5-2015

Psychometric instrumentation: reliability and validity of instruments used for clinical practice, evidence-based practice projects and research studies

Ann Mayo RN, DNSc, FAAN
University of San Diego, amayo@sandiego.edu

Psychometric Instrumentation: Reliability and Validity of Instruments Used for Clinical Practice, Evidence-based Practice Projects and Research Studies

Ann M. Mayo, RN; DNSc; CNS; FAAN

Professor

Hahn School of Nursing & Health Science

Beyster Institute of Nursing Research

University of San Diego

Psychometric Instrumentation: Reliability and Validity of Instruments Used for Clinical Practice, Evidence-based Practice Projects and Research Studies

Abstract

Clinical Nurse Specialists and other advanced practice nurses (APNs) are relied upon for their expert data based decision-making to ensure excellent clinical practice, high quality evidence-based practice projects, and efficacious research studies. Using measurement instruments to collect data that lack reliability and validity can compromise decision-making leading to deleterious results for patients, nurses, and healthcare organizations. The ultimate goal is to choose reliable instruments that produce valid data so that clinical trends, project evaluations, and research findings are trustworthy.

Determining the quality of a given instruments is most often done through a critique process. Critiquing potential instruments for use is time consuming and requires knowledge of the scientific principles and theories of psychometric instrumentation. Over a number of future articles, select instruments will be critiqued, making relevant reliability and validity information on those instruments available to readers. The current issue of the Clinical Nurse Specialist journal presents this first article in the series. The purpose of this first article is to provide background information the various types of reliability and validity testing that will be discussed across a series of future instrument critiques. This background article may be used as a reference for those subsequent critique articles.

Psychometric Instrumentation: Reliability and Validity of Instruments Used for Clinical Practice, Evidence-based Practice Projects and Research Studies

Background

Reliability and validity are important psychometric properties to be considered by clinical nurse specialists (CNSs) and other advanced practice nurses when selecting measurement instruments for clinical practice, evidence-based practice (EBP) projects and research studies. Instruments must be reliable and produce valid results so that clinical assessments, project evaluations, and research results are trustworthy. For example, decreasing numbers of stage II pressure ulcers in an evidence-based practice project should be due to an actual decrease in stage II pressure ulcers rather than nurses misinterpreting an unclear ranking process on the indicator data collection form. Likewise, in research, differences in mental status scores between males and females should be due to actual differences, not due to the wording of the items that are biased against one gender group.

Ideally, every measurement instrument should undergo some form of psychometric testing before it is utilized in a clinical setting, evidence-based practice project, or research study. Psychometric testing of instruments is the application of specific research methods designed to evaluate the amount of error contained within an instrument (reliability) or within the data produced by using the instrument (validity). Measurement error is an important consideration for reliability and validity. Simply stated, less measurement error equates to better reliability and validity.

A number of specific psychometric tests may be conducted to estimate the reliability and validity. The purpose and type of instrument determine the choice of tests. For

example, some tests are chosen based upon whether an instrument is used to simply provide a score for an individual (i. e, number on a pain scale) or to categorize an individual into a predetermined group based on normed scores (i. e., mild, moderate, or severe cognitive impairment) (Streiner & Norman, 2008). This article will provide background information on reliability and validity and related psychometric testing approaches. Subsequent articles over a number of future issues will present critiques of specific instruments. This current article may be used as a reference for specific tests discussed in those instrument critiques.

Reliability

The reliability of an instrument is evaluated based on its ability to be free from error. Problems with reliability appear when instruments are not stable over time or between users. As a result, the instruments are inconsistent in their performance. Methods for determining reliability estimate how much measurement error is present. And, reliability testing only provides an estimate of reliability; it is impossible to prove the exact extent an instrument is reliable. This is because, theoretically, any score obtained through the administration of a measurement instrument (the observed score) is comprised of two components; the true score and the error score (obtained score = true score + error score).

Sources of Error

Inaccurate items or items left out of the development of an instrument diminish reliability. The true score of an instrument could only be known if every possible item in the domain of interest could be included in an instrument. However, this is not practically possible. Therefore, this is one source of error that cannot be totally eliminated (McDowell, 2006, Waltz, et al 2010).

It is assumed that the reliability of an instrument increases with the number of items sampled from that possible universe of items. In other words, the more items contained within the instrument, the higher the reliability; the fewer number of items, the lower the reliability (Waltz, et al 2010). Therefore, some instrument developers attempt to increase reliability by increasing the number of items. A balance must be struck however, so that the instrument is not perceived to be over burdensome with too many items.

Another source of error can occur because individual patients, raters, or participants completing the instrument could be tired or distracted. If instruments were administered enough times, over and over again, these random errors would cancel each other out. However, in practical terms, numerous administrations are not possible. Therefore, this is yet another source of error that could be assumed to be a part of every observed (obtained) score on an instrument (McDowell, 2006).

Consistency of Instrument Performance

Consistency of instrument performance is an important concept related to reliability. Reliability, in terms of consistency, simply means that similar scores are obtained between different time frames or between different raters or users. As long as conditions are the same, similar scores should be produced from an instrument time over time. For example, determining if a patient is a fall risk should produce the same rating day after day if the condition of the patient has not changed. Re-administering an instrument in such a way would be an example of test-retest reliability in a psychometric instrumentation study. Similarly, inter-rater reliability determines if two raters (i. e., nurses) of fall risk obtain similar scores when assessing the same patient at the same time. A well-designed instrument will have high inter-rater reliability.

Interestingly, reliability has very little to do with if the users of the instrument correctly interpret the meaning of the items. Users may perceive the meaning of the items incorrectly; however, if they consistently assign the same meaning to those items their inaccurate scores would remain consistent and the instrument would be deemed reliable. Understanding the meaning of the items has more to do with the validity of the instrument. Therefore, while reliability is necessary for a strong instrument, its presence does not mean an instrument will provide valid (or accurate) scores.

There are particular *types* of instrument consistency and these include stability, equivalence, internal consistency, and consistency of ratings. Psychometric testing can determine the degree an instrument is stable, equivalent to another reliable instrument, and has internal consistency. Consistency of rating is an important reliability estimate when instruments are used to rate behaviors or objects. Instruments for which a mean score can be calculated typically use correlational statistics for reliability testing. Instruments designed to categorize a concept of interest into groups such as stage I/II or high/medium/low require different statistical tests such as percent agreement, Kappa, or Spearman rho

Stability. Stability of an instrument means that across repeated administrations (when nothing changes in the individuals being measured or administration procedure), the scores should remain consistent. Only when an instrument has been determined to be stable should it be used to actually measure change such as in intervention studies. Additionally, information about instrument stability is also important to consider when a clinical indicator or research variable is measured using a standardized instrument for which norms have been set (Walker, 2010, et al). Examples of such instruments include

health literacy tests or tests of cognitive ability. An unstable instrument would not be capable of generating reliable norms. Nor would an unstable instrument be able to reliably classify patients or research participants into groups. When an unstable instrument misclassifies a patient or research participant serious consequences can be the result.

Psychometric testing to determine stability (test-retest) involves using the same participants while administering the same instrument at different times, usually twice. The interval between the testing times is determined by the nature of the measure. In instances when rapid change in a condition is possible, shorter internals would be more appropriate for psychometric testing of stability. When mean scores for an instrument can be calculated, psychometric data is analyzed using a Pearson product-moment correlation (Waltz, et al 2010). A correlation of .70 determines that an instrument has acceptable stability. Prior to the Pearson product-moment correlation, a paired t-test can be used as a preliminary screen to verify that there is not a significant difference in the mean scores between the first testing period and the second testing period. When the purpose of an instrument is to categorize objects or person regarding the concept of interest, percent agreement or a Kappa is the appropriate psychometric analytical test (Waltz, et al 2010).

Equivalence. A parallel forms procedure is used to determine equivalence reliability. Determining equivalence is important for newly developed instruments. A newly developed instrument may be compared to an older gold standard instrument in order to determine equivalence reliability. Like the test-retest procedure, a single group of participants is used for the psychometric testing. However in this instance, the group is provided with two different, but assumed equal instruments, at the same time. The same statistical tests that are used to determine stability are also used for equivalence testing.

Theoretically, high reliability coefficients indicate that the two "forms" sample the universe of all possible items equally well. In other words, the items on each form are considered equivalent. A newly developed instrument can be used with higher confidence when it has been determined to be a reliable instrument through equivalence testing.

Internal Consistency. Prior to using an instrument it is important to know if all of the items in an instrument are measuring the same concept (inter-correlated). If an instrument is designed so that all of the items are measuring the same concept, the item scores should be correlated (or associated if the level of measurement is categorical). Internal consistency is based on the correlations or associations between different items on the same measure or between different subsets for larger instruments.

Data are analyzed for internal consistency using an alpha coefficient, a test of item inter-correlation. A high alpha means that each item is a good indicator of the other items. When instruments are used for research purposes it is recommended that alphas should be 0.70 to 0.80. For clinical purposes, alphas should be at least 0.90 (Bland & Altman, 1997). When an instrument's data are dichotomous (e. g., yes/no, true/false) then Kuder-Richardson (KR 20 and KR 21) are the statistical tests to be used.

The split half method is a less common approach in determining internal consistency. Typically, an instrument's items will be divided in half in order to determine if one half correlates with the other. Essentially this creates two instruments, each with fewer items that the original instrument. Remembering that fewer items in an instrument result in weaker reliability, the Spearman–Brown prophecy formula is the appropriate statistical test to be used to predict reliability after changing the length of an instrument.

Consistency of Ratings. Consistency of ratings or performance can be determined by testing *interrater* and *intrarater* reliability. *Interrater* reliability evaluates the consistency of rating between different raters. It answers the following question: Would the same rating scores have been obtained if a different person had made the assessment or judged the performance? This is an important question to answer when clinical nurse specialists and other advance practice nurses lead teams of nurses in evidence-based practice, use different nurse educators to conduct nurse competency testing, or have different research assistants collect research data. In psychometric testing, the consistency rating of an instrument is determined when two scores are compared that were collected by two equally qualified scorers. For sets of scores with interval or ratio level data, a Pearson product-moment correlation of .70 is considered acceptable reliability. Percentage of agreement, index of scorer consistency (i. e., Kappa), or Spearman's rho is the appropriate statistical test when the instrument scores categorize the concept of interest. Higher percent agreement and Kappa scores (i. e., .60-1.00) indicate that the scorers have good strength of agreement (Stemler, 2004). Finally, when interrater 'variability' is also of concern (i. e., not just how much in agreement people are, but rather do their ratings vary in the same way) then the intra-class correlation is the correct statistic (McDowell, 2006; Waltz, et al 2010).

Intrarater reliability evaluates the consistency of rating between the same rater, but at different times. It answers the following question: Would one specific person obtain the same score at two different times? This is important information to know when clinical or research data is to be collected by the same person on different occasions. The same

psychometric statistical tests can be used to determine intrarater reliability as interrater reliability.

In summary, many times issues with reliability occur because instruments are inconsistent in their performance. The particular *types* of consistency (stability, equivalence, internal consistency, and consistency of ratings among and between raters) can be estimated by statistical tests. CNSs and other advanced practice nurses should pay special attention to reliability estimates if they are using instruments to measure change after implementing a practice change or research study intervention, or categorizing clinical data (i. e., staged pressure ulcers) or persons (moderate vs advanced dementia) into groups.

Validity

Technically, validity is about the *interpretation of scores* generated from an instrument (Furr & Bacharach, 2014). While an instrument may be reliable (i. e., the architecture is strong), it may not be valid when used for certain purposes or with a particular group of respondents (i. e., different ethnic groups). So, when deciding to use an instrument in clinical practice or for an evidence-based practice project or research study, the instrument should be chosen with a specific purpose and a particular group of respondents in mind. For example, a traditional numeric visual analog scale (VAS) used for measuring pain acuity generates scores that would be deemed to have poor validity when used with young children. On the other hand, the FACES pain assessment scale generates scores that have higher validity in young children. Additionally, if the purpose of an instrument is to evaluate pain management (i. e., perception that everything is being done to control pain), then an instrument designed for that purpose should be used. A VAS

would not be appropriate because it measures pain acuity and therefore does not generate highly valid scores for pain management. Rather, a different instrument, one that measures pain management and not pain acuity, should be used if an evaluation of pain management is what is needed.

Because the validity of scores is directly related to a purpose and the population in which the instrument is to be used, clinicians and EBP project leaders must match the instrument to their purpose and population. Caution should be exercised if there is no published evidence that the instrument under review is valid for the intended purpose or in the population of interest. Practice changes implemented based on indicator data for which the validity is unknown can increase the chance for negative patient outcomes as well as the legal liability of nurses and organizations. For researchers, their work is designed to generate new knowledge (usually in new populations) and so the researcher may not know the validity of instruments. Because of this it is recommended that validity be evaluated every time an instrument is used for research purposes. Legal liability can be limited for researchers because research participants are informed they are volunteering for research and provide some form of consent (verbal or written).

Poor validity can be the result of instrument problems (mismatch to purpose or population) that are designed into the instrument or administration procedures and therefore happen every time the instrument is administered. In other words, poor validity is the result of primarily systematic errors. Content validity and other approaches are used to estimate validity.

Content Validity

Content validity determines how relevant the items are to the concept to be measured. A clearly articulated conceptual definition is the first step in assuring the items are clear and completely represent the concept of interest. Next, expert opinion (face validity) and/or formal ratings of item importance, adequacy, and clarity (content validity index) can be used to determine the preliminary quality of items (Fain, 2009; McDowell, 2006; Waltz, et al, 2010). Objectively, content validity is determined by gathering evidence about construct validity. Procedures such as convergent (concurrent and predictive) and divergent validity, as well as, factor and item analysis provide evidence of construct validity (McDowell, 2006).

As with other psychometric properties, construct validity cannot be proven. However, a systematic approach to provide estimates of construct validity begins with information about how that concept is related to other similar concepts and not related to dissimilar concepts. Hypotheses regarding those relationships are tested using correlational and group difference statistics.

Convergent validity. Convergent validity determines if the scores from the instrument of interest correlate to the scores from another instrument already known to measure the same or similar concept of interest (McDowell, 2006). Concurrent and predictive validity are two approaches used to determine convergent validity (Furr & Bacharach, 2014). Concurrent validity determines if the instrument scores are correlated with other pertinent indicators or variables collected at the same time. For example, it would be important to know if a dyspnea survey instrument is correlated with number of cigarettes smoked per day. Predictive validity testing determines if the instrument scores can forecast performance on a pertinent indicator or variable in the future. For example,

NCLEX scores may be used to predict success of graduate nurses. Some authors use the term criterion-related validity, rather than convergent validity, to categorize concurrent and predictive validity (Waltz, et al, 2006).

Divergent validity. Divergent validity determines if scores from the instrument of interest are different from scores produced by another instrument known to measure a quite different concept. For example, a group of children would be administered two instruments at the same time, a new one intending to measure happiness and the other one known to measure sadness. If scores deviate from each other, divergent validity will have been demonstrated. Alternately, the contrasted group approach, using only the instrument of interest, can determine if scores from two very different groups are unrelated. For example, the instrument of interest (happiness) is administered to a known happy group of children and to a known sad group of children. Divergent validity would be demonstrated by if significant group differences in scores (t-test or ANOVA) were demonstrated (Waltz, et al, 2010).

Factor analysis. Factor analysis is yet another way to determine construct validity. Factors are groupings of items that match the multiple dimensions of the concept. Factors are identified through statistical patterns of correlations (specifically, shared variance) between the common items. (Waltz, et al, 2010). Conducting a factor analysis will determine not only if factors exist within the instrument, but also how many and which items belong to which factor. There are two types of factor analysis, confirmatory factor analysis (CFA) and experimental factor analysis (EFA) (Albright & Park, 2009).

If an instrument developer has the different factor categories in mind as the instrument is constructed, then a CFA procedure would likely be performed to confirm

those factors. A model is conceptualized and tested regarding how the data, collected for this type of psychometric testing, fits the model. In other words, the results would reveal to what degree the data support the model identified by the developer. A number of goodness of fit indices can be used to determine the degree of fit (e. g., goodness-of-fit index, adjusted goodness-of-fit index, normed fit index, standardized root mean squared residual) (Albright & Park, 2009; Waltz, et al, 2010).

Conversely, if there is not a set of a priori categories conceptualized for items, then an EFA would be conducted and the new factors would be revealed through correlation patterns. Each pattern or grouping of correlated items is termed a factor. The number of factors contained in an instrument is determined by eigenvalues (greater than 1 or other statistical criteria) and scree plots (visual observations of specific patterns). Factor loadings (the correlations) can be statistically rotated obliquely (allowed to correlate) or orthogonally (not allowed to correlate) using specialized statistical software for improved interpretation of the groupings. Instrument factors are typically termed subscales when the instrument is ready to use. For scoring purposes each subscale will have its own total subscale score (Waltz, et al, 2010).

Item analysis. Item analysis is an important approach in determining the validity of data generated by norm-referenced tests and instruments. CNSs and other advanced practice nurses often administer norm-referenced tests. For example, CNSs may administer norm-referenced tests during annual nurse competency testing.

An instrument developer commonly uses item-level analyses when individual items are examined separately to determine how well each item can discriminate *higher* versus *lower* test takers or if an individual item can predict test success. Used and interpreted

appropriately, an instrument developer may be able to reduce the number of test items to a critical few and still obtain valid test results.

A number of procedures are used in the initial and ongoing development of a test resulting in an item *p* level, discriminate index, a chi square (obtained through the use of an item-response chart), or a differential item function. An item *p* level indicates the proportion of correct answers, with *p* levels closer to 1.00 indicating easy items. Discrimination index values (D values) greater than +0.20 indicate an item's ability to discriminate and thereby predict performance on an entire test. An item-response chart is used to determine if a significant difference exists between the upper and lower 25% of the test takers. And finally, a differential item function identifies biased items that affect the probability of an item predicting success on a test taken by equally capable test takers (Waltz, et al, 2010).

Other Approaches to Determine Validity

Meta analysis has also been used to determine validity (Waltz, et al, 2010). Meta analysis is similar to a systematic review but adds statistical procedures to identify common research study results patterns across studies. In the case of determining validity, it can be used to determine if a number of studies that used the instrument of interest provided expected research results. For example, if a stress reduction intervention reduced stress to a similar degree as measured by the same new instrument in 10 studies, validity would be assumed.

Conducting descriptive studies to determine validity can also be helpful. Examples include observing individuals as they complete the instrument, interviewing those individuals to determine their interpretation of the items, and studying how judges apply

criteria when the purpose of the instrument is to classify either people or objects (Waltz, et al, 2010).

Summary

It is important for CNSs and other advanced practice nurses to consider the reliability of instruments and if those instruments generate valid data for clinical practice, evidence-based practice projects, and research studies. Psychometric testing uses specific research methods to evaluate the amount of error associated with any particular instrument. Reliability estimates explain more about how well the instrument is designed, while validity estimates explain more about scores that are produced by the instrument. An instrument may be architecturally sound overall (reliable), but the same instrument may not be valid. For example, if a specific group does not understand certain well-constructed items, then the instrument does not produce valid scores when used with that group. Many instrument developers may only conduct reliability testing once; yet continue validity testing in different populations over many years. All CNSs should be advocating for the use of reliable instruments that produce valid results. CNSs may find themselves in situations where reliability and validity estimates for some instruments are unknown. In such cases, CNSs should engage key stakeholders to sponsor nursing researchers to pursue this most important work.

References

Albright, J. J. & Park, H. M. (2009). Confirmatory Factor Analysis using Amos, LISREL, Mplus, SAS/STAT CALIS. University Information Technology Services Center for Statistical and Mathematical Computing Indiana University. *http://www.indiana.edu/~statmath*

Furr, R. M. & Bacharach, V. R. (2008). Pyschometrics: An Introduction. Thousand Oaks: Sage

McDowell, I. (2006). *Measuring Health: A Guide to Rating Scales and Questionnaires (3rd ed).* New York: Oxford University Press

Oermann, M. H. & Gaberson, K. B. (2014). Evaluation and Testing in Nursing Education, 4th edition. New York: Springer Publishing Company

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation*, 9(4). Retrieved February 7, 2015 from http://PAREonline.net/getvn.asp?v=9&n=4

Streiner, D. L. & Norman, G. R. (2008). Chapter 4. Scaling responses. In *Health Measurement Scales: A Practical Guide to Their Development and Use*, 4th ED. Oxford University Press. ISBN: 9780199231881

Waltz, C. F, et al (2010). Measurement in Nursing & Health, 4th Ed. NY: Springer Publishing Company