DOI: 10.1111/j.1475-6773.2005.00445.x

Integrating Validity Theory with Use of Measurement Instruments in Clinical Settings

P. Adam Kelly, Kimberly J. O'Malley, Michael A. Kallen, and Marvella E. Ford

Objective. To present validity concepts in a conceptual framework useful for research in clinical settings.

Principal Findings. We present a three-level decision rubric for validating measurement instruments, to guide health services researchers step-by-step in gathering and evaluating validity evidence within their specific situation. We address *construct precision*, the capacity of an instrument to measure constructs it purports to measure and differentiate from other, unrelated constructs; *quantification precision*, the reliability of the instrument; and *translation precision*, the ability to generalize scores from an instrument across subjects from the same or similar populations. We illustrate with specific examples, such as an approach to validating a measurement instrument for veterans when prior evidence of instrument validity for this population does not exist.

Conclusions. Validity should be viewed as a property of the interpretations and uses of scores from an instrument, not of the instrument itself: how scores are used and the consequences of this use are integral to validity. Our advice is to liken validation to building a court case, including discovering evidence, weighing the evidence, and recognizing when the evidence is weak and more evidence is needed.

Key Words. Decision making, measurement, psychometrics, survey research, veterans

Validity of measurement instruments is crucial to successful health outcomes measurement and to the health decision making that follows, a point often stated in the clinical research literature. The importance of validity is particularly evident when one contemplates measurement in a setting with unique population characteristics, such as the Veterans Affairs (VA) health care system. However, as described elsewhere in this supplement (Sechrest 2005), validity of measures is no simple matter. In fact, validity is so complex that over time health services researchers have taken a seemingly endless

variety of approaches to assessing it. In an attempt to introduce some semblance of order, the clinical research literature includes numerous articles on how to assess the validity of a measurement instrument (e.g., Devellis 1996; Elasy and Gaddy 1998; Hays, Anderson, and Revicki 1998; Tennant 2000) as well as how *not* to (e.g., Knapp 1985, 1990; Knapp and Brown 1995). In addition, the Scientific Advisory Committee of the Medical Outcomes Trust has published a list of "Instrument Review Criteria" (1995) to provide guidance for readers in their quest to assess the "appropriateness" (i.e., validity) of a measurement instrument for a given setting.

Our goal in this article is not to rewrite the best practices for assessing validity of measurement instruments. Instead, we organize some current validity concepts for health services researchers and transfer these concepts from the theoretical realm of psychometrics to the real-world context of clinical measurement. In addition, these concepts are framed in a model that emphasizes collecting and weighing evidence from multiple sources before reaching a conclusion on validity. But first, we clarify our definition of validity. We concur with Messick (1989, 1995) that validity is a property of *inferences*, not of instruments or their scores. That is, it is not the instrument itself that is to be declared "valid" (or "invalid"), nor is it the scores from the instrument. Rather, it is the inferences about individuals, drawn from the *interpretation* and/or clinical *use* of those individuals' scores in the *specific clinical context*, that is to be judged valid (or invalid). For example, we would discourage the statements (a) "The XYZ Anxiety Scale is valid," or (b) "The scores on the XYZ Anxiety Scale are valid." Instead, we would suggest the statements (c) "The *interpretation* of a high XYZ Anxiety Scale score as being indicative of a patient with a high anxiety level, in the *specific clinical context* of this chronic care facility for veterans, is valid," and/or (d) "The clinical implication (i.e., *use*) of a high XYZ score for a patient, in the *specific clinical context* of this chronic care facility for veterans, is the relocation of the patient to a private room and removal of all caffeine from the patient's dietary intake."

Address correspondence to P. Adam Kelly, Ph.D., M.B.A., Measurement Excellence and Training Resource Information Center (METRIC), Health Services Research and Development Service, U.S. Department of Veterans Affairs, 2002 Holcombe Blvd. (152), Houston, TX 77030. Kimberly J. O'Malley, Ph.D., is with Pearson Educational Measurement, Austin, TX. Michael A. Kallen, Ph.D., M.P.H., is with Measurement Excellence and Training Resource Information Center (METRIC), Health Services Research and Development Service, U.S. Department of Veterans Affairs, Houston, TX. Marvella E. Ford, Ph.D., M.S.W., is with the Department of Biostatistics, Bioinformatics & Epidemiology, Medical University of South Carolina, Charleston, SC.

An important difference distinguishes statements (a) and (b) from statements (c) and (d). Unlike statements (a) and (b), statements (c) and (d) are not "givens" at all, but instead are arguments that must be substantiated with evidence collected over time. In general, the *types* of evidence cited in manuscripts (e.g., criterion-related, concurrent, and predictive validity; Cronbach's α; test–retest reliability) are correctly cited and calculated. Thus we sense no necessity to challenge the definitions or theoretical underpinnings of the *types* of evidence cited in the literature. Rather, as Knapp and Brown (1995) and others do, we challenge the *reasons* authors give, if any, for choosing to focus on particular types of validity evidence, and we recommend that health services researchers adopt a more systematic process for gathering and examining validity evidence in the future.

Extending beyond Knapp and Brown (1995) and others, we propose a systematic process, in the form of a decision rubric for guiding investigators through three levels of validity evidence gathering (Figure 1). The decision rubric assumes that the investigator already has settled on (a) a question of interest, (b) a well-defined underlying concept, or "construct," to measure in order to answer the question, and (c) a well-developed rationale for why the construct chosen answers the question. The decision rubric directs her to examine three decisions: *construct precision*, *quantification precision*, and *translation precision*.

The first level of the decision rubric, construct precision, addresses the adequacy with which the construct of interest is represented by a measurement instrument. That is, does the construct supposedly targeted by the measurement instrument, in fact, target with adequate focus—what psychometricians refer to as "unidimensionality"—and clarity? For example, some instruments still in use today, such as the original SF-36 health survey, contain item wording that may confuse or distract patients.[1] Furthermore, how does the investigator know that the construct targeted by the instrument and the construct of interest are, in fact, one and the same? For example, some instruments with highly recognizable names actually measure things other than what their names imply.[2] In these circumstances, there is a need to perform construct validation procedures, such as (1) examining the correlation of scores from the chosen instrument with scores from other instruments known to measure similar constructs as well as those known to measure unrelated constructs and (2) triangulating the constructs evident in the instrument with data from relevant alternate sources, such as structured interviews of patients and prior construct validation work on the instrument. Procedures such as these are actually individual components of the

Figure 1: A Decision Making Rubric for Clinical Use of Measurement Instruments

Integrating Validity Theory with Clinical Use of Measurement Instruments

Q: Where do I start?

A: Assume you have identified a question, and in order to answer it, you must measure something (usually, a construct). Respond to these questions:

- *What is it you must measure?*
- *How will measuring this construct answer your question?*
- *Why are you focusing on this and not other related constructs? (In fact, list the other related constructs)*

Be precise!

~~~~~~~~~~~~~~~~

Q: OK, I've got a specific construct in mind, and a rationale for measuring it to answer my question. Now what?

A: When contemplating how/in whom/with what to measure, we recommend you follow this decision rubric, focusing in turn on *construct precision*, *quantification precision*, and *translation precision:*

Construct Precision

How precise is your definition of what you want to measure? Even though you think you've identified what it is you want to measure, investigate it further. Constructs can be complex. Consider the following:

Question	*Solution*
➢ *Is the construct you want to measure unidimensional, or is it actually a set of related constructs? How do you know this?*	➢ *Literature review, theory-centered critical analysis, factor analysis, multitrait-multimethod analysis*
➢ *What about the writing quality of the items? Are they written "right?" (technically accurate, readable, elicit anticipated responses)*	➢ *Your own review; expert review (if available), small-sample try-out*
~~~~~~~~~~~~~~~~

Figure 1: *Continued*

Quantification Precision

How dependable is the measurement in producing the same results for the same sample, over and over?

Question	*Solution*
➢ *How consistent, responsive to change, and focused on a cut-point (if this is relevant) is the instrument?*	➢ *Reliability analysis, sensitivity analysis/ determination of MCID, specificity analysis*
➢ *Is it possible to reduce patient/provider burden without losing this information?*	➢ *Prudent choice of instruments/subscales/ items; modern measurement (CAT)*
➢ *How likely am I to see during my career lifespan the availability of CATs in my field?*	➢ *Modern measurement is a topic or rapidly growing interest in medicine; we expect many new CATs to become commercially available in the near future*

Translation Precision

Also known as generalizability. How well does the measurement translate from your study sample to other relevant samples/groups/populations?

Question	*Solution*
➢ *To what extent is the meaning of items similar across different population groups?*	➢ *Follow a comprehensive screening process for selecting the appropriate instrument for your population (e.g., the list of 17 "questions to answer" provided in our presentation)*

For more information, please visit us at www.MeasurementExperts.org
Measurement Excellence and Training Resource Information Center (METRIC)
U.S. Department of Veterans Affairs, Health Services Research & Development Service

"multitrait-multimethod matrix" (MTMM) approach to construct validation (Campbell and Fiske 1959).

A construct validation procedure that has grown in popularity is confirmatory factor analysis (CFA), a technique that "makes fully explicit the conceptual relations [across traits and methods] that are only implicit in the traditional bivariate analysis of the MTMM matrix" (Ferketich, Figuerdo, and Knapp 1991, p. 319). Advantages of CFA include (1) quantification of *indirect* relationships between predictor and outcome variables as well as direct ones, potentially yielding more and better validity evidence, (2) simultaneous analysis of larger numbers of variables than is practical using MTMM, and (3) availability of statistical indices that aid in interpreting the goodness-of-fit of a CFA model to the data. CFA is now accessible through user-friendly interfaces in commercial statistical analysis software packages, and measurement-related articles that include CFA have become common in the health services research literature. We encourage investigators to become familiar with the basics of this analysis technique. Further discussion of performing and interpreting CFA in a construct validation context is provided in Figuerdo, Ferketich, and Knapp (1991).

And what of so-called "content validity?" It is actually subsumed within the concept of "construct precision" as well. Messick (1989, 1995) argues that *content* validity evidence is in fact *construct* validity evidence, in that the content of the construct of interest should be comprehensively sampled, and thus well represented, by the content of the instrument. All content in the instrument should map back to the instrument's construct(s) unambiguously. There should be no "orphan" content lacking a clear linkage to the construct(s), as any such content would represent an unintentional measurement of a separate, additional construct.

The second level of the decision rubric addresses the issue of *quantification precision.* This encompasses two considerations, (1) the *reliability* of a measurement instrument and (2) the consequences of patient and investigator burden imposed by using the instrument. Reliability is "the consistency of . . . measurements when the testing procedure is repeated on a population of individuals or groups," as defined in the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education 1999, p. 25). The *Standards* lists a variety of axioms on reliability, including two particularly germane to clinical measurement: (1) a higher level of reliability is necessary when scores are to be used to make decisions concerning individuals rather than groups and (2) a higher degree of reliability is necessary when

scores are to be used to make decisions that have extreme and/or irreversible consequences (i.e., high-stakes decisions). Because situations relevant to one or both of these axioms are encountered frequently in the clinical milieu, reliability should be a key issue in selecting measurement instruments for use in clinical research.

Returning to our earlier example, consider the following three scenarios, each representing a different intended use of the scores from the XYZ Anxiety Scale. In Scenario 1, the scores from the XYZ Anxiety Scale will be used to classify patients a priori into two groups, "Low Anxiety" and "High Anxiety," in preparation for a multisite clinical study of a new educational intervention using nondrug relaxation techniques. In Scenario 2, XYZ scores will be used to classify patients a priori, but this time in preparation for a 6-month test of a new oral sedative. Lastly, in Scenario 3, XYZ scores will be used to determine who should receive a sedative immediately after the patient completes the XYZ Anxiety Scale. If existing literature reports a Cronbach's α coefficient for the XYZ Anxiety Scale averaging 0.70 across comparable settings, we would likely characterize that level as *adequate* evidence of reliability for Scenarios 1 and 2 (participation in multisite clinical research), but *inadequate* for scenarios with potential consequences as extreme as those in Scenario 3. In short, high reliability is a *necessary* condition for high-stakes uses.

In addition to the reliability of a measurement instrument, an investigator should take into account how the burden placed on patients—and herself—from using the instrument will impact quantification precision. Specifically, patient burden may arise from (1) the amount of time needed to complete an instrument and (2) any unease experienced in responding to generic questions, paradoxically designed to be appropriate for a wide range of patients yet frequently irrelevant or even inappropriate for many. Although it is often possible to alleviate some of the negative impacts of *time* burden by providing the patient with an incentive (e.g., a nominal cash payment or small gift), any burden arising from *discomfort* in answering irrelevant and/or inappropriate personal questions is less easily alleviated and can result in a loss of the patient's willingness to continue responding in a focused, honest way, in turn degrading the precision of quantification. For the investigator, burden arises from the opportunity cost of time spent finding an instrument, administering it to patients, entering responses, scoring them, and interpreting the scores. An excessive time burden may result in unwelcome stress for the investigator and, potentially, inadvertent degradation of the data. Computerized administration and scoring of an instrument, when available, may reduce the investigator's burden significantly, but often does little to reduce patient burden.

What if it were possible to reduce the burden on investigator and patients while simultaneously maintaining or even improving quantification precision? This is no longer wishful thinking. Over several decades, measurement has undergone a quiet evolution in which fundamental principles have been changed or even abandoned. An example of this is the relatively recent dismissal of a longstanding rule of psychometrics: "Longer tests are more reliable than shorter tests." The "old psychometrics," known as classical test theory (CTT), is theory-based, sample- and test-specific, and focused on respondents' total-score performance on a test; while "modern psychometrics," known as item response theory (IRT), is model-based, ability level-specific and focused instead on respondents' performance on each individual "item," or test question. With CTT, the test administrator must compute total scores for respondents in order to make score interpretations that are comparable across respondents. Thus, respondents must answer all items on the instrument, because omissions can and often do affect score interpretation. Alternatively, IRT introduces a world of customized item-by-item presentation, as each item from the instrument is mapped independently to an underlying continuum of respondent ability, ranging from lowest to highest level of ability. This mapping enables an ordering of items from "easiest" to "hardest" and a ranking of the respondent on the continuum after each item response, thereby eliminating the need for a total score.

As described elsewhere in this supplement (Cook, O'Malley, and Roddey 2005), IRT is the logic behind computer-adaptive testing (CAT): The computer program starts by presenting an item of average difficulty and, based only on the respondent's response *to that item*, the computer then selects the next item to administer—more difficult if the respondent got the first one right, less difficult if he got it wrong—and administers it. This cycle of item selection and administration repeats iteratively until any one of several stopping rules is triggered, for example, once the average change in a respondent's estimated location on the ability continuum fails to exceed some minimum after three consecutive items. By design, respondent burden arising from answering irrelevant or inappropriately easy or difficult items can largely be eliminated with CAT. The CAT stopping rules reduce respondent time burden as well, because administration ends once the computer has estimated a respondent's score to a prespecified level of precision, relieving the respondent from answering additional items. Moreover, the adaptive capacity of CAT item administration enables the generation of respondent ability estimates that often exhibit smaller standard errors of measurement than those obtained from total scores from the same instrument. Therefore, when CAT is an option for the

investigator, quantification precision actually can be *improved* with shorter tests.[3]

Through the first two levels of the decision rubric, strategies for ensuring and improving precision of the constructs measured and quantitative data collected have been addressed. We have introduced concepts and techniques that help clarify *which* constructs, and *how much* of each, are measured by an instrument. The third level of the decision rubric addresses the issue of *translation precision*, the generalizability of instrument scores to other populations, locations, and time periods of interest—for example, from a sample of patients to other similar groups of patients. As generalizing results logically implies also generalizing—or "translating"—inferences drawn from those results, we contend that translation precision constitutes another critically important consideration in assessing validity. For example, if the investigator wishes to extrapolate results (and inferences) to other patient groups, the instrument must be generalizable to those groups in a clearly demonstrable way. On the other hand, if the population the investigator seeks to measure consists of only the current study sample or is nearly identical in all respects to the study sample, the generalizability of the instrument is not at issue. What are the issues of interest when considering generalizability? Several of the most prominent ones are the sampling technique chosen (especially random versus convenience sampling); stratification of the sample by demographic variables such as patient gender, race, ethnicity, socioeconomic status, and highest educational level attained; choice of geographic location(s) for using the instrument; and time of day, month, and year that data collection is performed. Of course, each of these issues will carry differentially more or less importance given the particular context of the investigator and the study she proposes.

Once again turning to our example, the XYZ Anxiety Scale, imagine a chronic care facility for veterans where the majority of patients are of a different racial or ethnic group than the majority of clinicians and other health care providers. In this scenario, discordance exists in the racial or ethnic backgrounds of patients and providers, which could function as a source of anxiety for the patients. That is, patients may become anxious about the manner in which they will be perceived by their providers, and may worry that the quality or level of communication they receive will not be adequate or appropriate. Let us further assume that the XYZ Anxiety Scale was developed and psychometrically tested in several large, suburban medical complexes populated mostly by racially and ethnically homogeneous patients and providers, such that there are no items on the XYZ Anxiety Scale that *overtly*

measure racial/ethnic discordance-related anxiety. Not surprisingly, patients' responses to the scale could be *covertly* biased by the existence of racial/ethnic discordance in the facility, resulting in scores on the XYZ Anxiety Scale that do not accurately reflect the true reasons for the anxiety levels seen among these patients. At a minimum, an investigator planning to use the XYZ Anxiety Scale in this scenario would want to find or develop additional items that measure specific sources of anxiety for these patients. Otherwise, the biased scores could lead to flawed interpretations and inappropriate subsequent actions. In contrast, if the research question included estimating the general anxiety level of patient populations from various medical facilities throughout the region without regard for the source of that anxiety, then the XYZ Anxiety Scale might well be superior to context-specific anxiety scales. In fact, in such a scenario, administering an instrument that measures context-specific anxiety may actually create unnecessary burdens on patients and the investigator.

It is important to answer several key questions before selecting a measurement instrument for use with a specific population. First, has the instrument been used previously in the population of interest? Second, is it a general measure of the construct of interest, or a disease-specific measure? Third, what are the reliability coefficients of the instrument's components and/or the overall reliability of the instrument for the population of interest? Fourth, what validity evidence has been presented to date with respect to the way the instrument's scores are interpreted and used? Fifth, has this validity evidence been reported for the population group of interest? Finally, is the conceptual framework underlying the instrument relevant to and appropriate for the population of interest? These questions are extracted from "17 Questions," shown in Figure 2, to encapsulate the major points of McDowell and Newell (1996).

In summary, we have listed in Table 1 several recent examples of research studies, conceptual papers and research user guides that address issues pertaining to validation of measurement instruments. There is wide variability in the use of terms across the listed works, so we have distilled some of the most important validity considerations into the following take-away points:

- Ensure that the measurement instrument samples from patient tasks and/or behaviors that are relevant to the construct(s) of interest.
- Read the items and response options and ask yourself, "How well do I think the responses to these items reflect the construct?" Then use experts' judgment to help ensure you have it right.

Figure 2: "17 Questions," Adapted from McDowell and Newell (1996)

17 Questions to Answer
in Evaluating Generalizability of Instruments across Populations

(From McDowell, I., and C. Newell. 1996. Measuring Health. Oxford, England: Oxford University Press.)

1. What is the name of the measurement tool?
2. What is/are the name(s) of the measurement tool developer(s)?
3. In which population(s) and age group(s) was the tool developed? (Ex: Caucasians, men, college students, etc.)
4. To which population(s) has the tool been applied since its original publication?
5. Has the tool been tested in the population(s) of interest?
6. Has the tool been tested in older populations? If so, in which age cohort(s) was it tested?
7. Is the tool a general measure or a disease-specific measure?
8. What is the reading level of the measurement tool?
9. What are barriers to using the measurement tool?
10. What are the reliability coefficients of the tool components or the overall reliability of the tool, by population?
11. In what ways has the tool been validated?
12. Is the tool self- or professionally-administered?
13. What is the average length of time for completing the tool?
14. What is the conceptual approach to the topic area (such as psychological well-being)?
15. Is this conceptual approach relevant to/appropriate for the population of interest?
16. Is the original purpose of the tool appropriate for use in the proposed study?
17. What are the published citations of the measurement tool?

- Gather evidence of patients' engagement in different tasks and behaviors that are also reflective of the construct(s)—for example, by using "think-aloud" exercises or interviews.
- Look for high correlations with measures of like constructs, low correlations with measures of unlike ones.
- Look for consistency in responses across samples for which little to no change in the underlying construct(s) should have occurred.

Table 1: Examples of Publications That Present Validity Issues and/or Validation Techniques

Manuscript Title, Journal, Year	*"Type" of Validity*	*Validation Technique(s)*
"Psychometric Evaluation of Selected Pain Intensity Scales for Use with Cognitively Impaired and Cognitively Intact Older Adults," *Rehabilitation Nursing*, 2005	"Concurrent"	Spearman correlation across scales
"The Psychometric Properties of Five Scoring Methods Applied to the Script Concordance Test," *Academic Medicine*, 2005	"Differential"	Correlation with multiple-choice exam scores
	"Predictive"	Association with clinical performance
"Research Burnout: A Refined Multidimensional Scale," *Psychological Reports*, 2004	"Factorial"	Confirmatory factor analysis
	"Nomological"	Not provided
"A Level-of-Functioning Self-Report Measure for Consumers with Severe Mental Illness, *Psychological Services*, 2002	Construct	Correlation across persons
	"Further"	Correlation across self-reports, case manager ratings, interviewer ratings
"Sensation and Distress of Pain Scales: Reliability, Validity, and Sensitivity," *Journal of Nursing Measurement*, 2001	Convergent	Correlation across scales
	Construct	Correlation across scales
	Discriminant	Correlation across scales
"Development and Initial Validation of the Obsessive Beliefs Questionnaire," *Behavioral Research & Therapy*, 2001	Convergent	Partial correlation with other instruments
	"Known-groups"	MANOVA, ANOVA across groups, instruments
"Ensuring Content Validity: An Illustration of the Process," *Journal of Nursing Measurement*, 2001	Content	Agreement (κ) of experts; focus group of nurses

- Lastly, determine who will interpret and use the scores. Then talk to them, learn their views and the actions they plan to take once they obtain the scores, and develop a best- and worst-case scenario of the implications for patients.

Validity theory should become better integrated into the day-to-day practice of clinical measurement. Investigators can follow a clear set of guidelines for measurement decision making, such as the three-level decision rubric described in this manuscript. However, by definition, validation is an ongoing act of gathering evidence, *not* a deterministic process. There is no concrete yes-or-no determination of validity based solely on any one set of numbers.

Rather, validation is akin to building a court case, except that the investigator is collecting *all* the evidence, both "for" and "against" validity. Once the evidence is in, the investigator, perhaps with input from colleagues, plays the role of judge and jury, weighing the evidence "for" and "against" validity and pronouncing judgment. We further note that the evidence may not be exclusively quantitative; in fact, much of the evidence can be in the form of verbal statements or even observations. We recognize that the process of validation is arduous and that developing and gathering the necessary evidence can be an expensive undertaking. However, as we mentioned earlier, we have found that much of the investigative work published on instrument validation is on the right track, just incomplete. Following a detailed procedure for measurement decision making, such as our proposed rubric, will not necessarily increase the validation workload significantly. Rather, following such a procedure provides needed focus on measurement issues most likely to impact validity in a clinical context. The payoff should be more trustworthy measurement and, by extension, better informed decision making in a given clinical setting.

ACKNOWLEDGMENTS

The research reported here was supported by the U.S. Department of Veterans Affairs, Veterans Health Administration, Health Services Research and Development Service, Measurement Excellence and Training Resource Information Center (METRIC; RES 02-235). Dr. Kelly is supported by a career development award from the U.S. Department of Veterans Affairs, Veterans Health Administration, Health Services Research and Development Service. The views expressed in this manuscript are those of the authors and do not necessarily reflect the views of the U.S. Department of Veterans Affairs.

The material in this manuscript was presented as a workshop at the 27th annual meeting of the Society for General Internal Medicine on May 14, 2004. We thank Dr. David Kuykendall and two anonymous reviewers for their helpful comments and keen insight in framing the manuscript for the readership of this journal.

NOTES

1. The SF-36v2™ health survey (Ware, Kosinski, and Dewey 2000) features improvements in item wording including replacement of the terms "full of pep" and

"downhearted and blue" that many respondents found distracting in the original SF-36.

2. The Charlson index (Charlson et al. 1987) for "classifying prognostic comorbidity in longitudinal studies" is actually an index of both comorbidity and complication diagnoses.
3. There is one aspect of CAT that makes some health services researchers uncomfortable: pre- and postintervention assessments may involve different items. For example, if progress has occurred as a result of an intervention, then the items most appropriate for the posttreatment assessment may have higher difficulties, and thus be different items, from those used at preintervention. While this is psychometrically sound, it is often dismaying to health services researchers whose expectation is that an identical set of items would be used both before and after treatment.

REFERENCES

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. 1999. *Standards for Educational and Psychological Testing.* Washington, DC: American Educational Research Association.

Campbell, D. T., and D. W. Fiske. 1959. "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix." *Psychological Bulletin* 56: 81–105.

Charlson, M. E., P. Pompei, K. L. Ales, and C. R. McKenzie. 1987. "A New Method of Classifying Prognostic Comorbidity in Longitudinal Studies: Development and Validation." *Journal of Chronic Diseases* 40 (5): 373–83.

Cook, K. F., K. J. O'Malley, and T. S. Roddey. 2005. "Dynamic Assessment of Health Outcomes: Time to Let the C." *Health Services Research* DOI: 10.1111/j.1475-6778.2005.00446.x. Available at www.blackwell-synergy.com

DeVellis, R. F. 1996. "A Consumer's Guide to Finding, Evaluating, and Reporting on Measurement Instruments." *Arthritis Care and Research* 9 (3): 239–45.

Elasy, T. A., and G. Gaddy. 1998. "Measuring Subjective Outcomes: Rethinking Reliability and Validity." *Journal of General Internal Medicine* 13: 757–61.

Ferketich, S. L., A. J. Figuerdo, and T. R. Knapp. 1991. "The Multitrait-Multimethod Approach to Construct Validity." *Research in Nursing & Health* 14: 315–20.

Figuerdo, A. J., S. L. Ferketich, and T. R. Knapp. 1991. "More on MTMM: The Role of Confirmatory Factor Analysis." *Research in Nursing & Health* 14: 387–91.

Hays, R. D., R. T. Anderson, and D. Revicki. 1998. "Assessing Reliability and Validity of Measurement in Clinical Trials." In *Quality of Life Assessment in Clinical Trials*, edited by M. Staquet, R. D. Hays, and P. M. Fayers, pp. 169–82. Oxford, U.K.: Oxford University Press.

Knapp, T. R. 1985. "Validity, Reliability, and Neither." *Nursing Research* 34 (3): 189–92.

———. 1990. "Treating Ordinal Scales as Interval Scales: An Attempt to Resolve the Controversy." *Nursing Research* 39 (2): 121–3.

Knapp, T. R., and J. K. Brown. 1995. "Ten Measurement Commandments That Often Should Be Broken." *Research in Nursing & Health* 18: 465–9.

McDowell, I., and C. Newell. 1996. *Measuring Health.* Oxford, England: Oxford University Press.

Messick, S. 1989. "Validity." In *Educational Measurement*, 3d edition, edited by R. Linn, pp. 13–103. New York: Macmillan.

———. 1995. "Validity of Psychological Assessment: Validation of Inferences from Persons' Responses and Performances as Scientific Inquiry into Score Meaning." *American Psychologist* 50 (9): 741–9.

Scientific Advisory Committee of the Medical Outcomes Trust. 1995. "SAC Instrument Review Criteria (Volume 3, Number 4)" [accessed on March 28, 2005]. Available at http://www.proqolid.org/public/34sacrev.htm

Sechrest, L. 2005. "Validity of Measures Is No Simple Matter." *Health Services Research* DOI: 10.1111/j.1475-6773.2005.00443.x. Available at www.blackwell-synergy.com.

Tennant, A. 2000. "Measuring Outcome." *British Medical Bulletin* 56 (2): 287–95.

Ware, J. E., M. Kosinski, and J. E. Dewey. 2000. *How to Score Version Two of the SF-36 Health Survey.* Lincoln, RI: QualityMetric Inc.