The American Journal of Medicine (2006) 119, 166.e7-166.e16

ELSEVIER

THE AMERICAN JOURNAL of MEDICINE®

REVIEW

Current Concepts in Validity and Reliability for Psychometric Instruments: Theory and Application

David A. Cook, MD, MHPE, Thomas J. Beckman, MD, FACP

Division of General Internal Medicine, Mayo Clinic College of Medicine, Rochester, Minn.

ABSTRACT

Validity and reliability relate to the interpretation of scores from psychometric instruments (eg, symptom scales, questionnaires, education tests, and observer ratings) used in clinical practice, research, education, and administration. Emerging paradigms replace prior distinctions of face, content, and criterion validity with the unitary concept "construct validity," the degree to which a score can be interpreted as representing the intended underlying construct. Evidence to support the validity argument is collected from 5 sources:

- Content: do instrument items completely represent the construct?
- Response process: the relationship between the intended construct and the thought processes of subjects or observers
- Internal structure: acceptable reliability and factor structure
- Relations to other variables: correlation with scores from another instrument assessing the same construct
- Consequences: do scores really make a difference?

Evidence should be sought from a variety of sources to support a given interpretation. Reliable scores are necessary, but not sufficient, for valid interpretation. Increased attention to the systematic collection of validity evidence for scores from psychometric instruments will improve assessments in research, patient care, and education. © 2006 Elsevier Inc. All rights reserved.

KEYWORDS: Construct validity; Reproducibility of results; Educational measurement; Medical education; Quality of life; Questionnaire

Physicians must be skilled in assessing the quality of outcomes reported in the literature and obtained from instruments in clinical practice. Frequently these outcomes are assessed using instruments such as scales, questionnaires, education tests, and observer ratings that attempt to measure factors such as symptoms, attitudes, knowledge, or skills in various settings of medical practice (Table 1).[1-9] For the purposes of this article, we will refer to all such instruments as psychometric. The term "validity" refers to the degree to which the conclusions (interpretations) derived from the results of any assessment are "well-grounded or justifiable; being at once relevant and meaningful."[10] However, the skills required to assess the validity of results from psychometric assessments are different than the skills used in appraising the medical literature[11] or interpreting the results of laboratory tests.[12] In a recent review of clinical teaching assessment, we found that validity and reliability were frequently misunderstood and misapplied.[13] We also have noted that research studies with sound methods often fail to present a broad spectrum of validity evidence supporting the primary outcome.[6,14-16] Thus, we recognized a need for further discussion of validity in the context of psychometric instruments and how this relates to clinical research and practice.

Methods for evaluating the validity of results from psychometric assessments derive from theories of psychology and educational assessment,[17,18] and there is extensive literature in these disciplines. However, we are not aware of

Requests for reprints should be addressed to: David A. Cook, MD, MHPE, Baldwin 4-A, Division of General Internal Medicine, Mayo Clinic College of Medicine, 200 First Street SW, Rochester, MN 55905.

E-mail address: cook.david33@mayo.edu.

0002-9343/$ -see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.amjmed.2005.10.036

recent reviews for physicians. Furthermore, within the psychologic literature there is variation in terminology and practices. In an attempt to establish a unified approach to validity, the American Psychological Association published standards that integrate emerging concepts.[19] These standards readily translate to medical practice and research and provide a comprehensive approach for assessing the validity of results derived from psychometric instruments. This article will discuss this model and its application to clinical medicine, research, and education. Reliability, a necessary element of validity, will also be discussed within this framework.

CLINICAL SIGNIFICANCE

- Best clinical, research, and educational practice requires sound assessment methods. This article presents an innovative framework for evaluating the validity of scores from instruments such as symptom scales, questionnaires, education tests, and observer ratings.
- Validity is viewed as a carefully structured argument assembling evidence from a variety of sources to support or refute proposed interpretations of instrument scores.
- A thorough understanding of this framework will transform how physicians approach validity.

VALIDITY, CONSTRUCTS, AND MEANINGFUL INTERPRETATION OF INSTRUMENT SCORES

Validity refers to "the degree to which evidence and theory support the interpretations of test scores entailed by the proposed uses of tests."[19] In other words, validity describes how well one can legitimately trust the results of a test as interpreted for a specific purpose.

Many instruments measure a physical quantity such as height, blood pressure, or serum sodium. Interpreting the meaning of such results is straightforward.[20] In contrast, results from assessments of patient symptoms, student knowledge, or physician attitudes have no inherent meaning. Rather, they attempt to measure an underlying construct, an "intangible collection of abstract concepts and principles."[21] The results of any psychometric assessment have meaning (validity) only in the context of the construct they purport to assess.[17] Table 2 lists constructs (inferences) for selected instruments.[3,5,8,22] Because the validity of an instrument's scores hinges on the construct, a clear definition of the intended construct is the first step in any validity evaluation. Note that many of the constructs listed in Table 2 would benefit from more precision and clarity.

Validity is not a property of the instrument, but of the instrument's scores and their interpretations.[17,19] For example, an instrument originally developed for depression screening might be legitimately considered for assessing anxiety. In contrast, we would expect cardiology board examination scores to accurately assess the construct "knowledge of cardiology," but not "knowledge of pulmonary medicine" or "procedural skill in coronary angiography." Note that the instruments in these examples did not change—only the score interpretations.

Because validity is a property of inferences, not instruments, validity must be established for each intended interpretation. In the example above, the depression instrument's scores would require further study before use in assessing anxiety. Similarly, a patient symptom scale whose scores provided valid inferences under research study conditions or in highly selected patients may need further evaluation before use in a typical clinical practice.

Table 1 Examples of psychometric instruments used in medical practice

Medical setting	Type of instrument	Specific examples
Clinical practice	Symptom or disease severity scale	AUA-SI symptom score for BPH[1]
	Screening tool	CAGE screen for alcoholism,[2] PRIME-MD[3] screen for depression
Research	Symptom or disease severity scale	AUA-SI,[1] KCCQ[4]
	Quality of life inventory	LCSS[5]
	Questionnaire (survey)	Survey of teens regarding tobacco use[6]
Education	Written examination	USMLE Step 1,[7] locally developed multiple-choice exam
	Objective structured clinical examination or standardized patient examination	USMLE Step 2 Clinical Skills,[7] locally developed test of interviewing skill
	Learner or teacher assessment	Mini-CEX,[8] SFDP-26[9]
	Course evaluation	Locally developed evaluation form
Administration	Questionnaire (survey)	Staff or patient satisfaction survey

AUA-SI = American Urological Association Symptom Index; PRIME-MD = Primary Care Evaluation of Mental Disorders; USMLE = United States Medical Licensing Exam; Mini-CEX = Mini-clinical evaluation exercise; SFDP-26 = Stanford Faculty Development Program Questionnaire; KCCQ = Kansas City Cardiomyopathy Questionnaire; LCSS = Lung Cancer Symptom Scale; BPH = benign prostatic hypertrophy.

Table 2 Potential inferences and sources of validity evidence for scores from selected psychometric instruments

			Potential sources of information for each validity evidence category				
Instrument type	Sample instrument	Intended inference from scores*	Content	Response process	Internal structure	Relations to other variables	Consequences
Multiple-choice exam	Internal medicine certifying exam	"Competence in the diagnosis and treatment of common conditions . . . and excellence in the broad domain of internal medicine"[22]	Test blueprint; qualifications of question writers; well-written questions	Clarity of instructions; student thought process as he or she answers the questions; test security and scoring	Internal consistency; item discrimination	Correlation with clinical rotation grades, scores on other tests, or long-term follow-up of patient outcomes	Method of determining exam pass/fail score; differential pass/fail rates among examinees expected to perform similarly
Clinical performance evaluation	Mini-CEX	"Clinical competence of candidates for certification"[8]	Test blueprint; qualifications of question writers; well-written questions	Rater training; rater thought process as he or she observes performer; test scoring	Inter-rater reliability; factor analysis to identify distinct dimensions of clinical performance	Correlation with scores on other performance assessments	Method of determining pass/fail score; differential pass/fail rates among examinees expected to perform similarly
Patient assessment	PRIME-MD	This patient has one or more "of 18 possible current mental disorders"[3]	Qualifications of question writers; well-written questions; evidence that questions adequately represent domain	Language barrier; patient thought process as he or she answers the questions	Test-retest reliability; internal consistency	Correlation with clinically diagnosed depression; scores from other depression assessments, or health care use	Method of determining score thresholds; improvement in patient outcomes after implementation of this instrument
Questionnaire	Lung Cancer Symptom Scale	"Physical and functional dimensions of quality of life"[5]	Well-written questions; evidence that questions adequately represent domain	Language barrier; patient thought process as he or she answers the questions	Internal consistency; factor analysis	Correlation with an objective assessment of quality of life, eg, hospitalization	Improvement in patient outcomes after implementation of this instrument

Mini-CEX = Mini-clinical evaluation exercise; PRIME-MD = Primary Care Evaluation of Mental Disorders.

*Intended inference as represented by instrument authors in cited publication.

A Conceptual Approach to Validity

We often read about "validated instruments." This conceptualization implies a dichotomy—either the instrument is valid or it is not. This view is inaccurate. First, we must remember that validity is a property of the inference, not the instrument. Second, the validity of interpretations is always a matter of degree. An instrument's scores will reflect the underlying construct more accurately or less accurately but never perfectly.

Validity is best viewed as a hypothesis or "interpretive argument" for which evidence is collected in support of proposed inferences.[17,23,24] As Downing states, "Validity requires an evidentiary chain which clearly links the interpretation of . . . scores . . . to a network of theory, hypotheses, and logic which are presented to support or refute the reasonableness of the desired interpretations."[21] As with any hypothesis-driven research, the hypothesis is clearly stated, evidence is collected to evaluate the most problematic assumptions, and the hypothesis is critically reviewed, leading to a new cycle of tests and evidence "until all inferences in the interpretive argument are plausible, or the interpretive argument is rejected."[25] However, validity can never be proven.

Validity has traditionally been separated into 3 distinct types, namely, content, criterion, and construct validity.[26] However, contemporary thinking on the subject suggests that these distinctions are arbitrary[17,19] and that all validity should be conceptualized under one overarching framework, "construct validity." This approach underscores the reasoning that an instrument's scores are only useful inasmuch as they reflect a construct and that evidence should be collected to support this relationship. The distinct concepts of content and criterion validity are preserved as sources of validity evidence within the construct validity rubric, as discussed below.

Sources of Validity Evidence

Messick[17] identifies 5 sources of evidence to support construct validity: content, response process, internal structure, relations to other variables, and consequences. These are not different types of validity but rather they are categories of evidence that can be collected to support the construct validity of inferences made from instrument scores. Evidence should be sought from several different sources to support any given interpretation, and strong evidence from one source does not obviate the need to seek evidence from other sources. While accruing evidence, one should specifically consider two threats to validity: inadequate sampling of the content domain (construct underrepresentation) and factors exerting nonrandom influence on scores (bias, or construct-irrelevant variance).[24,27] The sources of validity evidence are discussed below, and examples are provided in Table 2.

Content. Content evidence involves evaluating the "relationship between a test's content and the construct it is intended to measure."[19] The content should represent the truth (construct), the whole truth (construct), and nothing but the truth (construct). Thus, we look at the construct definition, the instrument's intended purpose, the process for developing and selecting items (the individual questions, prompts, or cases comprising the instrument), the wording of individual items, and the qualifications of item writers and reviewers. Content evidence is often presented as a detailed description of steps taken to ensure that the items represent the construct.[28]

Response Process. Reviewing the actions and thought processes of test takers or observers (response process) can illuminate the "fit between the construct and the detailed nature of performance . . . actually engaged in."[19] For example, educators might ask, "Do students taking a test intended to assess diagnostic reasoning actually invoke higher-order thinking processes?" They could approach this problem by asking a group of students to "think aloud" as they answer questions. If an instrument requires one person to rate the performance of another, evidence supporting response process might show that raters have been properly trained. Data security and methods for scoring and reporting results also constitute evidence for this category.[21]

Internal Structure. Reliability[29,30] (discussed below and in Table 3) and factor analysis[31,32] data are generally considered evidence of internal structure.[21,31] Scores intended to measure a single construct should yield homogenous results, whereas scores intended to measure multiple constructs should demonstrate heterogenous responses in a pattern predicted by the constructs. Furthermore, systematic variation in responses to specific items among subgroups who were expected to perform similarly (termed "differential item functioning") suggests a flaw in internal structure, whereas confirmation of predicted differences provides supporting evidence in this category.[19] For example, if Hispanics consistently answer a question one way and Caucasians answer another way, regardless of other responses, this will weaken (or support, if this was expected) the validity of intended interpretations. This contrasts with subgroup variations in total score, which reflect relations to other variables as discussed next.

Relations to Other Variables. Correlation with scores from another instrument or outcome for which correlation would be expected, or lack of correlation where it would not, supports interpretation consistent with the underlying construct.[18,33] For example, correlation between scores from a questionnaire designed to assess the severity of benign prostatic hypertrophy and the incidence of acute urinary retention would support the validity of the intended inferences. For a quality of life assessment, score differences among patients with varying health states would support validity.

Consequences. Evaluating intended or unintended consequences of an assessment can reveal previously unnoticed

Table 3 Different ways to assess reliability*

Source of reliability	Description	Measures	Definitions	Comments
Internal consistency	Do all the items on an instrument measure the same construct? (If an instrument measures more than one construct, a single score will not measure either construct very well. We would expect high correlation between item scores measuring a single construct.) Note: Internal consistency is probably the most commonly reported reliability statistic, in part because it can be calculated after a single administration of a single instrument. Because instrument halves can be considered "alternate forms," internal consistency can be viewed as an estimate of parallel forms reliability.	Split-half reliability	Correlation between scores on the first and second halves of a given instrument	Rarely used in practice because the "effective" instrument is only half as long as the actual instrument; the Spearman-Brown† formula can adjust this result
		Kuder-Richardson	Similar concept to split-half, but accounts for all items	Assumes all items are equivalent, measure a single construct, and have dichotomous responses
		Cronbach's alpha	A generalized form of the Kuder-Richardson formulas	Assumes all items are equivalent and measure a single construct; can be used with dichotomous or continuous data
Temporal stability	Does the instrument produce similar results when administered a second time?	Test-retest reliability	Administer the instrument to the same person at different times	Usually quantified using correlation (eg, Pearson's r)
Parallel forms	Do different versions of the "same" instrument produce similar results?	Alternate forms reliability	Administer different versions of the instrument to the same individual at the same or different times	Usually quantified using correlation (eg, Pearson's r)
Agreement (inter-rater reliability)	When using raters, does it matter who does the rating? Is one rater's score similar to another's?	Percent agreement	Percent of identical responses	Does not account for agreement that would occur by chance
		Phi	Simple correlation	Does not account for chance
		Kappa	Agreement corrected for chance	
		Kendall's tau	Agreement on ranked data	
		Intraclass correlation coefficient	Uses analysis of variance to estimate how well ratings from different raters coincide	
Generalizability theory	How much of the error in measurement is the result of each factor (eg, item, item grouping, subject, rater, day of administration) involved in the measurement process?	Generalizability coefficient	Complex model that allows estimation of multiple sources of error	As the name implies, this elegant method is "generalizable" to virtually any setting in which reliability is assessed; for example, it can determine the relative contribution of internal consistency and inter-rater reliability to the overall reliability of a given instrument

For more details regarding the concepts in this table, please see references.[30,37-41]

This table adapted from Beckman TJ, Ghosh AK, Cook DA, Erwin PJ, Mandrekar JN. How reliable are assessments of clinical teaching? A review of the published instruments. *J Gen Intern Med*. 2004;19:971; used with permission from Blackwell Publishing.

*"Items" are the individual questions on the instrument. The "construct" is what is being measured, such as knowledge, attitude, skill, or symptom in a specific area.

†The Spearman Brown "prophecy" formula allows one to calculate the reliability of an instrument's scores when the number of items is increased (or decreased).

sources of invalidity. For example, if a teaching assessment shows that male instructors are consistently rated lower than females it could represent a source of unexpected bias. It could also mean that males are less effective teachers. Evidence of consequences thus requires a link relating the observations back to the original construct before it can truly be said to influence the validity of inferences. Another way to assess evidence of consequences is to explore whether desired results have been achieved and unintended effects avoided. In the example just cited, if highly rated faculty ostracized those with lower scores, this unexpected negative outcome would certainly affect the meaning of the scores and thus their validity.[17] On the other hand, if remediation of faculty with lower scores led to improved performance, it would support the validity of these interpretations. Finally, the method used to determine score thresholds (eg, pass/fail cut scores or classification of symptom severity as low, moderate, or high) also falls under this category.[21] Evidence of consequences is the most controversial category of validity evidence and was the least reported evidence source in our recent review of instruments used to assess clinical teaching.[34]

Integrating the Evidence. The words “intended” and “predicted” are used frequently in the above paragraphs. Each line of evidence relates back to the underlying (theoretical) construct and will be most powerful when used to confirm relationships stated a priori.[17,25] If evidence does not support the original validity argument, the argument “may be rejected, or it may be improved by adjusting the interpretation and/or the measurement procedure”[25] after which the argument must be evaluated anew. Indeed, validity evaluation is an ongoing cycle of testing and revision.[17,31,35]

The amount of evidence necessary will vary according to the proposed uses of the instrument. Circumstances requiring a high degree of confidence in the accuracy of interpretations (eg, high-stakes board certification or the primary outcome in a research study) will mandate more evidence than settings where a lower degree of confidence is acceptable. Some instrument types will rely more heavily on certain categories of validity evidence than others.[21] For example, observer ratings (eg, medical student clinical assessments) should show strong evidence of internal structure characterized by high inter-rater agreement. Interpretations for multiple-choice exams, on the other hand, should have abundant content evidence. Both types of instrument would, of course, benefit greatly from multiple sources of evidence. Interpretations informing important decisions in any setting should be based on substantial validity evidence from multiple sources. Recent authors have proposed that the validity arguments for directly observable attributes (eg, handwashing habits) and those for observations intended to reflect a latent or theoretical trait (eg, feelings about disease prevention) are inherently different.[18,25] If accepted, this model will provide additional guidance regarding the relative importance of the various evidence sources.[36]

What About Face Validity?

Although the expression “face validity” has many meanings, it is usually used to describe the appearance of validity in the absence of empirical testing. This is akin to estimating the speed of a car based on its outward appearance or the structural integrity of a building based on a view from the curb. Such judgments amount to mere guesswork. The concepts of content evidence and face validity bear superficial resemblance but are in fact quite different. Whereas content evidence represents a systematic and documented approach to ensure that the instrument assesses the desired construct, face validity bases judgment on the appearance of the instrument. Downing and Haladyna note, “Superficial qualities . . . may represent an essential characteristic of the assessment, but . . . the appearance of validity is not validity.”[27] DeVellis[37] cites additional concerns about face validity, including fallibility of judgments based on appearance, differing perceptions among developers and users, and instances in which inferring intent from appearance might be counterproductive. For these reasons, we discourage use of this term.

RELIABILITY: NECESSARY, BUT NOT SUFFICIENT, FOR VALID INFERENCES

Reliability refers to the reproducibility or consistency of scores from one assessment to another.[19] Reliability is a necessary, but not sufficient, component of validity.[21,29] An instrument that does not yield reliable scores does not permit valid interpretations. Imagine obtaining blood pressure readings of 185/100 mm Hg, 80/40 mm Hg, and 140/70 mm Hg in 3 consecutive measurements over a 3-minute period in an otherwise stable patient. How would we interpret these results? Given the wide variation of readings, we would be unlikely to accept the average (135/70 mm Hg), nor would we rely on the first reading alone. Rather, we would probably conclude that the measurements are unreliable and seek additional information. Scores from psychometric instruments are just as susceptible to unreliability, but with one crucial distinction: It is often impractical or even impossible to obtain multiple measurements in a single individual. Thus, it is essential that ample evidence be accumulated to establish the reliability of scores before using an instrument in practice.

There are numerous ways to categorize and measure reliability (Table 3).[30,37-41] The relative importance of each measure will vary according to the instrument type.[30] Internal consistency measures how well the scores for individual items on the instrument correlate with each other and provides an approximation of parallel form reliability (see below). We would expect that scores measuring a single construct would correlate highly (high internal consistency). If internal consistency is low, it raises the possibility that the scores are, in fact, measuring more than one construct. Reproducibility over time (test-retest), between different versions of an instrument (parallel forms), and between raters (inter-rater) are other measures of reliability. The

Appendix contains more information on interpretation of these measures.

Generalizability theory[42] provides a unifying framework for the various reliability measures. Under this framework the unreliability of scores can be attributed to various sources of error (called *facets*), such as item variance, rater variance, and subject variance. Generalizability studies use analysis of variance to quantify the contribution of each error source to the overall error (unreliability) of the scores, just as analysis of variance does in clinical research. For further reading, see Shavelson and Webb's[43] primer on generalizability theory.

We emphasize that although reliability is prerequisite to validity, it is not sufficient.[29] This contrasts with what we have observed in the literature, where reliability is frequently cited as the sole evidence supporting a "valid instrument."[13,34] As noted above, evidence should be accumulated from multiple sources to support the validity of inferences drawn from a given instrument's scores. Reliability constitutes only one form of evidence. It is also important to note that reliability, like validity, is a property of the score and not the instrument itself.[30] The same instrument, used in a different setting or with different subjects, can demonstrate wide variation in reliability.[29,41]

PRACTICAL APPLICATION OF VALIDITY CONCEPTS IN SELECTING AN INSTRUMENT

Consumers of previously developed psychometric instruments in clinical practice, research, or education need to carefully weigh the evidence supporting the validity of the interpretations they are trying to make. Scores from a popular instrument may not have evidence to justify their use. Many authors cite evidence from only one or two sources, such as reliability or correlation with another instrument's scores, to support the validity of interpretations. Such instruments should be used with caution. To illustrate the application of these principles in selecting an instrument, we will systematically evaluate an instrument to assess symptoms of benign prostatic hypertrophy in English-speaking men.

First we must identify potential instruments. Reviewing articles from a MEDLINE search using the terms "prostatic hyperplasia" and "symptom" reveals multiple instruments used to assess benign prostatic hypertrophy symptoms.[1,44-48] The American Urological Association Symptom Index[1] (AUA-SI, also known as the International Prostate Symptom Score) seems to be by far the most commonly used instrument. After confirming our impression with a local expert, we select this instrument for further review.

Content evidence for AUA-SI scores is abundant and fully supportive.[1] The instrument authors reviewed both published and unpublished sources to develop an initial item pool that reflected the desired content domain. Word choice, time frame, and response set were carefully defined. Items were deleted or modified after pilot testing.

Some response process evidence is available. Patient debriefing revealed little ambiguity in wording, except for one question that was subsequently modified.[1] Scores from self-administration or interview are similar.[49]

Internal structure is supported by good to excellent internal consistency and test-retest reliability,[1,49,50] although not all studies confirm this.[51] Factor analysis confirms two theorized subscales.[50,52]

In regard to relations to other variables, AUA-SI scores distinguished patients with clinical benign prostatic hypertrophy from young healthy controls,[1] correlated with other indices of benign prostatic hypertrophy symptoms,[53] and improved after prostatectomy.[54] Another study found that patients with a score decrease of 3 points felt slightly improved.[51] However, a study found no significant association between scores and urinary peak flow or postvoid residual.[55]

Evidence of consequences is minimal. Thresholds for mild, moderate, and severe symptoms were developed by comparing scores with global symptom ratings,[1] suggesting that such classifications are meaningful. One study[56] found that 81% of patients with mild symptoms did not require therapy over 2 years, again supporting the meaning (validity) of these scores. More meaningful evidence of consequences might come from a study comparing the outcomes of men whose treatment was guided by the AUA-SI, compared with men whose treatment was guided by clinical judgment alone, but we are not aware of such a study.

In summary, AUA-SI scores are well supported by evidence of content, internal structure, relations to other variables, and to a lesser extent response process, whereas evidence of consequences is minimal. These scores are likely to be useful, although their meaning (consequences on patient care) could be studied further. For completeness we ought to similarly evaluate some of the other available instruments. Also, because validity and reliability evidence may not generalize to new settings, we should collect confirmatory data in our own clinic.

PRACTICAL APPLICATION OF VALIDITY CONCEPTS IN DEVELOPING AN INSTRUMENT

When developing psychometric instruments, careful attention should again be given to each category of validity evidence in turn. To illustrate the application of these principles, we will discuss how evidence could be planned, collected, and documented when developing an assessment of clinical performance for internal medicine residents.

The first step in developing any instrument is to identify the construct and corresponding content. In our example we could look at residency program objectives and other published objectives such as Accreditation Committee for Graduate Medical Education competencies,[57] search the literature on qualifications of ideal physicians, or interview faculty and residents. We also should search the literature for previously published instruments, which might be used verbatim or adapted. From the themes (constructs) identified we would develop a blueprint to guide creation of individual questions. Questions would ideally be written by faculty trained in question writing and then checked for clarity by other faculty.

For response process, we would ensure that the response format is familiar to faculty, or if not (eg, if we use computer-based forms), that faculty have a chance to practice with the new format. Faculty should receive training in both learner assessment in general and our form specifically, with the opportunity to ask questions. We would ensure security measures and accurate scoring methods. We could also conduct a pilot study in which we ask faculty to "think out loud" as they observe and rate several residents.

In regard to internal structure, inter-rater reliability is critical so we would need data to calculate this statistic. Internal consistency is of secondary importance for performance ratings,[30] but this and factor analysis would be useful to verify that the themes or constructs we identified during development hold true in practice.

For relations to variables, we could correlate our instrument scores with scores from another instrument assessing clinical performance. Note, however, that this comparison is only as good as the instrument with which comparison is made. Thus, comparing our scores with those from an instrument with little supporting evidence would have limited value. Alternatively, we could compare the scores from our instrument with United States Medical Licensing Examination scores, scores from an in-training exam, or any other variable that we believe is theoretically related to clinical performance. We could also plan to compare results among different subgroups. For example, if we expect performance to improve over time, we could compare scores among postgraduate years. Finally, we could follow residents into fellowship or clinical practice and see whether current scores predict future performance.

Last, we should not neglect evidence of consequences. If we have set a minimum passing score below which remedial action will be taken, we must clearly document how this score was determined. If subgroup analysis reveals unexpected relationships (eg, if a minority group is consistently rated lower than other groups), we should investigate whether this finding reflects on the validity of the test. Finally, if low-scoring residents receive remedial action, we could perform follow-up to determine whether this intervention was effective, which would support the inference that intervention was warranted.

It should now be clear that the collection of validity evidence requires foresight and careful planning. Much of the data described above will not be available without conscious effort. We encourage developers or researchers of psychometric instruments to systematically use the 5 sources of validity evidence as a framework when developing or evaluating instruments.

CONCLUSION

A clear understanding of validity and reliability in psychometric assessment is essential for practitioners in diverse medical settings. As Foster and Cone note, "Science rests on the adequacy of its measurement. Poor measures provide a weak foundation for research and clinical endeavors."[18] Validity concerns the degree to which scores reflect the intended underlying construct, and refers to the interpretation of results rather than the instrument itself. It is best viewed as a carefully structured argument in which evidence is assembled to support or refute proposed interpretations of results. Reproducible (reliable) results are necessary, but not sufficient, for valid inferences to be drawn. Although this review focused on psychometric instruments, many of the concepts discussed here have implications for other health care applications such as rater agreement in radiology,[58] illness severity scales,[59,60] data abstraction forms, and even clinical pathways.[61] Increased attention to the systematic collection and appraisal of validity evidence will improve assessments in research, education, and patient care.

ACKNOWLEDGMENTS

We thank Steven M. Downing, PhD (University of Illinois at Chicago, Department of Medical Education), for his insights and constructive critique.

APPENDIX: INTERPRETATION OF RELIABILITY INDICES AND FACTOR ANALYSIS

Reliability is usually reported as a coefficient[41] ranging from 0 to 1. The reliability coefficient can be interpreted as the correlation between scores on two administrations of the same instrument, and in fact test-retest and alternate form reliability are usually calculated using statistical tests of correlation. The reliability coefficient can also be interpreted as the proportion of score variance explained by differences between subjects (the remainder being explained by a combination of random and systematic error). A value of 0 represents no correlation (all error), whereas 1 represents perfect correlation (all variance attributable to subjects). Acceptable values will vary according to the purpose of the instrument. For high-stakes settings (eg, licensure examination) reliability should be greater than 0.9, whereas for less important situations values of 0.8 or 0.7 may be acceptable.[30] Note that the interpretation of reliability coefficients is different than the interpretation of correlation coefficients in other applications, where a value of 0.6 would often be considered quite high.[62] Low reliability can be improved by increasing the number of items or observers and (in education settings) using items of medium difficulty.[30] Improvement expected from adding items can be estimated using the Spearman-Brown "prophecy" formula (described elsewhere).[41]

A less common, but often more useful,[63] measure of score variance is the standard error of measurement (SEM) (not to be confused with the standard error of the mean, which is also abbreviated SEM). The SEM, given by the equation SEM = standard deviation × square root (1-reliability),[64] is the "standard deviation of an individual's observed scores"[19] and can be used to develop a confidence interval for an individual's true score (the true score is the score uninfluenced by random error). For example, 95% of an individual's scores on retesting should fall within 2 SEM of the individual's true score. Note, however, that the observed score only estimates the true score; see Harvill[64] for further discussion.

Agreement between raters on binary outcomes (eg, heart murmur present: yes or no?) is often reported using kappa, which represents agreement corrected for chance.[40] A different but related test, weighted kappa, is necessary when determining inter-rater agreement on ordinally ranked data (eg, Likert scaled responses) to account for the variation in intervals between data points in ordinally ranked data (eg, in a typical 5-point Likert scale the "distance" from 1 to 2 is likely different than the distance from 2 to 3). Landis and Koch[65] suggest that kappa less than 0.4 is poor, from 0.4 to 0.75 is good, and greater than 0.75 is excellent.

Factor analysis[32] is used to investigate relationships between items in an instrument and the constructs they are intended to measure. Some instruments intend to measure a single construct ("symptoms of urinary obstruction"), whereas others try to assess multiple constructs ("depression," "anxiety," and "personality disorder"). Factor analysis can determine whether the items intended to measure a given construct actually "cluster" together into "factors" as expected. Items that "load" on more than one factor, or on unexpected factors, may not be measuring their intended constructs.

References

1. Barry MJ, Fowler FJ Jr, O'Leary MP, et al. The American Urological Association symptom index for benign prostatic hyperplasia. *J Urol.* 1992;148:1549-1557.
2. Ewing JA. Detecting alcoholism: the CAGE questionnaire. *JAMA.* 1984;252:1905-1907.
3. Spitzer RL, Williams JB, Kroenke K, et al. Utility of a new procedure for diagnosing mental disorders in primary care. The PRIME-MD 1000 study. *JAMA.* 1994;272:1749-1756.
4. Green C, Porter C, Bresnahan D, Spertus J. Development and evaluation of the Kansas City Cardiomyopathy Questionnaire: a new health status measure for heart failure. *J Am Coll Cardiol.* 2000;35:1245-1255.
5. Hollen P, Gralla R, Kris M, Potanovich L. Quality of life assessment in individuals with lung cancer: testing the Lung Cancer Symptom Scale (LCSS). *Eur J Cancer.* 1993;29A(Suppl 1):S51-S58.
6. Bauer UE, Johnson TM, Hopkins RS, Brooks RG. Changes in youth cigarette use and intentions following implementation of a tobacco control program: findings from the Florida Youth Tobacco Survey, 1998-2000. *JAMA.* 2000;284:723-728.
7. National Board of Medical Examiners. United States Medical Licensing Exam Bulletin. Produced by Federation of State Medical Boards of the United States and the National Board of Medical Examiners. Available at: http://www.usmle.org/bulletin/2005/testing.htm. Accessed March 7, 2005.
8. Norcini JJ, Blank LL, Duffy FD, Fortna GS. The mini-CEX: a method for assessing clinical skills. *Ann Intern Med.* 2003;138:476-481.
9. Litzelman DK, Stratos GA, Marriott DJ, Skeff KM. Factorial validation of a widely disseminated educational framework for evaluating clinical teachers. *Acad Med.* 1998;73:688-695.
10. Merriam-Webster Online. Available at: http://www.m-w.com/. Accessed March 10, 2005.
11. Sackett DL, Richardson WS, Rosenberg W, Haynes RB. *Evidence-Based Medicine: How to Practice and Teach EBM.* Edinburgh: Churchill Livingstone; 1998.
12. Wallach J. *Interpretation of Diagnostic Tests.* 7th ed. Philadelphia: Lippincott Williams & Wilkins; 2000.
13. Beckman TJ, Ghosh AK, Cook DA, Erwin PJ, Mandrekar JN. How reliable are assessments of clinical teaching? A review of the published instruments. *J Gen Intern Med.* 2004;19:971-977.
14. Shanafelt TD, Bradley KA, Wipf JE, Back AL. Burnout and self-reported patient care in an internal medicine residency program. *Ann Intern Med.* 2002;136:358-367.
15. Alexander GC, Casalino LP, Meltzer DO. Patient-physician communication about out-of-pocket costs. *JAMA.* 2003;290:953-958.
16. Pittet D, Simon A, Hugonnet S, Pessoa-Silva CL, Sauvan V, Perneger TV. Hand hygiene among physicians: performance, beliefs, and perceptions. *Ann Intern Med.* 2004;141:1-8.
17. Messick S. Validity. In: Linn RL, editor. *Educational Measurement, 3rd Ed.* New York: American Council on Education and Macmillan; 1989.
18. Foster SL, Cone JD. Validity issues in clinical assessment. *Psychol Assess.* 1995;7:248-260.
19. 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.
20. Bland JM, Altman DG. Statistics notes: validating scales and indexes. *BMJ.* 2002;324:606-607.
21. Downing SM. Validity: on the meaningful interpretation of assessment data. *Med Educ.* 2003;37:830-837.
22. 2005 Certification Examination in Internal Medicine Information Booklet. Produced by American Board of Internal Medicine. Available at: http://www.abim.org/resources/publications/IMRegistrationBook.pdf. Accessed September 2, 2005.
23. Kane MT. An argument-based approach to validity. *Psychol Bull.* 1992;112:527-535.
24. Messick S. Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *Am Psychol.* 1995;50:741-749.
25. Kane MT. Current concerns in validity theory. *J Educ Meas.* 2001;38:319-342.
26. American Psychological Association. *Standards for Educational and Psychological Tests and Manuals.* Washington, DC: American Psychological Association; 1966.
27. Downing SM, Haladyna TM. Validity threats: overcoming interference with proposed interpretations of assessment data. *Med Educ.* 2004;38:327-333.
28. Haynes SN, Richard DC, Kubany ES. Content validity in psychological assessment: a functional approach to concepts and methods. *Psychol Assess.* 1995;7:238-247.
29. Feldt LS, Brennan RL. Reliability. In: Linn RL, editor. *Educational Measurement, 3rd Ed.* New York: American Council on Education and Macmillan; 1989.
30. Downing SM. Reliability: on the reproducibility of assessment data. *Med Educ.* 2004;38:1006-1012.
31. Clark LA, Watson D. Constructing validity: basic issues in objective scale development. *Psychol Assess.* 1995;7:309-319.
32. Floyd FJ, Widaman KF. Factor analysis in the development and refinement of clinical assessment instruments. *Psychol Assess.* 1995;7:286-299.
33. Campbell DT, Fiske DW. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol Bull.* 1959;56:81-105.

34. Beckman TJ, Cook DA, Mandrekar JN. What is the validity evidence for assessments of clinical teaching? *J Gen Intern Med.* 2005;20:1159-1164.
35. Smith GT, McCarthy DM. Methodological considerations in the refinement of clinical assessment instruments. *Psychol Assess*. 1995;7: 300-308.
36. Kane MT. Content-related validity evidence in test development. In: Downing SM, Haladyna TM, editors. *Handbook of Test Development.* Mahwah, NJ: Lawrence Erlbaum Associates; 2006,131-153.
37. DeVellis RF. *Scale Development: Theory and Applications.* 2nd ed. Thousand Oaks, CA: Sage Publications; 2003.
38. Nunnally JC, Berstein IH. *Psychometric Theory.* 3rd ed. New York: McGraw-Hill; 1994.
39. McMillan J, Schumacher S. *Research in Education: A Conceptual Introduction.* 5th ed. New York: Addison Wesley Longman; 2001.
40. Howell D. *Statistical Methods for Psychology.* 5th ed. Pacific Grove, CA: Duxbury; 2002.
41. Traub RE, Rowley GL. An NCME instructional module on understanding reliability. *Educational Measurement: Issues and Practice.* 1991;10(1):37-45.
42. Brennan RL. *Generalizability Theory.* New York: Springer-Verlag; 2001.
43. Shavelson R, Webb N. *Generalizability Theory: A Primer.* Newbury Park: Sage Publications; 1991.
44. Boyarsky S, Jones G, Paulson DF, Prout GR Jr. New look at bladder neck obstruction by the Food and Drug Administration regulators: guidelines for investigation of benign prostatic hypertrophy. *Trans Am Assoc Genitourin Surg.* 1976;68:29-32.
45. Madsen PO, Iversen P. A point system for selecting operative candidates. In: Hinman F, editor. *Benign Prostatic Hypertrophy.* New York: Springer-Verlag; 1983,763-765.
46. Fowler FJ Jr, Wennberg JE, Timothy RP, Barry MJ, Mulley AG Jr, Hanley D. Symptom status and quality of life following prostatectomy. *JAMA*. 1988;259:3018-3022.
47. Hald T, Nordling J, Andersen JT, Bilde T, Meyhoff HH, Walter S. A patient weighted symptom score system in the evaluation of uncomplicated benign prostatic hyperplasia. *Scand J Urol Nephrol.* 1991; 138(suppl):59-62.
48. Donovan JL, Abrams P, Peters TJ, et al. The ICS-"BPH" Study: the psychometric validity and reliability of the ICSmale questionnaire. *Br J Urol*. 1996;77:554-62.
49. Barry MJ, Fowler FJ, Chang Y, Liss CL, Wilson H, Stek M Jr. The American Urological Association symptom index: does mode of administration affect its psychometric properties? *J Urol.* 1995;154: 1056-1059.
50. Welch G, Kawachi I, Barry MJ, Giovannucci E, Colditz GA, Willett WC. Distinction between symptoms of voiding and filling in benign prostatic hyperplasia: findings from the Health Professionals Follow-up Study. *Urology*. 1998;51:422-427.
51. Barry MJ, Williford WO, Chang Y, et al. Benign prostatic hyperplasia specific health status measures in clinical research: how much change in the American Urological Association symptom index and the benign prostatic hyperplasia impact index is perceptible to patients? *J Urol*. 1995;154:1770-1774.
52. Barry MJ, Williford WO, Fowler FJ Jr, Jones KM, Lepor H. Filling and voiding symptoms in the American Urological Association symptom index: the value of their distinction in a Veterans Affairs randomized trial of medical therapy in men with a clinical diagnosis of benign prostatic hyperplasia. *J Urol*. 2000;164:1559-1564.
53. Barry MJ, Fowler FJ Jr, O'Leary MP, Bruskewitz RC, Holtgrewe HL, Mebust WK. Correlation of the American Urological Association symptom index with self-administered versions of the Madsen-Iversen, Boyarsky and Maine Medical Assessment Program symptom indexes. *J Urol.* 1992;148:1558-1563.
54. Schwartz EJ, Lepor H. Radical retropubic prostatectomy reduces symptom scores and improves quality of life in men with moderate and severe lower urinary tract symptoms. *J Urol*. 1999;161:1185-1188.
55. Barry MJ, Cockett AT, Holtgrewe HL, McConnell JD, Sihelnik SA, Winfield HN. Relationship of symptoms of prostatism to commonly used physiological and anatomical measures of the severity of benign prostatic hyperplasia. *J Urol*. 1993;150:351-358.
56. Kaplan SA, Olsson CA, Te AE. The American Urological Association symptom score in the evaluation of men with lower urinary tract symptoms: at 2 years of followup, does it work? *J Urol*. 1996;155: 1971-1974.
57. Program Requirements for Residency Education in Internal Medicine. Produced by Accreditation Council for Graduate Medical Education. Available at: http://www.acgme.org/. Accessed December 22, 2003.
58. Kundel H, Polansky M. Measurement of observer agreement. *Radiology*. 2003;228:303-308.
59. Knaus W, Wagner D, Draper E, et al. The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults. *Chest*. 1991;100:1619-1636.
60. Fine MJ, Auble TE, Yealy DM, et al. A prediction rule to identify low-risk patients with community-acquired pneumonia. *N Engl J Med*. 1997;336:243-250.
61. Marrie TJ, Lau CY, Wheeler SL, Wong CJ, Vandervoort MK, Feagan BG. A controlled trial of a critical pathway for treatment of community-acquired pneumonia. CAPITAL Study Investigators. *JAMA*. 2000;283:749-755.
62. Fraenkel JR, Wallen NE. *How to Design and Evaluate Research in Education.* New York, NY: McGraw-Hill; 2003.
63. Cronbach LJ. My current thoughts on coefficient alpha and successor procedures. *Educ Psychol Meas*. 2004;64:391-418.
64. Harvill LM. NCME Instructional module: standard error of measurement. *Educational Measurement: Issues and Practice*. 1991;10(2):33-41.
65. Landis J, Koch G. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33:159-174.