Validity threats: overcoming interference with proposed interpretations of assessment data

Steven M Downing 1 & Thomas M Haladyna 2

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

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

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

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

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

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

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

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

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

Correspondence: Steven M Downing PhD, Associate Professor of Medical Education, University of Illinois at Chicago, College of Medicine, Department of Medical Education (MC 591), 808 South Wood Street, Chicago, Illinois 60612-7309, USA. Tel: 00 1 312 996 6428; Fax: 00 1 312 413 2048; E-mail: sdowning@uic.edu
noted 2 major sources of validity threats: construct under-representation (CU) and construct-irrelevant variance (CIV). Construct under-representation refers to the undersampling or biased sampling of the content domain by the assessment instrument. Construct-irrelevant variance refers to systematic error (rather than random error) introduced into the assessment data by variables unrelated to the construct being measured. Both CU and CIV threaten validity evidence by reducing the ability to reasonably interpret assessment results in the proposed manner.

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

**Written examinations**

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

---

**Table 1 Threats to validity of assessments**

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

Key learning points

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

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

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

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

*Face validity* is not any type of true, scientific evidence of validity.
specifications well, so that some content areas are oversampled while others are undersampled; use of many items that test only low level cognitive behaviour, such as recall or recognition of facts, while the instructional objectives require higher level cognitive behaviour, such as application or problem solving; and, use of items that test trivial content that is unrelated to future learning. 5

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

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

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

What are the solutions to these types of CIV problems? On written achievement tests, items should be well crafted and follow the basic evidence-based principles of effective item writing. 9,11,12 The item format itself should not be an impediment to student assessment. The reading ability of students should not be a major factor in the assessment of the achievement construct. Most items should be targeted in difficulty to student achievement levels. All items which are empirically shown to be biased or which use language that might offend some cultural, racial or ethnic group should be eliminated from the test. Test items must be maintained securely and tests should be administered in proctored, controlled environments so that any potential cheating is minimised or eliminated. Instructors should not teach directly to the content of the test. Instead, teaching should be targeted at the content domain of which the test is a small sample. Finally, passing scores (or grading standards) should be established in a defensible manner, which is fair to all students.

Performance examinations

Objective structured clinical examinations or SP examinations increase the fidelity of the assessment and are intended to measure performance, rather than knowledge or skills. 13 Performance assessments
are closer in proximity to the actual criterion performance of interest, but these types of assessment also involve constructs, because they sample performance behaviour in a standardised, ideal environment. They are simulations of the real world, but are not the real world. The performance of students, rated by trained SPs in a controlled environment on a finite number of selected cases requiring maximum performance, is not actual performance in the real world; rather, inferences must be made from performance ratings to the domain of performance, with a specific interpretation or meaning attributed to the checklist or rating scale data. Validity evidence must be documented to support or refute the proposed meaning associated with these performance-type constructs.

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

Many SP examinations use trained lay SPs to portray actual patient medical problems and to rate student performance after the encounter, using standardised checklists or rating scales. The quality of the SP portrayal is extremely important, as is the quality of the SPs’ training in the appropriate use of checklists and rating scales. If the SPs are not sufficiently well trained to consistently portray the patient in a standardised manner, different students effectively encounter different patients and slightly different patient problems. The construct of interest is therefore misrepresented, because all students do not encounter the same patient problem or stimulus. Remedies for CU in SP examinations include the use of large numbers of representative cases, using well trained SP raters. Standardised patient monitoring, during multiday performance examinations, is critical, so that any slippage in the standard portrayal can be corrected during the time of the examination.

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

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

The remedies for CU in performance examinations are obvious, but may be difficult and costly to implement. Low reliability is a major threat to validity,9 thus using sufficient numbers of reliable and representative cases to adequately generalise to the proposed domain is critical. Generalisability must be estimated for most performance-type examinations, using generalisability theory.17,18 For high-stakes performance examinations, generalisability coefficients should be at least 0.80; the phi-coefficient is the appropriate estimate of generalisability for criterion-referenced performance examinations (which have absolute, rather than relative passing scores).16 Standardised patients should be well trained in their patient roles and their portrayals monitored throughout the time period of the examination to ensure standardisation. To control or eliminate CIV in performance examinations, checklists and rating scales must be well developed, critiqued, edited and tried out and be sufficiently accurate to provide reproducible data when completed by SPs who are well trained in their use. Methods to detect statistical bias in performance examination ratings should be implemented for high-stakes examinations.19 Performance cases should be
pretested with a representative group of students prior to their final use, testing the appropriateness of case difficulty and all other aspects of the case presentation. Standardised patient training is critical, in order to eliminate sources of CIV introduced by variables such as SP rater bias and student success at bluffing the SP. If passing scores or grades are assigned to the performance examination results, these scores must be established in a defensible, systematic, reproducible and fair manner.

Ratings of clinical performance

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

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

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

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

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

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

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

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

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

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

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

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

CONCLUSIONS

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

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

CONTRIBUTORS

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

FUNDING

There was no external funding for this project.

REFERENCES


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


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