the task is too difficult for this level of learner. By contrast, making decisions on the basis of the results of this assessment tool might very well produce valid (accurate) judgements about the clinical skills of senior neurosurgery fellows because the task is appropriate for their skill level. Thus, validity is not a characteristic of the assessment tool; nor is it a characteristic of the scores derived from the tool. Rather, validity is a characteristic of the *judgements* made on the basis of these scores about a specific *construct*. Thus, *construct* becomes a key issue in any discussion of validity.

CONSTRUCTS AND ASSESSMENT

Every assessment is intended to measure or quantify some underlying construct. A construct is a theoretical entity – something we believe exists and which can be described, but which may not be amenable to direct measurement. Common constructs in medicine include diagnostic reasoning, surgical procedures, cardiopulmonary arrest management, professionalism, teamwork, and the like. These are broad constructs, but constructs may also include specific activities such as suturing skills, in which attributes associated with the construct are transferable across contexts (e.g. intracorporeal, extracorporeal, laparoscopic, arthroscopic, endoscopic, etc.). Because constructs are not simple, concrete objects, a major issue lies in adequately defining them. Constructs such as professionalism have proven to be extraordinarily difficult to define, whereas constructs such as surgical procedures appear much more amenable to definition. Generally, if a construct has a component for which concrete quantitative measures are possible (such as the demonstration of a psychomotor skill) or well-defined indicator behaviours have been demonstrated as highly correlated to the construct (such as selecting the correct diagnosis answer on a case-based test), it is easier to define and therefore easier to assess performance within the construct.

The validity argument for any assessment effort focuses on *how well the assessment reflects the construct*. Unless the construct is clearly defined, it is impossible to provide convincing evidence that the assessment process is appropriately mapping the construct. Without a well-defined construct, you will have problems accruing and interpreting evidence that a learner has achieved associated goals or standards.

Ideally, assessment maps the entire construct with 100% confidence that all relevant aspects of the construct are accurately captured and reflected in the assessment data and all irrelevant information is

excluded by the assessment. For example, the construct for rapid sequence induction (RSI) must include the multiple components that comprise the construct (i.e. procedural knowledge, observance of universal precaution, knowledge of anatomy, knowledge of equipment, clinical reasoning skills, knowledge of medications, knowledge of instruments, knowledge of endotracheal [ET] tube placement, use of follow-up, etc.) so that assessment of the construct can be captured through the assessment of its respective components (see Fig. 1). However, even with a well-defined construct, assessments will seldom, if ever, be a perfect fit with the construct.

One way in which the validity of an assessment may be limited is when the assessment under-represents the construct. For example, if the construct of interest is RSI and the assessment refers to the successful use of a laryngoscope to place an ET tube in a manikin patient simulator, the construct will be under-represented because other aspects of the RSI construct, such as medications, follow-up care and clinical reasoning, are not included in this assessment (Fig. 2). Thus, it is clear that aspects of the construct are not reflected in the assessment results and therefore the validity of decisions made on basis of these results will be limited.

Furthermore, there is a complementary problem in which the assessment captures all the elements of the construct, but includes extraneous elements that are not considered part of the construct, such as, in the case of the RSI assessment, ventilator use, extubation and pain management (Fig. 3). This extraneous information will also limit the validity of judgements based on these results. In this assessment context, the learners are assessed on all aspects of RSI, as well as on pain management, ventilator settings and when to extubate the patient.

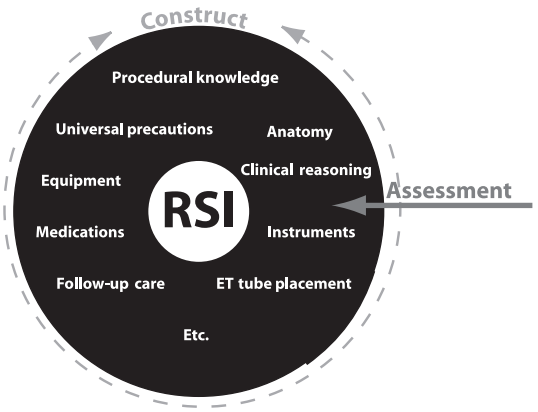

Figure 1 Construct represented perfectly by the assessment

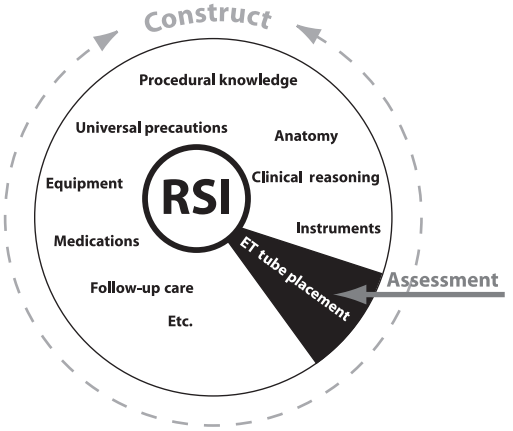

Figure 2 Construct under-representation as a threat to validity

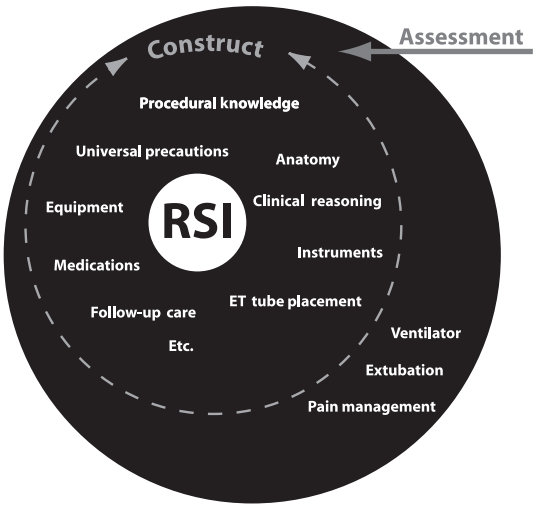

Figure 3 Construct over-representation as a threat to validity

Although these are important knowledge and skill components for the clinician, they are extraneous to this definition of the RSI construct.

In addition, many real-world assessments may miss aspects of the construct (construct under-representation), but include components that are irrelevant to the construct. An assessment for RSI that comprises only ET tube placement, ventilator settings, pain management and extubation is an example (Fig. 4).

The importance of carefully defining the construct emerges as we recognise that the mapping of assessment elements to the construct depends as much on the definition of the construct as it does on the nature and scope of the assessment. Therefore, validation refers to a very specific construct with

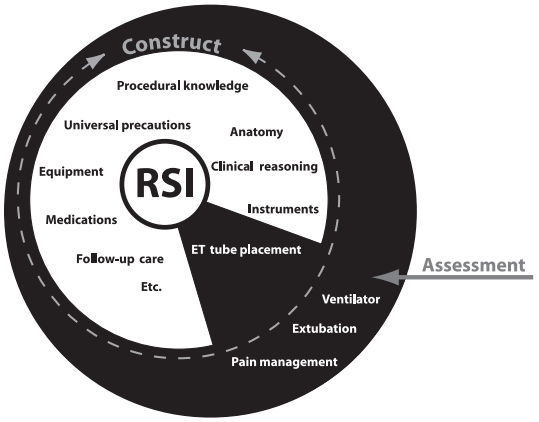

Figure 4 Construct over- and under-representation as threats to assessment validity

limited contextual relevance. For example, validation of assessment measures for learner performance on a specific box trainer exercise (e.g. suturing foam tubing together) may not provide data that support a valid judgement about learner performance on another box trainer exercise (e.g. moving beans between small canisters), or even the same exercises using a different platform (e.g. LapMentor™ virtual reality trainer). This is because assessment necessarily includes the contextual elements of the assessment environment and therefore context is inextricably linked to validity evidence. We shall elaborate this further and delineate how to secure validity evidence for instructional assessment and evaluation purposes.

VALIDITY AS AN EVIDENCE-BASED ARGUMENT

Validity is an evidence-based argument about the trustworthiness of a decision made on the basis of performance data collected in a specific context. As such, it is not 'valid' or 'invalid', but varies by degrees and will be more or less convincing to different audiences. The evidence for validity focuses on how well the assessment data delineate the underlying construct so that the results can be used to make effective decisions about the construct. Evidence that the judgements based on scores derived from an assessment procedure are valid fall into five broad categories: content of the assessment; response processes used in the assessment; the internal structure of the assessment; predicted relationships of assessment scores with other variables, and the consequences of the decisions made on the basis of the assessment data.[4,9] We will describe and provide examples of each type of validity evidence and provide a model from which to determine the extent of evidence required for particular assessment needs.

Evidence from assessment content

A primary source of validity evidence comes from the extent to which the content included in an assessment is relevant to the construct of interest. Consider the construct of laparoscopic surgery and its requisite component of being able to work in three-dimensional space by translating 2-D image representations on a monitor. Earlier, we provided an example of completing a sequence of task activities on a box trainer to assess this skill. Critical content for assessment of 2-D or 3-D translational ability includes information about the instruments, monitor, 3-D working space, and task activities that accurately reflect the laparoscopic surgical construct. An assessment that requires the learner to simply move a laparoscopic instrument within 3-D space without specific performance task requirements would underrepresent the construct, whereas including an assessment of the precision of sutures and knot integrity on a box trainer suturing task would over-represent the intended target.

Experts in the field who are knowledgeable about how the target construct is expressed in its relevant contexts typically provide content validity evidence. In the traditional validity framework, this was often referred to as 'face validity'. In addition to expert opinion, content validity evidence may be derived from a formal job or task analysis, a curriculum analysis, or scientific inquiry into the nature of the field.

Evidence from the response processes of an assessment

Not only must the *content* of an assessment represent the construct, but the cognitive and physical *processes* required by the assessment must also represent the construct. The common complaint that multiple-choice tests are poor (low validity) indicators of clinical skill is based in considerable part on the fact that selecting the best answer from five options presented on paper in the context of a brief, focused question involves a very different cognitive process from that used by a resident in a clinical setting with a real patient. The rapid growth in the use of simulation for clinical training and assessment is primarily driven by evidence that the processes used in simulated contexts are closely aligned with those used in applied clinical practice.

Like content evidence for an assessment, process validity evidence is often provided by the judgements of experts in the field. Process validity evidence may

also come from empirical results that show consistencies in the performance of tasks employing similar processes, or contrasts between the performances of tasks employing different processes from the underlying construct. Other sources of process validity evidence may be derived from asking examinees about how they responded to the task, or by analysing their judgement processes.

Evidence from the internal structure of the assessment

Assessment content and processes provide data about learner performance relevant to the construct, but these data typically need to be transformed into a score before any decisions can be based on them. The third source of validity evidence focuses on how this transformation is made and the match between a score and the structure of the underlying construct. For example, a computer-based simulator used to assess skills in laparoscopic surgery generates a performance score that is a composite of scores given for time, accuracy, efficiency of movement and tissue damage caused. The score assumes that all factors are of equal weight; that is, efficiency of movement and time are valued equally with tissue damage and accuracy. However, if the construct being assessed is competence in performing a laparoscopic surgical task that requires great delicacy and precision, tissue damage may be considered a more important factor than time. The important thing to consider is whether or not the score generated from the data accurately reflects the desired valuation embedded in the construct and therefore provides an accurate assessment of performance.

Internal structure validity evidence for an assessment is provided by explicit scoring algorithms and criteria and by clear explanations for how these algorithms relate to the underlying construct. Evidence may also be provided by the judgements of field experts about how components of the assessment data should be combined.

Evidence from relationships with other variables

A key source of evidence for the validity of assessment-based decisions is represented by the relationships between assessment results and other data and variables with predicted associations to the construct. These predicted associations can take many forms. One prediction would be that scores from an assessment are stable across settings, tasks and groups. This is evidence that the assessment is consistent in representing the construct and is often conveyed by reliability statistics (inter-rater reliability, internal consistency, test–retest reliability). Another predicted relationship might be that assessment scores increase with increasing experience or training in the examinees (i.e. an assessment differentiates among expert, intermediate and novice learners).

The predictions that underlie this evidence of validity depend greatly on the underlying theoretical construct and how it relates to other constructs. For example, we would probably predict that scores from a simulation-based assessment of central venous catheterisation indicating that the learners met the criterion standard of performance would be associated with a reduction in catheter-related bloodstream infections and complications related to catheter placement.

Another example would be that a simulation-based assessment of laparoscopic skills would have a positive correlation with ratings of actual performance in the operating room (OR) because both assessments are, presumably, assessing the same skill. The extent that these predicted correlations are observed provides supportive evidence for using the results of this assessment to decide which residents are ready to proceed to the OR and which should practise more. If this predicted relationship is not observed, we must then examine possible explanations, which could include the hypothesis that ratings by teaching staff are measuring something other than laparoscopic skills.

By contrast, ratings by teaching staff of a resident's laparoscopic skill in a simulation laboratory would probably be *unrelated* to ratings of the same resident's professionalism because these are considered different constructs. If the correlations between these ratings are indeed low, they can be construed as supportive evidence for the validity results of the ratings by teaching staff. If, however, they are highly correlated, they might be seen as evidence that these results are not, perhaps, measuring laparoscopic skills as originally thought.

Evidence from the consequences of assessment

The final source of validity evidence for decisions made on the basis of assessment results is, essentially, that the decisions 'work'. This evidence comes from monitoring the outcomes of decisions made on the basis of the scores – successes or errors – and evaluating the intended and unintended consequences of interpreting and using assessment scores. Consequential validity evidence includes issues of bias, fairness and justice, and relates to the consequences of actions taken as a result of the

assessment outcomes. That is, it considers the implications of decisions made on the basis of assessments that either over- or under-estimate actual competence in the target construct.

As an example, consider the consequences of using a series of assessments of laparoscopic skill that utilise varied assessment methods. The results of these assessments are combined and compared with a standard for defining 'competent'. Competent residents are then given more OR responsibility (the decision that is based on the assessment). If the result were a sudden increase in the number of adverse events or procedural errors on the part of the residents, it would be logical to conclude that the consequences of these decisions were undesirable and, thus, that the assessment and scoring process had not led to good (valid) decisions.

Similarly, the use of this assessment for deciding on OR responsibilities may have the intended consequence of motivating learners to practise specific techniques or the unintended consequence of leading them to ignore other, important learning activities. Each of these could be used as evidence for and against the validity of the decisions made on the basis of the scores on this assessment.

IMPLICATIONS FOR SIMULATION-BASED MEDICAL EDUCATION

As we have illustrated, validity is contextual. Because validity is not a stable attribute of a given assessment, but depends on how those assessment scores are collected, the characteristics of the examinees and how the scores are to be used, it should be clear that judgements about validity are limited to each context and application. Therefore, to say an assessment instrument is 'valid' or has been 'validated' is strictly true only for the context of the learners, the performance context, the content domain, and the stringency of decisions made on those assessment data.

Obviously, the closer one context is to another, the more readily we can argue that the validity evidence from the first context might apply to the second. Conversely, the further we take an assessment from its original implementation (e.g. taking an assessment developed for Year 1 medical students in the Netherlands and applying it to senior surgical residents in the USA), the more we should be concerned with documenting the validity of the results of the assessment for the intended purpose. This implication highlights the fallacy of describing an assessment as 'validated' as if it were appropriate for all times, in all places, for all audiences, and for all purposes.

The term 'validity' in this sense is semantically quite different from the term as used in standard English, when it refers to the effectiveness or feasibility of something. An instrument that is valid or has been validated signifies that it yields concrete evidence of sufficient discriminatory power from which to confidently make decisions about performance in the construct. Therefore, the assessment instrument itself is inextricably tied to the contextual aspects through which validity evidence has been collected.

It is important to recognise that not all forms of validity evidence are required for an assessment to have value or to be an acceptable measure of performance. From a research perspective, to say something is valid means that the data generated by the assessment describe the learner's performance with sufficient accuracy to support confidence in the decisions that will be made about the learner from it. The implications for error in the making of decisions determine the limits imposed on validity evidence. That is, high-stakes, summative and generalisable decisions require significantly more validity evidence than might be necessary for low-stakes, formative or narrowly applied decisions. For example, at all levels of medical education the implications for decisions made from performance data tend to occur at the institutional level rather than a regulatory level. Aside from agency-based medical licensing and specialty board examinations, the decisions made on the basis of learner assessments at the institutional level are generally not high stakes beyond meeting the base requirements for graduation from medical school. Therefore, the validity evidence for most training-based assessments that include content, process, internal structure and relational evidence are likely to be sufficient. High-stakes assessments, such as licensing and board examinations, require rigorous validity evidence including content, process, internal structure, relational and consequential evidence because the accuracy with which performance is quantified must be much more defensible. (See Fig. 5 for a model illustrating the relationships between assessment and validity for performance across the continuum of construct competence.)

DECISIONS AT DIFFERENT PHASES OF TRAINING REQUIRE DIFFERENT LEVELS OF VALIDITY EVIDENCE

As learners move through the acquisition of knowledge, skills and affective elements of a construct, they

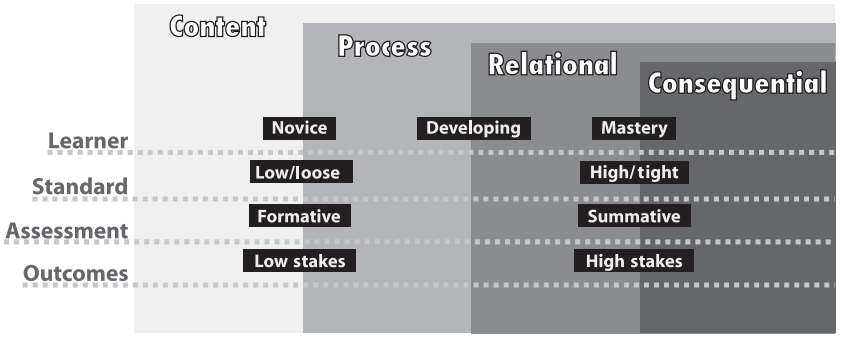

Figure 5 Validity evidence for assessments across the learning continuum

require assessment at multiple points along the continuum of instruction. These points of assessment, and the type of assessment, will vary depending on the curriculum itself, as well as on the learner's maturity within the field of instruction. In the early phases of learning, assessment is more formative and focused on providing feedback to the learner to guide his or her development in the construct. For example, an assessment comprised of a faculty member rating the quality of a resident's sutures by pulling on them to test their integrity would provide formative information to the learner to modify his or her perceptions of how well he or she had performed. Content and process validity evidence is acceptable for this level of assessment.

Moving through the continuum of learning, expertise grows until no new knowledge or skills are being learned, but, rather, the objective is mastery of the construct through practice. Assessment at this phase of instruction should provide evidence of content, process, internal structure and relational validity because it will be used to guide the learner's progress towards specific performance objectives. A faculty member's of the quality of a resident's sutures that notes deviations in their spacing from a model of acceptable performance would provide formative information to the learner to modify his or her performance accordingly.

The last phase of learning leads to demonstration that mastery has been achieved. Assessment at this phase on the continuum is summative and performance outcomes determine whether or not the learner has achieved the requisite standards in the construct. Because this type of assessment is typically considered high stakes, content, process, internal structure, relational and consequential validity evidence should accurately predict performance in the applied environment. Continuing with our suturing examples, the learner might complete a suturing exercise modelled to reflect the applied clinical context that is assessed by faculty members for multiple suture qualities (e.g. integrity, tension, placement, consistency) against an expert model, as well as whether or not the sutures held under imposed stressors or over an appropriate time period.

Each assessment point across the continuum requires an assessment instrument targeted to the learner's level of expertise in the construct and the expected performance outcomes at that point of the continuum. The model presented in Fig. 5 provides a template for identifying the type of assessment required for any given performance step along the continuum and can help in selecting or constructing an instrument with appropriate validity evidence. By specifying the level of the learner's expertise and expected standards of performance, the type of assessment and requisite validity evidence will be made clear.

In conclusion, a summary of the points to be considered when conceptualising and classifying

Table 1 Summary of points

Curricula cannot be 'valid' or 'validated'. Only assessment can have validity evidence
Validity evidence delineates how well an assessment maps to the underlying construct
Watch out for
Over-representation of the construct
Under-representation of the construct
Missing the construct completely
Validity evidence is contextual and may not be transferable
The decisions that will be made from assessment data determine the rigor of required validity evidence
High stakes decisions require more rigorous validity evidence than low stakes
Summative assessments require more rigorous validity evidence than formative assessments

validity evidence for simulation-based medical education is presented in Table 1.

Contributors: both authors contributed to the conception, design, drafting and revising of this paper and both approved the final manuscript for publication.
Acknowledgements: none.
Funding: none.
Conflicts of interest: none.
Ethical approval: not applicable.

REFERENCES

1 Peters JH, Fried GM, Swanstrom LL, Soper NJ, Sillin LF, Schirmer B, Hoffman K, SAGES FLS Committee. Development and validation of a comprehensive programme of education and assessment of the basic fundamentals of laparoscopic surgery. *Surgery* 2004;**135**: 21–7.
2 Dauster B, Steinberg AP, Vassiliou MC, Bergman S, Stanbridge DD, Feldman LS, Fried GM. Validity of the MISTELS simulator for laparoscopy training in urology. *J Endourol* 2005;**19**:541–5.
3 Scott DJ, Ritter EM, Tesfay ST, Pimental EA, Nagji L, Fried GM. Certification pass rate of 100% for fundamentals of laparoscopic surgery skills after proficiency-based training. *Surg Endosc* 2008;**22** (8):1887–93.
4 Downing SM. Validity: on the meaningful interpretation of assessment data. *Med Educ* 2003;**37**:830–7.
5 Downing SM, Haladyna TM. Validity threats: overcoming interference with proposed interpretations of assessment data. *Med Educ* 2004;**38**:327–33.
6 Messick S. Validity. In: Linn RL, ed. *Educational Measurement*, 3rd edn, New York, NY: American Council on Education, Macmillan 1989;13–103.
7 Messick S. Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *Am Psychol* 1995;**50**:741–9.
8 Kane MT. The assessment of professional competence. *Eval Health Prof* 1992;**15**:163–82.
9 American Educational Research Association/American Psychological Association/National Council on Measurement in Education (AERA/APA/NCME). *Standards for Educational and Psychological Testing*. Washington, DC: AERA 2000.

Received 19 March 2009; editorial comments to authors 25 May 2009; accepted for publication 26 June 2009