The Concept of Validity

Dr Wan Nor Arifin
Unit of Biostatistics and Research Methodology, Universiti Sains Malaysia.
wnarifin@usm.my

Outlines

Measurement validity and reliability
The classical view of measurement validity
The validity

Measurement validity and reliability

- Measurement is “the process observing and recording the observations that are collected as part of a research effort.” (Trochim, 2006)
- ***Measurement validity*** is "the degree to which the data measure what they were intended to measure", or in other words, how close the data reflect the true state of what being measured (Fletcher, Fletcher and Wagner, 1996). It is synonymous to **accuracy**.
- ***Measurement reliability*** means **repeatability**, **reproducibility**, **consistency** or **precision** (Fletcher, Fletcher and Wagner, 1996; Gordis, 2009; Trochim, 2006). It is “the extent to which repeated measurements of a stable phenomenon – by different people and instruments, at different times and places – get similar result” (Fletcher, Fletcher and Wagner, 1996).
- Think of concept that we want to measure as target (Trochim, 2006) as shown in Figure 1 below, how accurate and precise you can get to the center of the target/concept.

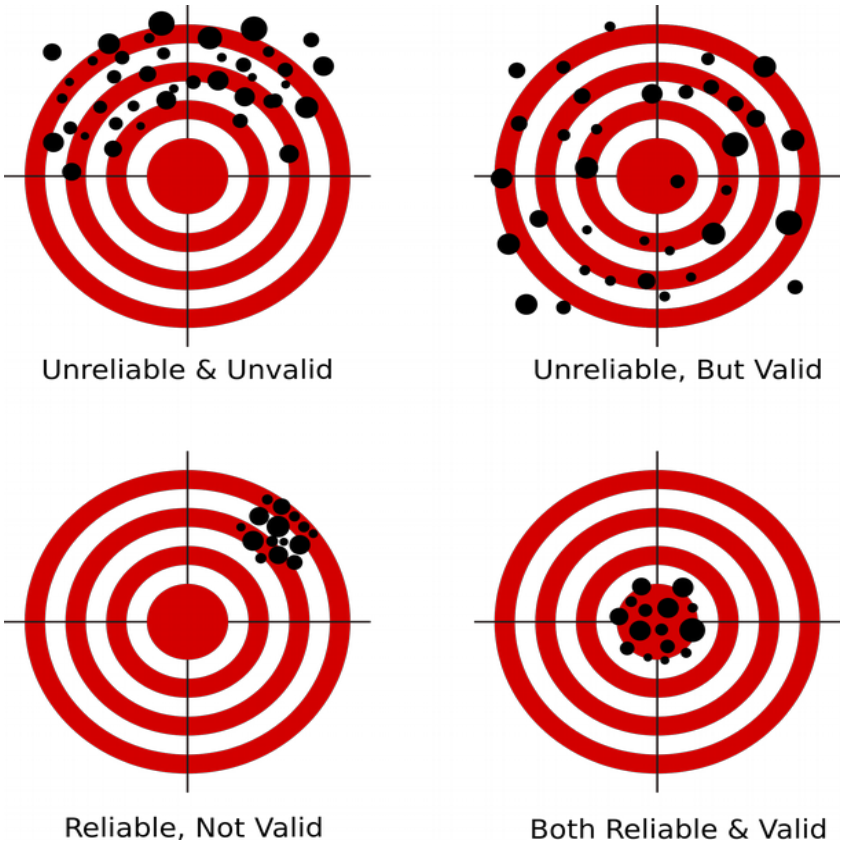

Figure 1: Validity and reliability. Image © Nevit Dilmen found at Wikimedia commons, licensed under the Creative Commons Attribution-Share Alike 3.0 Unported license.

The classical view of measurement validity

- Used to be divided into 3Cs (DeVellis, 1991; Fletcher, Fletcher and Wagner, 1996):
  1. Content validity.
  2. Criterion validity.
  3. Construct validity.
- Nowadays, validity is described differently under the unitary concept of validity (Cook, & Beckman, 2006; American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA & NCME], 1999).

The validity

- Validity is "the degree to which all the accumulated evidence supports the intended interpretation of test scores for the proposed purpose" (AERA, APA & NCME, 1999).
- The validity evidence can be obtained from five sources (AERA, APA & NCME, 1999; Cook, & Beckman, 2006):
  1. Content.
  2. Internal structure.
  3. Relations to other variables
  4. Response process.
  5. Consequences.

1. Content

- It describes how well a measure includes all the facets of an idea or concept, which a researcher intends to measure (Fletcher, Fletcher and Wagner, 1996).
- It "depends on the extent to which an empirical measurement reflects a specific factor of content" (Carmines and Zeller, 1979).
- It is "the extent to which a specific set of items reflect a content domain" (DeVellis, 1991).
- For example, if we want to measure anxiety, we should include symptoms like shaky hands, cold and clammy palms, stomach aches, palpitations and etc among the questions.
- We have covered briefly about approaches to development in "Questionnaire design" lecture. Now we are concerned with the draft of the questionnaire/measurement tool.
- Judged on three aspects (Streiner and Norman, 2008):
  1. Relevance: How relevant and related the items to the concept.
  2. Coverage: Adequate number of items to cover the concept.
  3. Representativeness: Number of items covering the item is proportionate to the importance of the concept.
- Judgment on these aspects is usually done by experts in related area (Streiner and Norman, 2008). We have covered the other further evaluation of a questionnaire in "Questionnaire design" lecture.
- For translated questionnaire, the goal is to achieve equivalence between original and translated version (**Will be discussed in detail later in translation)**.

2. Internal structure.

- It is concerned with the degree of the relationships among items and constructs as proposed or hypothesized (AERA, APA & NCME, 1999).
- **Construct** is "the concept or characteristic that a test is designed to measure" (AERA, APA & NCME, 1999).
- Recall: **Construct = Factor = Domain = Concept = Idea**
- Generally proven on the basis of analyses that can prove the correlatedness (i.e. correlations coefficients, factor loadings) and dimensionality (number of factors), of importance are (Cook, Thomas & Beckman, 2006)
  - Factor analysis (exploratory and confirmatory).

    - Reliability.
- The analyses are based on variables available ***internal*** to the test itself (i.e. the questions, items), hence the name internal evidence.
- **Will be discussed in detail on day 2 of the workshop.**

3. Relations to other variables

- It is concerned with the relationship of the measurement tool scores to other external variables, which may include other measurement tools/questionnaires, and other observable variables or criteria.

1. Convergent and discriminant evidence

- Correlation with other measures of similar concept (Streiner and Norman, 2008; Matthews, Zeidner and Roberts, 2007):
- **Good correlation** between a construct from the new measure and a related construct measuring the same concept from other measure is an evidence of *convergent validity*.
- For example, correlation between depression scale score from DASS and BDI score is supposed to be good (both are inventories to measure depression).
- **Poor correlation** between a construct from the new measure and an unrelated construct from other measure measuring different concept is an evidence of *discriminant validity*.
- For example, correlation between depression scale score from BDI and intelligence quotient (IQ) score is supposed to be poor (as both are intended to measure totally different concepts).
- The correlation is usually given by Pearson's correlation coefficients.

2. Test-criterion relationship

- This evidence of relationship indicates how well it correlate with directly observable variables (Fletcher, Fletcher and Wagner, 1996; Streiner and Norman, 2008).
- The criterion are of two types (Streiner and Norman, 2008):
    1. Concurrent,
        - A new tool is correlated/compared with a criterion (clinical judgment, gold standard, group).
        - Assessment done at **same time** (concurrent).
        - For example, 8am blood glucose level (new measurement tool) is used to distinguish between diabetic and non-diabetic patient based on established way of diagnosis of diabetes mellitus (criterion). Similarly, recall HIV rapid test vs the criterion ELISA test to establish HIV status.
        - In another example, the total scores of a tool should be able to differentiate between a number of groups that are supposed to be different based on the characteristics that the tool is supposed to measure (BDI that measures depression should be able to differentiate between depressed patients and healthy persons).
    2. Predictive,
        - A new tool is correlated/compared with a criterion, which is measured in the future.
        - Assessment done at **different time interval**: new tool (current) and criterion

(future).
        - For example, total score of a questionnaire on attitude towards statistics on admission to statistics course is used to predict whether students would pass of fail the course at first attempt.
        - In another example, a new scoring of cancer survival on diagnosis is compared against the outcome of the patient 5 year later.
        - Also consider bachelor CGPA on admission to master program vs CGPA for the master program.
- Analyses:
    - Depending on how you want to provide the evidence.
        - Different mean total scores between groups by?
        - Establish cut-offs in relation to the criterion by?

4. Response process

- It is concerned with the process of responding to the questions.
- May be done in cognitive debriefing (previous lecture) by probing the respondent as to how he comes up with a response per question.
- For interviewer rated, may observe how the interviewer/rater comes up with a rating.

5. Consequences

- It is concerned with the evidence regarding the intended and unintended consequences of the result from a measurement tool.
- For example, if a person is rated as depressed, what would be the consequence of that? Referral to psychiatric clinic (intended)? Losing job (unintended)? Etc.
- As an additional source of evidence to support the rest of evidence.

References

Brown, T. A. (2006). Confirmatory factor analysis for applied research. New York: The Guilford Press.

Carmines, E. G. & Zeller, R. A. (1979). Reliability and validity assessment (Sage university paper series on quantitative applications in the social sciences, series no. 17). Newsbury Park, CA: Sage Publications.

Cook, D. A., & Beckman, T. J. (2006). Current concepts in validity and reliability for psychometric instruments: theory and application. *The American journal of medicine*, *119*, 166.e7-166.e16.

DeVellis, R. F. (1991). Scale development: Theory and applications. California: Sage Publications.

Fletcher, R. H., Fletcher, S. W., & Wagner, E. H. (1996). Clinical epidemiology: the essentials (3rd ed.). Maryland: Williams & Wilkins.

Floyd, F. J. & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment,* **7(3),** 286.

Gordis, L. (2009). Epidemiology (4th ed.). Philadelphia: Saunders.

Kline, R. B. (2011). Principles and practice of structural equation modeling. 3rd ed. New York:

Guilford Publications.
Matthews, G., Zeidner, M. & Roberts, R. D. (2007). Emotional intelligence: Consensus, controversies, and questions. In*:* Matthews, G., Zeidner, M. and Roberts, R. D. (eds.), *The science of emotional intelligence: Knowns and unknowns*. New York: Oxford University Press.
Streiner, D. L. & Norman, G. R. (2008). Health measurement scales: a practical guide to their development and use. New York: Oxford University Press.
Trochim, W. M. K. (2006). Research methods knowledge base. [Online] Available at: http://www.socialresearchmethods.net