Reliability and Validity for Neuroscience Nurses

CE 2.0 ANCC Contact

Janice M. Buelow, Janice L. Hinkle, Molly McNett

ABSTRACT

The concepts of reliability and validity are important for neuroscience nurses to understand, particularly because they evaluate existing literature and integrate common scales or tools into their practice. Nurses must ensure instruments measuring specified concepts are both reliable and valid. This article will review types of reliability and validity—sometimes referred to collectively as a psychometric testing–of an instrument. Relevant examples in neuroscience are included to illustrate the importance of reliability and validity to neuroscience nurses.

Keywords: neuroscience, reliability, validity

Have you heard this story about reliability and validity? Joe walks into an ice cream shop every night at 5 P.M. and orders a milkshake. Every night at about 5:05 P.M., the soda clerk gives him a milkshake that tastes just like it did the night before. Joe loves this because he knows he can walk into this ice cream shop and always get a drink he likes that tastes the same way each time. One day, Joe invites a friend from work to join him. They both go into the shop and order what Joe thinks is the reliable milkshake. When it is delivered, however, his friend from work tastes the drink and declares, "This isn't a milkshake—it's an ice cream soda."

This story illustrates the concepts of reliability and validity. Although the clerk was reliably delivering the same drink night after night, he was not delivering a drink that actually fits the definition of a milkshake; therefore, the statement that the drink actually was a milkshake was not valid. For Joe in the ice cream shop, it may not make much difference that he was receiving an ice cream soda and not a milkshake, but for neuroscience nurses measuring physical concepts such as weight or temperature or behavioral concepts such as brain impairment or disability, the instrument measuring the concept clearly needs to be both reliable and valid. Consequently, two universal challenges of any measurement tool are reliability and validity.

Neuroscience nurses using tools in practice and researchers must ask themselves two important questions: What is the reliability of the measurement instrument? What is the validity of the measurement instrument? This article will review types of reliability and validity—sometimes referred to collectively as a psychometric testing of an instrument. Relevant examples are used to illustrate the importance of reliability and validity to neuroscience nurses.

Reliability

A measurement instrument that is reliable is one that is stable or consistent across time and across raters (Kerlinger, 1986). In statistical terms, "reliability" is the ability of an instrument to measure something consistently and repeatedly. However, understanding reliability in behavioral measures normally used by neuroscience nurses can be confusing. The reliability of a measure refers to its stability, internal consistency, and equivalence (Polit & Beck, 2016).

Stability

The stability of a scale is how well it measures the construct at different points in time. It is easiest to picture this when thinking about physical measures such as weight. When measuring weight, given that all other variables are the same (e.g., the amount of food consumed), if a scale weighs a person at 120 pounds today, that same scale should weigh that person at 120 pounds the next day.

"Test–retest reliability" is one method to determine stability of a scale over time and means that, when a test is given on two separate occasions, the results will be the same (Waltz, Strickland, & Lenz, 2010). When measuring both brain impairment behaviors

Question or comments about this article may be directed to Molly McNett, PhD RN CNRN, at mmcnett@metrohealth.org. She is the Director of Nursing Research, MetroHealth Medical Center, Cleveland, OH.

Janice M. Buelow, PhD RN, is Professor, Indiana University School of Nursing, Indianapolis, IN.

Janice L. Hinkle, PhD RN CNRN, is Fellow, Villanova University College of Nursing, Villanova, PA.

The authors declare no conflicts of interest.

DOI: 10.1097/JNN.0000000000000239

and disability, for example, if the scale used to measure each concept is administered to a group of people today, their answers should look similar 2 weeks from now if all other variables are the same. The statistical comparison measure used for test–retest reliability is the Pearson's *r* correlation coefficient; it can range from +1.00 to −1.00. A Pearson's *r* correlation coefficient of +1.00 indicates a perfect positive relationship, .00 indicates no relationship, and −1.00 indicates a perfect negative relationship (Munro, 2005). Cameron and colleagues (2008) used test–retest reliability to develop the Brain Impairment Behavior Scale (BIBS). Clinical team members tested the scale with 37 participants on two occasions 2 weeks apart. The correlation coefficients of .75, .88, .82, and .81, respectively, were reported for each of the four subscales, indicating strong positive relationships between the two administrations of the scale (Cameron et al., 2008). Similarly, 2-week test–retest was used in the psychometric testing of the Americanized version of the Guy's Neurological Disability Scale (GNDS). A Pearson's *r* correlation of .91 was reported; this indicates a strong relationship between the amount of disability measured at different times (Fraser & McGurl, 2007).

Internal Consistency

Internal consistency reliability is more complicated, because this type of reliability establishes how well each item (or question) on a scale measures the same construct. Internal consistency reliability is often measured with a statistical test called a Cronbach's alpha coefficient (Munro, 2005). This measures the extent to which items on an instrument fit together. Cronbach's alpha reliability coefficient normally ranges between 0 and 1.0. The closer the resulting number is to 1.0, the greater the internal consistency of the items on the scale. In behavioral measures, a 100% correlation would not be expected. As a rule of thumb, some professionals require a reliability of .70 (or 70%) or higher (obtained on a substantial sample) before they will use an instrument. Cameron and colleagues (2008) reported Cronbach's alpha coefficients ranging from .78 to .91 for the four domains of their 18-item BIBS. Because the values of the Cronbach's alpha coefficients all were greater than .70, each of the items included were considered to be measuring the same thing. For example, with a Cronbach's alpha of .89, each of the items in the apathy subscale appears to be measuring apathy. Another example of the measurement of internal consistency occurs in the work of Fraser and McGurl (2007), who reported Cronbach's alpha values for the entire GNDS for each administration of the scale. The Cronbach's alpha was .79 at time 1, .78 at time 2, and .80 at time 3, indicating good internal consistency (Fraser & McGurl, 2007).

Pearson's r correlation (range = + 1.00 to -1.00)
Cronbach's alpha (range = 0 to 1.0)
Cohen's kappa (range = 0 to 1).

Split-half reliability is another mechanism to evaluate internal consistency of a scale. This technique compares one half of a test with the other half based on the assumption that all items should be comparable in measuring one construct and the results should be similar. If there were 20 items on a measure, the first 10 items would be compared with the second 10 items. The Spearman–Brown correlation formula is used to determine split-half reliability.

Equivalence

The final component to consider when evaluating reliability is the degree of agreement among different raters using the same scale on the same patient or its equivalence. Interrater and intrarater reliability refers to how consistent ratings of a tool are between different raters or between the same rater at different time points (Waltz et al., 2010). If a tool is reliable, it should produce the same results, regardless of the rater. For example, if two nurses were scoring the same patient using the GNDS, it would be expected that their ratings on each item for the patient would be similar or identical. This is referred to as interrater reliability. Intrarater reliability refers to how well the same rater scores the patient at two different time points. Again, it would be expected that these scores would be the same if evaluated on the same patient within a very short period by the same person. Interrater and intrarater reliability is typically measured by a Cohen's kappa statistic, which evaluates degree of agreement among raters when controlling for chance. Kappa scores range from 0 to 1, with 1 indicating a perfect agreement. A kappa score of 0.6 is acceptable, with higher scores (0.75 and above) being indicative of high reliability among raters. Percentage of agreement among raters is often reported with the kappa statistic, with 100% agreement being the ultimate measure. Intraclass correlation coefficients may also be calculated when describing reliability among raters. Intraclass correlation coefficient describes the strength of the relationship between scores from different raters using the same scale.

Validity

Validity in behavioral measures refers to how well the instrument measures the construct it says it is

measuring (Kerlinger, 1986). For example, if an instrument is designed to measure disability, is it really measuring disability, or is it measuring impairment? Similar to reliability, there are different types of validity that can be established when evaluating a tool, which include content validity, criterion-related validity, construct validity, and factor analysis (Polit & Beck, 2016). The type of validity investigated depends on the purpose of the measure (Waltz et al., 2010).

Content validity is established by having a panel of experts familiar with the construct being measured judge the content of the instrument to establish how well they believe the items actually measure the content. Multiple judges usually are used, and their answers are compared to establish their level of agreement. Oftentimes, a content validity index is calculated to determine the relevance of each item of the tool in measuring the construct (Lynn, 1986). Scores from the panel of experts on each item are then evaluated based on the proportion of judges who agreed on the relevance of each item. Typically, content validity index scores ranging from 0.8 to 1.0 indicate high validity among an expert panel.

Criterion-related validity refers to how well an instrument compares with an established tool that measures the same construct. However, the validity of the established tool must already be recognized. The criterion-related validity of the instrument may also be established by examining if the tool accurately predicts an associated outcome of the construct or has predictive validity (Waltz et al., 2010). For example, the criterion validity of the Stroke Driver's Screening Assessment tool was established by comparing scores from the tool with results of on-the-road driving assessments (George & Crotty, 2010). Individuals who scored well on the Stroke Driver's Screening Assessment also performed well during the road driving assessments, supporting the validity of the instrument.

Construct validity refers to how well the instrument establishes the theoretical soundness of the instrument. This is established in multiple ways. When developing a behavioral instrument, authors usually hypothesize relationships between the new instrument and other established measures. For example, in disability, one might hypothesize that there would be a relationship between disability and activities of daily living. So a person who is more disabled would, in theory, have more difficulty managing his or her activities of daily living. This process of establishing construct validity is involved and generally requires multiple studies to accurately establish relationships among variables.

Factor analysis is a statistical process that is used to establish how individual items cluster around a given dimension. Subscales can be developed in this manner. Exploratory factor analysis often is used in the early stages of instrument development (Munro, 2005). A study by Cameron and colleagues (2008) reports a factor analysis of the BIBS. In the beginning, the instrument had 37 items, and four factors were identified when factor analysis was completed: apathy, comprehension/memory problems, depression/emotional distress, and irritability (Cameron et al., 2008). The technique of factor analysis was used in a slightly different manner in the testing of the GNDS. On the basis of previous research on the scale, a four-factor solution was tried, revealing a different configuration of items loading (clustering) on the factors. The authors concluded that the items on the Americanized version of the GNDS should not be conceptualized as falling together to form consistent subscales and recommended a 15-item version to be subject to further testing (Fraser & McGurl, 2007).

Conclusion

Neuroscience nurses should base interventions on evidence. To do so, it is important to become good consumers of research. When reviewing research articles, consider if the research findings are sound. If the measures that researchers use are not reliable and valid, their findings are not reliable or valid. Every time research is used, reliability and validity are some of the criteria on which neuroscience nurses should base their evaluation of research.

References

Cameron, J. I., Cheung, A. M., Streiner, D. L., Coyte, P. C., Singh, M. D., & Stewart, D. E. (2008). Factor structure and reliability of the brain impairment behavior scale. *Journal of Neuroscience Nursing*, *40*(1), 40–47.

Fraser, C., & McGurl, J. (2007). Psychometric testing of the Americanized version of the Guy's Neurological Disability Scale. *Journal of Neuroscience Nursing*, *39*(1), 13–19.

George, S., & Crotty, M. (2010). Establishing criterion validity of the Useful Field of View Assessment and Stroke Drivers' Screening Assessment: Comparison to the result of on-road assessment. *American Journal of Occupational Therapy*, *64*(1), 114–122.

Kerlinger, F. (1986). *Foundations of behavioral research* (3rd ed.). Orlando, FL: Harcourt Brace Jovanovich.

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

Munro, B. (2005). *Statistical methods for health care research* (5th ed.). Philadelphia, PA: Lippincott Williams & Wilkins.

Polit, D., & Beck, C. (2016). *Nursing research: Generating and assessing evidence for nursing practice* (10th ed.). Philadelphia, PA: Lippincott Williams & Wilkins.

Waltz, C. F., Strickland, L. O., & Lenz, E. R. (2010). *Measurement in nursing and health research* (4th ed.). New York, NY: Springer.

CE

Instructions:
- Read the article. The test for this CE activity can only be taken online at www.NursingCenter.com/CE/JNN. Tests can no longer be mailed or faxed. You will need to create (its free!) and login to your personal CE Planner account before taking online tests. Your planner will keep track of all your Lippincott Williams & Wilkins online CE activities for you.
- There is only one correct answer for each question. A passing score for this test is 13 correct answers. If you pass, you can print your certificate of earned contact hours and access the answer key. If you fail, you have the option of taking the test again at no additional cost.
- For questions, contact Lippincott Williams & Wilkins: 1-800-787-8985.

Registration Deadline: October 31, 2018

Disclosure Statement:
The authors and planners have disclosed that they have no financial relationships related to this article.

Provider Accreditation:
Lippincott Williams & Wilkins, publisher of *Journal of Neuroscience Nursing,* will award 2.0 contact hours for this continuing nursing education activity.

Lippincott Williams & Wilkins is accredited as a provider of continuing nursing education by the American Nurses Credentialing Center's Commission on Accreditation.

This activity is also provider approved by the California Board of Registered Nursing, Provider Number CEP 11749 for 2.0 contact hours. Lippincott Williams & Wilkins is also an approved provider of continuing nursing education by the District of Columbia, Georgia, and Florida, CE Broker #50-1223. Your certificate is valid in all states.

Payment:
- The registration fee for this test is $21.95.
- AANN members can take the test for free by logging into the secure "Members Only" area of http://www.aann.org to get the discount code. Use the code when payment is requested when taking the CE test at www.NursingCenter.com/CE/JNN.

For more than 85 additional continuing education articles related to Neurological topics, go to NursingCenter.com/CE.