Reliability and Validity: What's the Difference and Why Does it Matter?

TERRI S. ARMSTRONG, PhD, ANP-BC, FAANP, and IBRAHIMA GNING, DDS, DrPH, MPH

From University of Texas Health Science Center School of Nursing, Houston, Texas, and MD Anderson Cancer Center, Houston, Texas

Author's disclosures of potential conflict of interest are found at the end of this article.

Correspondence to: Terri S. Armstrong, PhD, ANP-BC, FAANP, Department of Family Health, University of Texas Health Science Center School of Nursing, 6901 Bertner Avenue, Room 791, Houston, TX 77030.
E-mail: Terri.S.Armstrong@uth.tmc.edu

J Adv Pract Oncol
2011;2:338–340

There has been increasing use of and support for patient-reported outcome (PRO) instruments to assess the impact of care on the patient (Brandrud et al., 2011; Cleeland & Sloan, 2010; US Department of Health and Human Services [HHS], 2006; Turk et al., 2006). The integrity of our analysis of care provided and study findings is predicated on having a questionnaire that is both valid and reliable (DeVon et al., 2007). Psychometrics is the subspecialty concerned with measurement procedures in psychological and social phenomena (DeVellis, 2003).

Patient-reported outcome instruments encompass "measurement of any aspect of a patient's health status that comes directly from the patient (without interpretation of the patient's responses by a physician or anyone else)," (HHS, 2006). The development of a PRO instrument is a multistep process. A full discussion of the components of instrument development is beyond the scope of this discussion. The reader is referred to several references for additional information related to the components of this process (Armstrong, Cohen, Eriksen, & Cleeland, 2005; DeVellis, 2003; HHS, 2006; Rothman, Beltran, Cappelleri, Lipscomb, & Teschendorf, 2007; Snyder, Watson, Jackson, Cella, & Halyard, 2007).

If you are considering using an instrument to measure outcomes in your patient population, you are encouraged to review the published literature on the development of the instrument in question for evidence of reliability and validity related to the intended use. Other considerations may include the ability to detect change (or sensitivity), feasibility, and interpretability. The following section briefly introduces the concepts of reliability and validity and provides guidelines for review of an instrument.

Reliability

According to Nunnally & Bernstein (1994), reliability refers to the ability of a PRO to consistently measure an attribute of interest (repeatable). Reliability does not assure that the instrument is measuring what it is intended to measure, only that one obtains consistent results when using it. Reliability is considered necessary for validity (i.e., the instrument must be reliable to be valid, but reliability does not assure

validity) (Nunnally & Bernstein, 1994). Common forms of reliability that are reported include test-retest and inter-rater (often referred to as stability reliability); Cronbach's alpha; and alternative form and parallel form (equivalence reliability) (DeVellis, 2003; Nunnally & Bernstein, 1994).

TEST-RETEST

Test-retest is the correlation between two scores indicating the stability of a questionnaire when administered to the same group at different times (DeVellis, 2003). The length of time between testing that is necessary is controversial and based on the construct being measured. In general, a time frame of 2 weeks to 1 month is generally accepted (Waltz, Strickland, & Lenz, 2005). In the case of time-dependent concepts such as symptoms, which could change within days or weeks, the time could be shorter with a second administration in 1 or 2 days. Research has shown equivalence between a 2-day or 2-week interval in a variety of health status instruments (Marx, Menezes, Horovitz, Jones, & Warren, 2003). The author should therefore state the time interval used and the criteria for minimal correlations. Typically a correlation of 0.3 is considered moderate, and 0.7 is considered high or strong (Cohen, 1988; DeVon et al., 2007).

CRONBACH'S ALPHA

Cronbach's alpha, the most common method of reliability testing, is the applicable method in a cross-sectional sample. It indicates how well items on a questionnaire fit together (DeVon et al., 2007). Cronbach's alpha should be calculated every time an instrument is used. Often it is calculated for the instrument as a whole and also for any groupings or factors within the questionnaire. A Cronbach's alpha of 0.7 is acceptable for new scales (DeVellis, 2003). High alphas (> 0.9) are sometimes thought to indicate redundancy (DeVellis, 2003; Knapp & Brown, 1995), whereas others recommend that for clinical use, the alpha should be 0.9 (Nunnally & Bernstein, 1994).

SEE PAGE 304

Use your smartphone to access the US Department of Health and Human Services document quoted on page 338.

Validity

Assessment of validity is an ongoing process, and an instrument is continually assessed with use in different populations and settings. Prior to using an instrument in a clinical setting or for research, reports of content validity, construct validity, and criterion-referenced validity should be assessed.

CONTENT VALIDITY

Content validity occurs if the complete range of an attribute is included in a questionnaire. Items to include in the questionnaire are determined by definition of the construct of interest, review of the literature and identification of domains of the construct, and review by experts and those experiencing the construct. Statistical techniques can then be employed to quantify the validity of the items. Determination of content validity is often a two-stage process: first items are reviewed and either confirmed, reworded, or deleted by content experts; then remaining items are scored and a content validity index is calculated (Armstrong et al., 2005; Lynn, 1986).

CONSTRUCT VALIDITY

The ability of a questionnaire to represent what it is attempting to measure is paramount to the use of the instrument related to the construct of interest. The construct is often an idea or theory comprising several concepts. During the development of an instrument, factor analysis is a group of techniques that can be employed to evaluate this (Nunnally & Bernstein, 1994). This is a complicated analysis, with a variety of methods used. Two primary types are principal component analysis and principal factor analysis. These techniques allow the researcher to explore the relationship between items and the amount of variance explained by the instrument.

Items on the instrument (termed factors) that represent an underlying latent concept are grouped together. The assessment allows the researcher to investigate the structure of the relationship between the items (how items are related to the latent or underlying concept). The number of factors is often determined by the number of groups that have an eigenvalue great-

er than 1 (this means that the group explains at least as much variance as a single item).

Often, the statistical analysis is completed, and then the researcher and statistician review the data for meaning and clarity. A name that represents the underlying concept represented by the group is then given to the factor. For example, nausea and vomiting are often grouped together in a factor and given the name "gastrointestinal." When considering the use of an instrument, a report of construct validity analysis should be reviewed for clarity of the factors and the amount of variance in the responses that can be explained by the solution.

CRITERION VALIDITY

Criterion validity reports on the relationship between the concept being measured by one questionnaire and attributes on another variable (DeVellis, 2003; Nunnally & Bernstein, 1994). This evaluation can include assessment of attributes known to be associated with the variable of interest or one that is opposite. For example, symptom severity is often thought to be associated with performance status (those with high symptom severity have a low performance status). This can be assessed when administered at the same time (concurrent validity) or in the future (predictive validity). Convergent validity can be assessed if two items to measure the same concept are highly correlated. Divergent validity occurs if a questionnaire designed to measure something opposite (for example, fatigue and vigor) is given at the same time, and these would not be expected to be correlated (Carmines & Zeller, 1979).

Feasibility and Applicability

The above discussion presents basic information related to the psychometric properties of reliability and validity of an instrument that should be reported on and assessed prior to use. In addition, there are issues related to feasibility and applicability that are equally important in the clinical setting. Consideration should be made regarding the burden of the instrument to the patient (length, time to complete), scoring, language, use in the population of interest, method of administration (paper and pencil, electronic), and ability to detect change if evaluating an intervention.

DISCLOSURE

The authors have no conflicts of interest to disclose.

REFERENCES

Armstrong, T. S., Cohen, M. Z., Eriksen, L., & Cleeland, C. (2005). Content validity of self-report measurement instruments: An illustration from the development of the Brain Tumor Module of the M.D. Anderson Symptom Inventory. *Oncology Nursing Forum, 32*(3), 669–676. doi:10.1188/05.ONF.669-676

Brandrud, A. S., Schreiner, A., Hjortdahl, P., Helljesen, G. S., Nyen, B., & Nelson, E. C. (2011). Three success factors for continual improvement in healthcare: An analysis of the reports of improvement team members. *British Medical Journal Quality & Safety, 20*(3), 251–259. doi:10.1136/bmjqs.2009.038604

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment*. Beverly Hills, CA: Sage.

Cleeland, C. S., & Sloan, J. A. (2010). Assessing the Symptoms of Cancer Using Patient-Reported Outcomes (ASCPRO): Searching for standards. *Journal of Pain and Symptom Management, 39*(6), 1077–1085. doi:10.1016/j.jpainsymman.2009.05.025

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). London: Routledge Academic.

DeVellis, R. F. (2003). *Scale development: Theory and applications* (2nd ed.). Thousand Oaks, CA: Sage Publications.

DeVon, H. A., Block, M. E., Moyle-Wright, P., Ernst, D. M., Hayden, S. J., Lazzara, D. J.,...Kostas-Polston, E. (2007). A psychometric toolbox for testing validity and reliability. *Journal of Nursing Scholarship, 39*(2), 155–164.

US Department of Health and Human Services. (2006). Guidance for industry: Patient-reported outcome measures: Use in medical product development to support labeling claims: Draft guidance. *Health and Quality of Life Outcomes, 4*, 79.

Knapp, T. R., & Brown, J. K. (1995). Ten measurement commandments that often should be broken. *Research in Nursing & Health, 18*(5), 465–469.

Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research, 35*(6), 382–385.

Marx, R. G., Menezes, A., Horovitz, L., Jones, E. C., & Warren, R. F. (2003). A comparison of two time intervals for test-retest reliability of health status instruments. *Journal of Clinical Epidemiology, 56*(8), 730–735. doi:10.1016/S0895-4356(03)00084-2

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Rothman, M. L., Beltran, P., Cappelleri, J. C., Lipscomb, J., & Teschendorf, B. (2007). Patient-reported outcomes: Conceptual issues. *Value Health, 10*(suppl 2), S66–S75.

Snyder, C. F., Watson, M. E., Jackson, J. D., Cella, D., & Halyard, M. Y. (2007). Patient-reported outcome instrument selection: Designing a measurement strategy. *Value Health, 10*(suppl 2), S76–S85.

Turk, D. C., Dworkin, R. H., Burke, L. B., Gershon, R., Rothman, M., Scott, J.,...Wyrwich, K. W. (2006). Developing patient-reported outcome measures for pain clinical trials: IMMPACT recommendations. *Pain, 125*(3), 208–215. doi:10.1016/j.pain.2006.09.028

Waltz, C. F., Strickland, O. L., & Lenz, E. R. (2005). *Measurement in nursing and health* (3rd ed.). New York: Springer.