A Psychometric Toolbox for Testing Validity and Reliability

Holli A. DeVon, Michelle E. Block, Patricia Moyle-Wright, Diane M. Ernst, Susan J. Hayden, Deborah J. Lazzara, Suzanne M. Savoy, Elizabeth Kostas-Polston

Purpose: To review the concepts of reliability and validity, provide examples of how the concepts have been used in nursing research, provide guidance for improving the psychometric soundness of instruments, and report suggestions from editors of nursing journals for incorporating psychometric data into manuscripts.

Methods: CINAHL, MEDLINE, and PsycINFO databases were searched using key words: validity, reliability, and psychometrics. Nursing research articles were eligible for inclusion if they were published in the last 5 years, quantitative methods were used, and statistical evidence of psychometric properties were reported. Reports of strong psychometric properties of instruments were identified as well as those with little supporting evidence of psychometric soundness.

Findings: Reports frequently indicated content validity but sometimes the studies had fewer than five experts for review. Criterion validity was rarely reported and errors in the measurement of the criterion were identified. Construct validity remains underreported. Most reports indicated internal consistency reliability (α) but few reports included reliability testing for stability. When retest reliability was asserted, time intervals and correlations were frequently not included.

Conclusions: Planning for psychometric testing through design and reducing nonrandom error in measurement will add to the reliability and validity of instruments and increase the strength of study findings. Underreporting of validity might occur because of small sample size, poor design, or lack of resources. Lack of information on psychometric properties and misapplication of psychometric testing is common in the literature. 

The foundation of all rigorous research designs is the use of measurement tools that are psychometrically sound. Confirmation of the validity and reliability of tools is a prerequisite for assuring the integrity of study findings. Knowing what type of psychometric properties to look for, what statistical tests mean, or what type of evidence is sufficient can be problematic. A discussion of how to assess the psychometric properties of tools used in nursing research and the presentation of examples of evidence found in the literature may be helpful to new and established investigators and to those hoping to publish research findings.

Background

Traditionally, support for the validity of instruments has been determined by examining construct, content, and criterion-related concepts. For this discussion, validity is defined as the ability of the instrument to measure the attributes of the construct under study. The definition proposed by Trochim (2001) is particularly informative; all
types of validity fall under the broad heading of construct but content and face validity are termed translational (as in translation of the construct); concurrent, predictive, convergent, and discriminant are types of criterion validity (Figure). Reliability is defined as a measure of true scores and includes an examination of stability and equivalence. Reliability is a necessary but not sufficient component of the validity of an instrument; it refers to the ability of an instrument to measure an attribute consistently.

The published literature, including texts and research articles, include extensive definitions of validity and reliability, though many pertain to comprehensive exploration of a single psychometric concept (Polit & Beck, 2006). No recent articles were found about how the concepts have been applied in nursing research along with useful guidelines for critiquing published data, designing studies to assess psychometric properties of instruments, and incorporating psychometric data into manuscripts for publication.

Therefore, the purpose of this article is to review the concepts of reliability and validity and provide examples in the healthcare literature to show how these concepts have been documented. Specific aims are to: (a) identify strengths and weaknesses in psychometric evaluation of instruments; (b) provide suggestions for improving the validity and reliability of instruments; and (c) report guidance from editors of nursing research journals for incorporating psychometric data into manuscripts. Guidelines for the evaluation of specific psychometric properties of instruments are offered.

Methods

The CINAHL, MEDLINE, and PsycINFO, electronic databases were searched using key words and concepts associated with validity and reliability including Cronbach’s coefficient alpha, content validity index, and correlation coefficients. Articles were eligible for inclusion if they were published in the past 5 years and included reports of psychometric properties, including statistical tests and levels of significance. Articles covering a broad array of specialty areas were retrieved. Studies indicating strong psychometric properties were identified as well as those indicating adequate validity and reliability of instruments but with little supporting evidence. The goal was to find at least one good example of psychometric evaluations for each validity and reliability concept.

Validity

Construct Validity

Construct validity is the degree to which an instrument measures the construct it is intended to measure (Cronbach & Meehl, 1955). It is supported if the instrument’s items are related to its operationally defined theory and concepts. For example, an instrument intended to measure anxiety is constructively valid if all items in the tool have the capability to exclusively measure concepts that are theoretically and structurally related to anxiety. However, if the instrument also has the capability to measure closely related concepts such as stress or depression, it might not have adequate construct validity as a measure of anxiety. An instrument might be “construct valid” but not capable of measuring the intended construct. Close attention to construct validity is a fundamental requirement in nursing research studies, because the study variables are often abstract and thus difficult to measure quantitatively (Hulley et al., 2001). Validity is not an all-or-none proposition. Support for construct validity comes through evidence from each subcategory.

A variety of ways exists to evaluate the construct validity of an instrument, including contrasted groups, hypothesis testing, factor analysis, and the multitrait-multimethod (MT-MM) approach. The MT-MM approach will be described under criterion-related validity. In the contrasted-groups approach, two groups known to be high and low in the construct being measured are sampled. The mean scores
of the two groups should differ significantly in the expected direction if the instrument is valid. An example would be a functional status measure in a group of age- and gender-matched patients with severe chronic obstructive pulmonary disease (COPD) compared to healthy controls. The patients with COPD would be expected to score lower than would the healthy group. If this is the case, the tool is capable of differentiating between the contrasting groups of participants.

Hypothesis testing is based on a theoretical framework and indicates the expected direction of scores on the measure. Construct validity is supported if the scores reflect the framework as hypothesized. Suppose an investigator develops a tool designed to measure nurse burnout. The investigator hypothesizes that nurses working in units with high patient acuity have a higher incidence of burnout than do nurses on units with low patient acuity. To test the hypothesis, the tool is administered to a sample of nurses working in an intensive care unit (ICU) and a medical-surgical unit. A higher level of burnout among ICU nurses than among the medical-surgical unit nurses might be considered evidence for construct validity of the newly developed tool.

Factor analysis is a statistical method commonly used during instrument development to analyze relationships among large numbers of variables. A factor is a combination of test items that are believed to belong together. Related items define a part of the construct and are grouped together. Unrelated items do not define the construct and should be deleted from the tool (Munro, 2005). Exploratory factor analysis (EFA) helps investigators identify the various factors that define the construct. EFA is used to identify the greatest variance in scores with the smallest number of factors, expressed statistically as an eigenvalue >1.0. Confirmatory factor analysis (CFA) generally follows EFA and includes theoretical knowledge to further test the construct validity of a tool. CFA validates the extent to which the statistical model fits the actual data (Waltz, Strickland, & Lenz, 2005). Psychometric experts disagree on the number of participants necessary for factor analysis, but generally a minimum of five per variable is recommended (Munro, 2005).

Heo and associates (2005) examined the psychometric properties of the Minnesota Living with Heart Failure Questionnaire (LHFQ). Although initial testing indicated the reliability and validity of the LHFQ, they determined through interitem correlations and factor analysis that several items did not enhance the psychometric soundness of the instrument. Five items (working to earn a living, sexual activities, hospitalization, medical costs, and side effects from medications) were deleted from the instrument. The resulting item-total correlation coefficients improved to greater than .30, and interitem correlation coefficients were greater than .30 and less than .70. EFA indicated a two-factor construct and explained 50% of the variance with factor loadings greater than .40 on 15 of the 16 items. Convergent validity was shown in the high correlations between the LHFQ and the New York Heart Association (NYHA) functional classification.

Translational Validity

Face validity. Face validity means that the instrument looks, on the face of it, as if it measures the construct of interest. It is the easiest way to claim support for construct validity and, as a result, is frequently reported in the literature. Face validity is a subjective assessment, so it is the weakest form of validity (Trochim, 2001). It is not a form of validity in the true sense of indicating that the tool measures the construct of interest; however, it does provide insight into how potential participants might interpret and respond to the items. Investigators seek experts (Netemeyer, Bearden, & Sharma, 2003) or lay people (Schultz & Whitney, 2005) to review the instrument for grammar, syntax, organization, appropriateness, and confirmation that it appears to flow logically.

Content validity. Content validity is indicated if the items in the tool sample the complete range of the attribute under study. To develop a pool of scale items, a researcher first defines the construct of interest and its dimensions by searching the literature, seeking expert opinions, performing population sampling (Carmines & Zeller, 1979; Netemeyer et al., 2003), or through qualitative research (Hogan, Greenfield, & Schmidt, 2001). A panel of content experts is then asked to review the potential scale items and validate that they are appropriate indicators of the construct (Schulz & Whitney, 2005). The early stages of instrument development should include the largest pool of potential items possible, which can be reduced, based on content reviews (Netemeyer et al., 2003).

Content analysis can be difficult when the construct of interest is highly abstract, but methods have been proposed to quantify the process (Lawshe, 1975; Lynn, 1986). Lawshe proposed a method wherein experts rate each item on a 3-point scale.

With the information in Table 1, the content validity ratio (CVR) can be computed with scores ranging from 0 (no agreement) to 2 (perfect agreement). A table of minimum CVR scores for item inclusion was developed based on a one-tailed test at the .05 level of significance (Lawshe, 1975). The score for the entire instrument, called the Content Validity Index (CVI), can be calculated by determining the mean CVR for all of the retained items.

Lynn (1986) proposed a two-step method for determining content validity. In the developmental stage, individual items are evaluated by content experts. A 4-point scale, ranging from 1—not relevant to 4=very relevant and succinct, is used for determining whether items should be retained or rejected (Table 1). In phase two, the judgment phase, a CVI is computed on the remaining items. The summary CVI is the proportion of experts whose endorsement is required to establish content validity beyond the .05 level of significance. Complete agreement must exist among the reviewers to retain an item with seven or fewer experts (Lynn, 1986). Establishing strong support for content validity is a challenge because many reviewers are needed to avoid an inflated estimate of validity that often results when experts endorse most items.
Psychometric Toolbox

Table 1. Two Methods of Calculating the Content Validity Ratio (CVR) and the Content Validity Index (CVI)

<table>
<thead>
<tr>
<th>Author</th>
<th>Lawshe (1975)</th>
<th>Lynn (1986)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rating scale</td>
<td>Scale used for rating items:</td>
<td>Scale used for rating items:</td>
</tr>
<tr>
<td></td>
<td>0: Not necessary</td>
<td>1: Irrelevant</td>
</tr>
<tr>
<td></td>
<td>1: Useful</td>
<td>2: Extremely relevant</td>
</tr>
<tr>
<td></td>
<td>3: Essential</td>
<td>4: Extremely relevant</td>
</tr>
<tr>
<td>Calculations</td>
<td>To calculate the CVR (a score for individual scale items):</td>
<td>This item would not meet the .83 level of endorsement required to</td>
</tr>
<tr>
<td></td>
<td>CVR = ( \frac{n-e}{N/2} )</td>
<td>establish content validity using a panel of 6 experts at the .05 level of</td>
</tr>
<tr>
<td></td>
<td>Note: ( n ) = the number of experts who rated an item as “essential.”</td>
<td>significance. Therefore, it would be dropped.</td>
</tr>
<tr>
<td></td>
<td>( N ) = the total number of experts. Example: 8 of 10 experts rated an item as essential. The CVR would be (8 − 5)/5 = .60.</td>
<td>The CVI for the entire scale is the proportion of the total number of items</td>
</tr>
<tr>
<td></td>
<td>The CVI is the mean CVR for all retained items.</td>
<td>deemed content valid. Example: 77 of 80 items were deemed content valid. The CVI would be: 77/80 = .96</td>
</tr>
<tr>
<td>Acceptable range</td>
<td>Depends on the number of reviewers</td>
<td>Depends on the number of reviewers</td>
</tr>
</tbody>
</table>

An innovative method of content validation through the use of experiential experts was described by Stewart, Lynn, and Mishel (2005). Data gathered through qualitative inquiry were used to develop a 22-item scale intended to measure uncertainty in children and adolescents. Six children, aged 8 to 16 years, who were undergoing chemotherapy for the treatment of cancer, were asked to evaluate the items and rate them as “good,” “okay,” or “bad” to indicate their agreement with the experience of feeling uncertain. Items rated as “good” or “okay” by at least five of the six reviewers were retained. Input from four content experts was then sought to further support the validity of the items, because children had not previously been used as content experts. The nurse experts used Lynn’s (1986) criteria for establishing content validity. All items were rated as 3 (relevant with minor revisions) or 4 (relevant as presented) and were retained. The nurse experts made no other suggestions for changes. Studies were identified in the literature that included five or fewer content experts, with satisfactory CVIs. This result would require a perfect agreement among all reviewers if Lynn’s criteria were followed.

Criterion Validity

Criterion-related validity pertains to evidence of a relationship between the attributes in a measurement tool with its performance on some other variable. For example, responses on a healthy lifestyle tool could be validated by correlations with biological measures such as lipid levels, oxygen consumption, or biometric measures (the criteria). Criterion variables might be any established standard of the phenomenon of interest. Criterion-related validity is indicated when measures on the predictor and the criterion variables are correlated and the strength of the correlation substantially supports the extent to which the instrument estimates performance on each criterion (Waltz et al., 2005). Shultz and Whitney (2005) noted that criterion contamination, i.e., error in measurement of the criterion, is a threat to validity. Criterion contamination leads to an exaggerated correlation between instrument and criterion variables, and thus to a faulty criterion-related validity estimate (Waltz et al., 2005). For instance, criterion scores might be influenced if investigators have knowledge of participants’ performance on instruments.

Predictive validity. The degree to which test scores predict performance on some future criterion is referred to as predictive validity (Shultz & Whitney, 2005). High correlations between the original measure and criterion variables reinforce the conclusion that the tool is a valid predictor of the specified criteria. While reviewing research designs in the nursing literature, we often found limitations in psychometric measurement when critically analyzing predictive criterion-related validity. Some limitations were related to the misapplication of the psychometric concept and analyses; others were related to limitations inherent in the data available for measurement. Nibert, Young, and Adamson’s (2002) study is a model case showing the inability to measure predictive criterion-related validity because of limitations in the type of data available for analysis, in this case dichotomous rather than interval data. The study was designed to explore the extent to which the Health Education Systems, Inc. (HESI) Exit Exam scores predicted both registered nurse (RN) and practical nurse (PN) success rates on the National Certification Licensure Examination (NCLEX). Also of interest was the degree of risk for failure of the NCLEX-RN and NCLEX-PN associated with specific scoring intervals.

Before 1994, the NCLEX-RN and NCLEX-PN Exams were norm-referenced, paper and pencil examinations. Each candidate received a numeric score indicating either a pass or fail result. In 1994, the National Council of State Boards of Nursing (NCSBN) made a policy decision to change to
computerized adaptive testing (CAT) and to report only a pass or fail result. Although only a pass or fail result is reported to candidates and institutions of higher learning, a data metric underlies the scoring and overall performance on the NCLEX-RN and NCLEX-PN (T. R. O’Neill, personal communication, July 28, 2006). The CAT was designed to spread the difficulty of test items over the entire exam and to estimate the degree of difficulty after each question has been answered, based on the response to the question (O’Neill, Marks, & Reynolds, 2005). Limiting the NCLEX-RN and NCLEX-PN results to pass or fail has affected the ability of researchers to perform correlation analyses (and establish predictive validity) because the results were reduced from a ratio level to a nominal level.

**Concurrent validity.** Concurrent criterion-related validity is confirmed when scores on a tool are correlated to a related criterion at the same point in time (Carmines & Zeller, 1979). For example, concurrent criterion-related validity was supported in a study to examine the extent to which the Oral Mucositis Index-20 (OMI) indicated changes associated with developing and resolving oral mucositis in patients undergoing bone marrow and stem cell transplant (McGuire et al., 2002). The criterion variable, mucous membrane and tongue (mucosal and anatomic site of changes), was identified by the researchers as the most relevant for diagnosis of mucositis. Significant correlations (p=.01) were found between total OMI-20 scores and the criterion variable of clinically observed oral inflammation at four key times in the study, including Time 5: \( r=.398 \), Time 8: \( r=.545 \), Time 11: \( r=.644 \), and Time 14: \( r=.622 \). Results indicated concurrent validity because the OMI-20 was not only able to show changes associated with the development or resolution of oral mucositis, but also show distinctions between the various stages of mucositis severity.

**Convergent and discriminant validity.** Convergent and discriminant validity are related concepts. Think of these concepts as being at opposite ends of a spectrum. Convergent validity is a correspondence or convergence between constructs that are theoretically similar. The interitem correlation coefficients would be high in an instrument that has convergent validity. Conversely, discriminant validity is the instrument’s capability to differentiate or discriminate between constructs that are theoretically different. Interitem correlation coefficients would be low in an instrument with discriminant validity. A researcher could conclude convergent validity of a tool designed to measure children’s pain if interitem correlations are high or if similar results are achieved with other validated tools used to measure children’s pain. Convergent validity would be supported if scores were similar. Discriminant validity would be supported if correlations were low between the pain tool and a measure of comfort or well being.

The multitrait-multimethod (MT-MM) approach is another method for identifying convergent and discriminant validity. Interestingly, this approach appears in psychometric texts but it is rarely reported in the literature, perhaps because the approach is complex and it requires extensive testing and resources. The MT-MM method can be used whenever two or more constructs are being measured with two or more methodologies. A matrix shows the degree of correlation and the relationships between traits. Different measures of the same construct should correlate highly with each other (converge) and different constructs should show low correlation with each other (discriminate).

Johnson and Rogers (2006) explored the construct validity of the Purposeful Action Medication-Taking Questionnaire (MTQ: Purposeful Action), an instrument designed to measure the reasons people decide to accept medication treatment. This 12-item instrument has two subscales: treatment benefits and medication safety. CFA and MT-MM methods were used to examine construct validity. CFA indicated that the treatment benefit and medication safety subscales were a good model fit, as shown by a relative chi-square value of 2.09. The MT-MM confirmed convergent and discriminant validity by showing positive correlations between the Hamilton Health Belief Model subscales Perceptions of Hypertension and Individual Health and the MTQ: Purposeful Action, Treatment Benefits Subscale, as would be expected. Discriminant validity was confirmed by a low correlation between the MTQ: Purposeful Action and the Lifestyle Busyness Questionnaire (LBQ). The LBQ is used to measure a patient’s consistency in keeping to general routines or level of busyness, and it contrasts with a purposeful action.

Measures of validity can be strengthened by prior predictions of how the instrument should perform if it is actually measuring the construct of interest (Oldridge, Lim, & Guyatt, 2003). Labeling a correlation coefficient as “high” or as offering sufficient evidence for validity is difficult because criterion standards are often difficult to identify; but generally correlations above .50 are infrequent (Oldridge et al., 2003). DeVon and Ferrans (2003) examined the psychometric properties of four quality-of-life tools and found that correlations in support of validity ranged from .11 to .88. Fifty-three percent of the correlations were higher than .50. Confidence intervals are also useful in making decisions about the strength of correlations. The confidence interval is the range in which the truth (or point estimate) most plausibly lies (DiCenso & Guyatt, 2005). Therefore, a moderate correlation might be a strong evidence of validity if the score (point estimate) falls within a narrow confidence interval.

In the literature review, we identified studies in which construct validity was supported by strong item-total correlations (> .30), inter-item correlations (.30 to .70), and factor loadings (> .40). However, we also identified correlations as low as .11 that authors considered adequate to label a tool valid. Adequate content validity is frequently claimed but rarely is the method specified. This lack of reporting shows the importance of critiquing the quality of the data and the conclusions.
Reliability

Reliability is a necessary but not sufficient component of the validity of an instrument. It pertains to the ability of an instrument to consistently measure an attribute. If a pulse oximeter showed a reading of 98% for a healthy person one minute and a reading of 68% the next minute, the oximeter would be considered unreliable, assuming no change in the person’s oxygen saturation. According to classical measurement theory, reliability can be expressed in the following equation: Obtained score = true score ± error score.

The true score can never be known because no measure is perfect. The error score cannot be known either; however, the amount of both random and systematic error can often be controlled for. Stability reliability is tested when the attributes under study are not expected to change. Equivalence reliability indicates whether all items in the tool reliably measure the attributes and if participants score similarly on like measures.

Stability Reliability

Test-retest. Test-retest reliability is estimated by administering the same test to the same group of respondents at different times. The correlation between the two scores, and often between individual questions, indicates the stability of the instrument. Time intervals between the original test and the retest are somewhat controversial. Studies were found with retesting as early as a few hours and as long as 6 months after baseline testing. However, 2 weeks to 1 month is the generally accepted time interval for retesting (Waltz et al., 2005). The time interval should be long enough that respondents do not remember their original responses, but not long enough for their knowledge of the material to have changed. As a rule, the longer the time, the lower the reliability (Trochim, 2003) and the more likely that knowledge or attitudes actually have changed.

Test-retest reliability is relevant for cognitive and trait scales that are not expected to change over time. It is not appropriate for states that are expected to change over time, such as attitude, mood, or knowledge following an intervention. The major weaknesses of test-retest measures of reliability are the memory reactivity effects. Respondents’ memories tend to decline as the time between tests lengthens. The reactivity effect is a change in test response related to the observer’s presence (Polit & Beck, 2004). Psychometric properties of the 15-item Bakas Caregiving Outcomes Scale (BCOS) were explored in a study of 147 family caregivers of stroke survivors. Test-retest reliability at 2 weeks was satisfactory with intraclass correlation coefficients of .66 (confidence interval=.42 to .81; Bakas, Champion, Perkins, Farran, & Williams, 2006).

Many reports in the literature included test-retest reliability coefficients, but not the time between test administrations. In addition, authors often did not specify their criteria for minimal correlations, making it difficult for readers to evaluate the claims of the authors and the rigor of the findings.

Equivalence Reliability

Coefficient alpha. Cronbach’s alpha coefficient is the most frequently used statistic to show internal consistency reliability; it is the most widely used by nurse researchers (Brink & Wood, 1998; Polit & Beck, 2004). Internal consistency indicates how well the items on a tool fit together conceptually. Coefficient alpha is the only reliability index that can be performed with one test administration, thus requiring much less effort than either the split-half, alternative form, or retest methods (Ferketich, 1990). Coefficient alpha is sample specific; in other words, it is a measure of the internal consistency for the test responses from the current participants. Therefore alpha coefficients should be computed each time the test is administered (Waltz et al., 2005). Higher coefficient alpha values can be achieved by adding items, provided they are correlated (DeVellis, 2003; Nunnally & Bernstein, 1994). If items are not correlated, the value of alpha is reduced. Inflated alpha values are achieved when computed for an entire scale, i.e., composed of two or more subscales. In this case, coefficient alpha should be computed for each subscale rather than for the entire scale (Nunnally & Bernstein, 1994).

A coefficient alpha of .70 is acceptable for new scales (DeVellis, 2003). Opinions differ about the ideal value of the coefficient alpha for instruments in clinical use, such as the Braden Pressure Sore Risk Scale or the Acute Physiology and Chronic Health Evaluation (APACHE II) severity of disease scale. Several authors have recommended that the coefficient alpha should be minimally .90, with an ideal value of .95 (Bland & Altman, 1997; Nunnally & Bernstein, 1994; Polit & Beck, 2004). Some authors (DeVellis, 2003) have contended that alpha coefficient values over .90 represent redundancies and indicate that the tool should be shortened. DeVellis (2003), however, warned that some redundancy is integral to internal consistency. Schim, Doorenbos, and Borse (2005) studied cultural competence variables, in a sample of 145 hospital-based care providers, using the 26-item Cultural Competence Assessment tool. Alpha for the overall tool was .89 and were .76 (cultural awareness) and .93 (cultural competence behavior) for the subscales. The overall alpha exceeds the minimum alpha for a new tool and nearly reaches the recommended level for clinical usage.

Alternative and parallel forms. Unlike retest reliability, alternative forms reliability pertains to different versions of an instrument to determine reliability of scores. Alternative forms can prevent participants from using knowledge gained from the first test in answering questions during subsequent administrations. Also known as parallel forms (DeVellis, 2003; Waltz et al., 2005), alternative forms reliability pertains to scores from two tests, each with different items from an item “pool” to test the same concepts. Both versions of the instrument must measure the same phenomenon and have scores with approximately equal means, variances,
The subscale correlations between Forms A and B ranged from .69 to .75 for chance locus of control. The investigators described the equivalence of the testing methods and reliability for interpretation of the psychometric data. Data do not dictate results; investigators are responsible for interpreting the data based on theoretical constructs and recognition of elements of bias.

Editors of highly respected nursing journals reporting original research frequently have addressed issues of validity and reliability in editorials. We interviewed four editors of widely read journals and asked them to identify specific issues related to the reporting of psychometric data. We were interested in their suggestions for enhancing the quality of psychometric reporting and increasing chances for publication. Three editors gave extensive comments about validity and reliability and their responses are included here.

Dr. Sue Hegyvary, editor of the Journal of Nursing Scholarship, commented that “Validity and reliability are basic requirements for research. Good articles include such information but others do not, to the detriment of those articles, because findings are not credible unless the data are credible.” She cautioned that, “Authors need to know that editors and reviewers must be diligent about the quality of data and analyses, and simply adding alpha coefficients in the manuscript is not sufficient. Authors often report alphas as low as .20 to .30 and proceed with the planned analysis as if the lack of reliability was not a concern. If the measure is not reliable, the author has to decide how to proceed, either deleting it from the analysis or offering an explanation for the insufficient reliability. If a nonreliable measure is central to the analysis, the manuscript probably will be rejected” (S.T. Hegyvary, personal communication, August 9, 2006).

Manuscripts are not usually rejected solely because of lack of information on reliability and validity, according to experts on the Psychometric Toolbox, because the modern researcher has access to the plethora of statistical software and novel psychometric tools. Munro (2005) said that alternative forms administered at the same time would show instrument equivalence, but stability could also be established by administering the instruments on two occasions.

Alternate forms reliability was described by Hubley and Wagner (2004) in a study to examine two different forms of the Multidimensional Health Locus of Control Scale (MHLCS). All participants completed Forms A & B of the MHLCS at the same sitting as a single 36-item test with approximately half (n = 124) receiving Form A followed by Form B and the remainder (n = 121) receiving Form B then Form A. The means and standard deviations of the MHLCS subscale scores across forms were comparable and the mean scores on the retest were nearly the same for both forms. The subscale correlations between Forms A and B ranged from r = .80 (p < .001) for powerful others locus of control, r = .75 (p < .001) for internal locus of control subscale, and r = .69 (p < .001) for chance locus of control. The investigators concluded that the results did not show parallel forms reliability because they wanted a correlation of at least .90 between forms.

Alternative or parallel forms reliability has been frequently reported in the literature, contrary to established psychometric definitions. Researchers have claimed parallel forms reliability of paper-and-pencil and Web-based tests, even though both had identical items. In these examples, investigators described the equivalence of the testing methods, not the equivalence of the instrument. In another study researchers measured risk behavior with test-retest and parallel forms reliability synonymously. Alternative forms reliability was also erroneously reported when administering long- and short-versions of a scale at the same time.

Discussion

To increase the validity of instruments, we recommend an assessment of content validity using the process proposed by Lynn (1986) or Lawshe (1975), while also cautioning that content validity is related to the number of underlying dimensions, the number of items, and the theoretical framework. Engaging adequate numbers of content experts (≥ 7) can help account for inflated CVIs based on high levels of interrater agreement. For a comprehensive review of techniques for measuring the CVI see Polit & Beck (2006). In addition, all but the most novel instruments can be tested against similar tools and provide support for criterion (concurrent) validity. If sample size is large enough (> 5 participants per variable), factor analyses can be done to examine construct validity of the tool (Munro, 2005).

To increase reliability of instruments we recommend computing alpha coefficients for each administration of a tool and making the tool as concise as possible without sacrificing necessary content. These are minimum standards that can be applied to every study. Test-retest procedures should be done if a stable attribute is measured. This procedure requires additional resources but returns good value, especially if the tool has potential for additional or long-term use.

A summary of recommendations is given in Table 2. Reporting psychometric properties of a tool based on prior study is a starting point for determining if the tool is useful for the questions and population under study, but it is usually not sufficient. Investigators must set parameters for decisions on reliability and validity and assume responsibility for interpretation of the psychometric data. Data do not dictate results; investigators are responsible for interpreting the data based on theoretical constructs and recognition of elements of bias.
<table>
<thead>
<tr>
<th>Psychometric concept and definition</th>
<th>Statistical test</th>
<th>Accepted standard</th>
<th>Common errors</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Validity</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Construct (overall)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Extent to which an instrument measures the construct under study</td>
<td>Exploratory Factor Analysis (EFA); Confirmatory Factor Analysis (CFA); Principal Components Analysis</td>
<td>Eigenvalues &gt;1.0; Generally, factor loadings &gt;.40; Approximately five subjects per variable or number of subjects exceeds number of variables by 50</td>
<td>Finding same number of factors as items on the tool; Too few participants for number of variables</td>
</tr>
<tr>
<td>Translational</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Face: the instrument, on the face of it, appears to measure the construct</td>
<td>None; tool &quot;looks&quot; like valid measure of the construct</td>
<td>None: not considered a &quot;true&quot; measure of validity. Tool is accepted at face value</td>
<td>Subjective measure; no criteria for acceptance</td>
</tr>
<tr>
<td>2. Content: Extent to which items in the tool sample the complete range of the attribute under study</td>
<td>Content Validity Ratio or Content Validity Index: ( CVR = \frac{\sum Z_x Z_y}{N} )</td>
<td>Depends on the number of expert reviewers</td>
<td>Tools with low CVR or CVI are called content valid; Inflated CVI is possible with high levels of agreement</td>
</tr>
<tr>
<td><strong>Criterion</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relationship linking the attributes in a tool with the performance on a criterion</td>
<td>Pearson Product Moment Correlations</td>
<td>Substantial and high: ( r \geq .45 ) is recommended by many authors</td>
<td>Inappropriate criterion chosen for comparison</td>
</tr>
<tr>
<td>1. Concurrent</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scores on the measurement tool are correlated to a related criterion at the same time</td>
<td>High Pearson Product Moment correlations</td>
<td>Substantial and high: ( r \geq .45 )</td>
<td>Tools with low correlations are labeled criterion valid</td>
</tr>
<tr>
<td>2. Predictive</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The degree to which test scores predict performance on some future criterion</td>
<td>High Pearson Product Moment correlations</td>
<td>Substantial and high: ( r \geq .45 )</td>
<td>Tools with low correlations are labeled criterion valid</td>
</tr>
<tr>
<td>3. Convergent</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Extent to which different measures of the same construct correlate with one other</td>
<td>High Pearson Product Moment correlations</td>
<td>Substantial and high: ( r \geq .45 )</td>
<td>Tools with low correlations are labeled criterion valid</td>
</tr>
<tr>
<td>4. Discriminant</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Extent to which measures of different constructs correlate with one other</td>
<td>Low Pearson Product Moment correlations</td>
<td>( \leq .45 )</td>
<td>Tools with high correlations are labeled criterion valid</td>
</tr>
<tr>
<td><strong>Reliability</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Internal consistency</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Extent to which performance on one item in an instrument is a good indicator of performance on any other item in the same instrument</td>
<td>Coefficient alpha (Gronbach’s alpha)</td>
<td>( \geq .90 ) for clinical tools; ( \geq .70 ) for research tools; Guideline based on underlying dimensions of the construct</td>
<td>Tools with lower coefficient alphas are called reliable</td>
</tr>
<tr>
<td>2. Test-retest</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Extent to which an instrument measures stable characteristics at two separate times</td>
<td>Interclass Correlation Coefficients; Pearson Product Moment correlations; ( t \text{test} )</td>
<td>High correlations; generally ( r \geq .70 ); No statistically significant difference in scores from pre to posttest</td>
<td>No report of statistical tests, level of significance, or confidence intervals</td>
</tr>
<tr>
<td>3. Alternative forms</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Extent to which different forms of an instrument yield comparable results when given to the same sample during a single administration</td>
<td>Pearson Product Moment Correlation Coefficient; Spearman Brown ( r \text{test} ) length has been changed (Both versions of the instrument must have equal means, variances, and alpha coefficients)</td>
<td>High correlations, generally ( r \geq .70 )</td>
<td>No report of statistical tests, level of significance, or confidence intervals</td>
</tr>
</tbody>
</table>
to Dr. Molly Dougherty, editor of Nursing Research, “If a manuscript is valuable and advances nursing knowledge in a priority area, the author will be required to clearly state the limitations of the instruments and the related limitations of the research.” She also noted that, “On occasion authors report retest reliability but do not explain how the data were collected.” In that case, readers cannot evaluate the reliability test (M. C. Dougherty, personal communication, August 28, 2006).

Dr. Vicki Conn, editor of the Western Journal of Nursing Research, said that addressing construct validity is the first critical issue in whether the instrument measures the intended construct, and it affects the chances for publication. She added that, “Manuscripts have been rejected because of insufficient evidence that they measure the intended construct. This is especially true of investigator-developed instruments without any psychometric testing.” She further noted that, “Often instruments that have been tested in the same population might not need further testing, but further psychometric testing is necessary if differences exist between the study population and the population sampled when the instrument was developed and tested. Psychometric properties of tools used in the current study are necessary to report because they are specific to the sample of participants. Readers of published studies need to know that the results are trustworthy” (V.S. Conn, personal communication, August 21, 2006).

These editors concurred that support for the validity and reliability of tools used in nursing research is necessary for readers to have confidence in study findings, and credible psychometric data are a prerequisite for publication.

Conclusions

The literature is replete with examples of valid and reliable tools used in nursing research as well as examples of misapplication of psychometric testing. Reports of studies frequently included information about content validity, but sometimes fewer than five experts reviewed items. Criterion validity was rarely reported, and errors in measurement of the criterion were identified. Construct validity remains underreported. Most researchers have reported internal consistency reliability (α), but few articles indicate any type of stability reliability testing. When retest reliability was asserted, time intervals and correlations were frequently not included.

Researchers and consumers of research might be comforted by structure and guidelines in research reports. Unfortunately, the confidence derived from following guidelines can be a false confidence. The strength of psychometric tests is dependent on many important factors such as underlying theory, the consequences of weak measures, and the implications of the findings. Therefore, we urge less focus on rigid adherence to statistical rules and more emphasis on evaluating the strength of psychometric tests, planning for future studies, and publishing data-based articles.

References

Psychometric Toolbox


