Validity and reliability in quantitative studies

Roberta Heale,[1] Alison Twycross[2]

10.1136/eb-2015-102129

[1]School of Nursing, Laurentian University, Sudbury, Ontario, Canada
[2]Faculty of Health and Social Care, London South Bank University, London, UK

Correspondence to:
Dr Roberta Heale,
School of Nursing, Laurentian University, Ramsey Lake Road, Sudbury, Ontario, Canada P3E2C6;
rheale@laurentian.ca

Evidence-based practice includes, in part, implementation of the findings of well-conducted quality research studies. So being able to critique quantitative research is an important skill for nurses. Consideration must be given not only to the results of the study but also the *rigour* of the research. Rigour refers to the extent to which the researchers worked to enhance the quality of the studies. In quantitative research, this is achieved through measurement of the validity and reliability.[1]

Validity

Validity is defined as the extent to which a concept is accurately measured in a quantitative study. For example, a survey designed to explore depression but which actually measures anxiety would not be considered valid. The second measure of quality in a quantitative study is *reliability*, or the accuracy of an instrument. In other words, the extent to which a research instrument consistently has the same results if it is used in the same situation on repeated occasions. A simple example of validity and reliability is an alarm clock that rings at 7:00 each morning, but is set for 6:30. It is very reliable (it consistently rings the same time each day), but is not valid (it is not ringing at the desired time). It's important to consider validity and reliability of the data collection tools (instruments) when either conducting or critiquing research. There are three major types of validity. These are described in table 1.

The first category is *content validity*. This category looks at whether the instrument adequately covers all the content that it should with respect to the variable. In other words, does the instrument cover the entire domain related to the variable, or construct it was designed to measure? In an undergraduate nursing course with instruction about public health, an examination with content validity would cover all the content in the course with greater emphasis on the topics that had received greater coverage or more depth. A subset of content validity is *face validity*, where experts are asked their opinion about whether an instrument measures the concept intended.

Construct validity refers to whether you can draw inferences about test scores related to the concept being studied. For example, if a person has a high score on a survey that measures anxiety, does this person truly have a high degree of anxiety? In another example, a test of knowledge of medications that requires dosage calculations may instead be testing maths knowledge.

Table 1 Types of validity

Type of validity	Description
Content validity	The extent to which a research instrument accurately measures all aspects of a construct
Construct validity	The extent to which a research instrument (or tool) measures the intended construct
Criterion validity	The extent to which a research instrument is related to other instruments that measure the same variables

There are three types of evidence that can be used to demonstrate a research instrument has construct validity:

1. Homogeneity—meaning that the instrument measures one construct.
2. Convergence—this occurs when the instrument measures concepts similar to that of other instruments. Although if there are no similar instruments available this will not be possible to do.
3. Theory evidence—this is evident when behaviour is similar to theoretical propositions of the construct measured in the instrument. For example, when an instrument measures anxiety, one would expect to see that participants who score high on the instrument for anxiety also demonstrate symptoms of anxiety in their day-to-day lives.[2]

The final measure of validity is *criterion validity*. A criterion is any other instrument that measures the same variable. Correlations can be conducted to determine the extent to which the different instruments measure the same variable. Criterion validity is measured in three ways:

1. Convergent validity—shows that an instrument is highly correlated with instruments measuring similar variables.
2. Divergent validity—shows that an instrument is poorly correlated to instruments that measure different variables. In this case, for example, there should be a low correlation between an instrument that measures motivation and one that measures self-efficacy.
3. Predictive validity—means that the instrument should have high correlations with future criterions.[2] For example, a score of high self-efficacy related to performing a task should predict the likelihood a participant completing the task.

Reliability

Reliability relates to the *consistency* of a measure. A participant completing an instrument meant to measure motivation should have approximately the same responses each time the test is completed. Although it is not possible to give an exact calculation of reliability, an estimate of reliability can be achieved through different measures. The three attributes of reliability are outlined in table 2. How each attribute is tested for is described below.

Homogeneity (internal consistency) is assessed using item-to-total correlation, split-half reliability, Kuder-Richardson coefficient and Cronbach's α. In split-half reliability, the results of a test, or instrument, are

Table 2 Attributes of reliability

Attributes	Description
Homogeneity (or internal consistency)	The extent to which all the items on a scale measure one construct
Stability	The consistency of results using an instrument with repeated testing
Equivalence	Consistency among responses of multiple users of an instrument, or among alternate forms of an instrument

divided in half. Correlations are calculated comparing both halves. Strong correlations indicate high reliability, while weak correlations indicate the instrument may not be reliable. The Kuder-Richardson test is a more complicated version of the split-half test. In this process the average of all possible split half combinations is determined and a correlation between 0–1 is generated. This test is more accurate than the split-half test, but can only be completed on questions with two answers (eg, yes or no, 0 or 1).[3]

Cronbach's α is the most commonly used test to determine the internal consistency of an instrument. In this test, the average of all correlations in every combination of split-halves is determined. Instruments with questions that have more than two responses can be used in this test. The Cronbach's α result is a number between 0 and 1. An acceptable reliability score is one that is 0.7 and higher.[1,3]

Stability is tested using test–retest and parallel or alternate-form reliability testing. Test–retest reliability is assessed when an instrument is given to the same participants more than once under similar circumstances. A statistical comparison is made between participant's test scores for each of the times they have completed it. This provides an indication of the reliability of the instrument. Parallel-form reliability (or alternate-form reliability) is similar to test–retest reliability except that a different form of the original instrument is given to participants in subsequent tests. The domain, or concepts being tested are the same in both versions of the instrument but the wording of items is different.[2] For an instrument to demonstrate stability there should be a high correlation between the scores each time a participant completes the test. Generally speaking, a correlation coefficient of less than 0.3 signifies a weak correlation, 0.3–0.5 is moderate and greater than 0.5 is strong.[4]

Equivalence is assessed through inter-rater reliability. This test includes a process for qualitatively determining the level of agreement between two or more observers. A good example of the process used in assessing inter-rater reliability is the scores of judges for a skating competition. The level of consistency across all judges in the scores given to skating participants is the measure of inter-rater reliability. An example in research is when researchers are asked to give a score for the relevancy of each item on an instrument. Consistency in their scores relates to the level of inter-rater reliability of the instrument.

Determining how rigorously the issues of reliability and validity have been addressed in a study is an essential component in the critique of research as well as influencing the decision about whether to implement of the study findings into nursing practice. In quantitative studies, rigour is determined through an evaluation of the validity and reliability of the tools or instruments utilised in the study. A good quality research study will provide evidence of how all these factors have been addressed. This will help you to assess the validity and reliability of the research and help you decide whether or not you should apply the findings in your area of clinical practice.

Twitter Follow Roberta Heale at @robertaheale and Alison Twycross at @alitwy

Competing interests None declared.

References

1. **Lobiondo-Wood G**, Haber J. *Nursing research in Canada. Methods, critical appraisal, and utilization.* 3rd Canadian edn. Toronto: Elsevier, 2013.
2. **Korb K.** Conducting Educational Research. Validity of Instruments. 2012. http://korbedpsych.com/R09eValidity.html
3. **Shuttleworth M.** Internal Consistency Reliability. 2015. https://explorable.com/internal-consistency-reliability
4. **Laerd Statistics.** Determining the correlation coefficient. 2013. https://statistics.laerd.com/premium/pc/pearson-correlation-in-spss-8.php