Validity of Measurement Tools

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

Last update: 14 May, 2023

Outlines

Introduction

- 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 the 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.

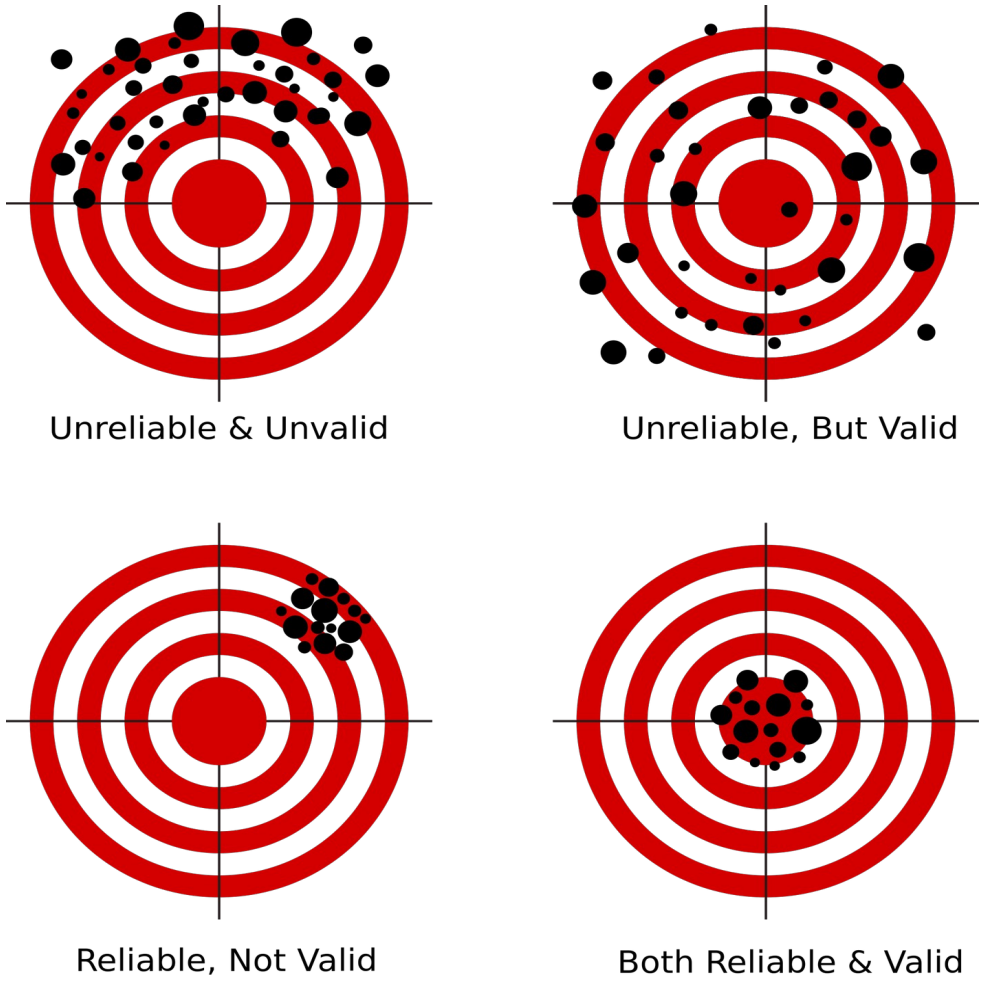

Figure 1: Validity and reliability

Classical view of measurement validity

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

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 5 sources (AERA, APA & NCME, 1999; Cook, & Beckman, 2006):
  1. Content
  2. Internal structure
  3. Relations to other variables
  4. Response process
  5. Consequences

Class activity 1 (30 min):

- Read and understand “Current concepts in validity and reliability for psychometric instruments” (Cook, Thomas & Beckman, 2006) [20 minutes] (**excluding Reliability**, pg. 166.e12)
- Briefly explain what you have learned [10 minutes].

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.

Translation

- For translated questionnaire, the goal is to achieve equivalence between original and translated version. **Five** key aspects of equivalence are (Streiner and Norman, 2008):

Aspects	**Description**	**Western**	**Malaysian**	**Adaptation**
Conceptual	Do respondents from two different populations and cultures understand the concept similarly?	Canning is child abuse.	Canning is way to teach children to behave properly.	Change to suitable items representing abuse in local culture.
Item	Whether the items are relevant and acceptable in target population.	Turning on heater. Use of furnace. Manual transmission for car. An apple a day, keeps doctor away.	Items not relevant in local setting.	Drop the items. Find suitable items conceptually.
Semantic	Concerns similarity in meaning attached to an item.	I get butterflies in my stomach.	Saya ada rama-rama dalam perut?	Saya rasa gelisah/cemas.
Operational	Equivalence of operational aspect of the measure, i.e. format of the measure, the instructions and mode of administration.	Direct question? Self-administered?	Indirect, politely phrased question? Interviewer guided?	Change the operational aspect of the questionnaire.
Measurement Equivalent	Concerns equivalence of psychometric properties of the measure, i.e. validity and reliability.	Factor analysis Reliability.	–	–

Class activity 2 (30 min):

- Three groups.
- Read and understand:
  - Group 1: “ABC of Content Validation and Content Validity Index Calculation” (Yusoff, 2019).
  - Group 2: “Guidelines for the Process of Cross-Cultural Adaptation of Self-Report Measures” (Beaton et al., 2000).
- Briefly explain what you have learned [10 minutes].

Self-study:

- “Is the CVI an Acceptable Indicator of Content Validity? Appraisal and Recommendations” (Polit, Beck & Owen, 2007).
- “Principles of Good Practice for the Translation and Cultural Adaptation Process for Patient-Reported Outcomes (PRO) Measures” (Wild et al., 2005).

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.

1. Factor analysis

- Factor analysis (exploratory and confirmatory):
  - In the context of validity evidence, a high (i.e. > 0.4 or 0.5) and statistically significant factor loading of an item under its corresponding factor indicates its association to the factor (Floyd and Widaman, 1995). This reflects the correlatedness between the items and defines their relationship with the respective factors.
  - As an evidence of dimensionality, the number of factors should reflect the proposed number. It is also possible to determine correlation between factors. A correlation between factors ≥ 0.85 indicates factors overlap (Brown, 2006), may indicate poor dimensionality in our context of validity.
  - Lastly, by looking at model fitness to the data, we are able to verify the validity of a theoretical model based on the data at hand.

2. Reliability

- Discussed below.

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.

Reliability

Class activity 3 (30 min):

- Read and understand “Current concepts in validity and reliability for psychometric instruments” (Cook, Thomas & Beckman, 2006) [20 minutes] (**Reliability part** from pg. 166.e12)
- Briefly explain what you have learned [10 minutes].

Overview

- Measurement reliability.
- In the current framework, part of validity evidence from internal structure source.
- Repeatability, reproducibility, consistency or precision (Flether, Flether and Wagner, 1996; Trochim, 2006).
- “the extent to which repeated measurements of a stable phenomenon – by different people and instruments, at different times and places – get similar result” (Flether, Flether and Wagner, 1996).
- For example:

    Measurement tool → Weighing scale – Brand XYZ, Brand ABC
    Phenomenon → Weight of a person A – 60kg.

    Data on weight of person A measured 5 times:

        Brand XYZ: *58*, *57*, *61*, *62*, *63*
        Brand ABC: *60*, *59*, *60*, *60*, *61*

    Which brand gives more reliable reading?

True Score Theory of Measurement Error

- When we measure something, we may not get the true reading of a phenomenon (e.g. weight), it is susceptible to error.
- More so with psychological state of mind or emotion e.g. perception, depression. Measured with questionnaires, inventories → giving scores to emotion.
- As such any observed reading/score is thought to be made of true reading and error.

  Observed reading = True reading + Error

  $X = T + e_x$

  to put it in another way:

  Variance of observed reading = Variance of true reading + Variance of error

  $VAR(X) = VAR(T) + VAR(E_x)$

- Measurement error in this theory considered as random error.

Theory of Reliability

- Going back to our true score theory, reliability is defined as:

$$\text{Reliability}\left(\rho_{XX}\right)=\frac{\text{T}}{\text{X}}=1-\frac{e_X}{\text{X}}$$

or in term of variability:

$$\rho_{XX}=\frac{\text{VAR(T)}}{\text{VAR(X)}}=1-\frac{\text{VAR}(\text{E}_x)}{\text{VAR(X)}}$$

thus, the implication of the formula is that:

  - When VAR(T) = VAR(X), in ideal case of zero error, $\rho_{XX}=1$ → Perfectly reliable.
  - When VAR(T) = 0, in which VAR(Ex) = 1, $\rho_{XX}=0$ → Nothing measured except error!

- Since we are unable to know exactly the true reading and also the error, we are also unable to know how reliable a measurement tool is using the formulas mentioned above.

- Reliability is however can estimated in term of correlation.
- Recall the definition of of reliability as “repeatability”. If we measure weight and able to get the same result again and again, we can conclude that the weighing scale is reliable.
- When the variables are correlated to each other, the correlation shows the amount of “truth” shared between the variables.
- Remember that correlation is given by,

$$r_{XX} = \frac{\mathrm{COV}(X_1, X_2)}{\sqrt{\mathrm{VAR}(X_1)\mathrm{VAR}(X_2)}}$$

  with the lower part of equation becomes,

$$\sqrt{\mathrm{VAR}(X_1)}\sqrt{\mathrm{VAR}(X_2)} = \mathrm{VAR}(X) \quad \text{when} \quad \mathrm{VAR}(X_1) = \mathrm{VAR}(X_2)$$

  with the upper part, $\mathrm{COV}(X_1, X_2)$ comparable to that of VAR(T).
- Following that basic way of establishing the reliability, specific ways of estimating reliability were developed, depending on types of reliability.

Types of Reliability

- Generally divided into four types (Trochim, 2006; Kline, 2011):
    1. Test-retest reliability
    2. Parallel-forms reliability
    3. Interrater reliability
    4. Internal consistency reliability

Test-Retest

- It is the reliability of a tool when used on the same group on two different occasions after a time interval (Trochim, 2006).
- If the tool scores between these two occasions are correlated, it is assumed that random error due to temporal/time factor is minimal (Kline, 2011).
- The interval between the two measurements is usually between 2 to 14 days (Streiner and Norman, 2008).
- This type of reliability indicates the stability of the tool over time, given the measured concept is also stable (e.g. personality).
- For continuous numerical outcome Pearson's correlation and intraclass correlation coefficient is suitable to assess the reliability. For categorical outcome, Cohen's kappa can be used.

Parallel-Forms

- When a questionnaire has a large number of items designed to represent a construct.
- These items are considered equal to each other in term of its representativeness of the construct.
- The items are randomly allocated into two parallel sets that are considered equivalent to each other, so as it does not matter whether set 1 or 2 is used that it would give similar representation of the domain.
- Correlation between these two sets is the parallel-forms reliability (Trochim, 2006).
- Pearson's correlation between the total score of the two sets is the reliability.

Interrater

- Interrater reliability examines the effect of different raters/observers on scores/outcomes (Streiner and Norman, 2008)
- It is relevant when human factor is an important part of an assessment, and thus contributes to observed score variability (Trochim, 2006, Kline, 2011).
- For example, in measurement of blood pressure manually using mercury sphygmomanometer, in which the consistency/agreement of the readings of staffs taking the measurement is questioned, interrater reliability is to be determined. As we are concerned with continous numerical outcome, **intraclass correlation coefficien**t is suitable.
- In another example, we are interested to know the agreement between two radiologists on the presence or absence of cancerous lesion on X-ray films. As here we are concerned with categorical outcome, **kappa coefficient** is suitable.

Internal Consistency

- It is the degree to which responses are consistent across the items within a construct i.e. measure the same thing (Kline, 2011) in similar direction for a particular subject. In other words, how homogenous the items in a construct in term of their variance.
- Low internal consistency means that the items are heterogeneous within a construct i.e. do not measure the same concept, thus the total score is not the best way to summarize the construct (Kline, 2011).
- When scores for items within a construct are almost similar in values and in similar direction, they are positively correlated to each other, thus would indicate that they measure common thing.
- In comparison to the rest of reliability types, it only requires measurement on a single occasion.
- There are several ways of estimating internal consistency of a set of questionnaire, as discussed below.

Internal Consistency

Average Item-Item Correlation

- The average of all bivariate correlations between the items is taken as reliability coefficient.

Average Item-Total Correlation

- Sum up the score of the items → total score.
- Calculate the bivariate correlations between each item to the total score.
- The average of these correlations is taken as reliability coefficient.

Split-Half

- Not to be confused with parallel-forms reliability.
- In parallel-forms reliability, two forms separated earlier on before administered to respondents.
- Considered as parallel or equivalent halves of each other.
- Split-half → randomly split the questionnaire before analysis, i.e. after being administered to respondents.
- Correlation between total score of the split-half questionnaires is taken as reliability coefficient.

Internal Consistency

- It is the degree to which responses are **consistent** across the items within a construct i.e. measure the same thing (Kline, 2011) in **similar direction** for a particular subject. In other words, how **homogenous** the items in a **construct** in term of their variance.
- **Low internal consistency** means that the items are **heterogeneous** within a construct i.e. do not measure the same factor, thus the total score is not the best way to summarize the construct (Kline, 2011).
- When responses for items within a construct are **positively correlated** to each other, they may measure the same factor. In this case, **high internal consistency** is obtained.
- In comparison to the rest of reliability types, it only requires measurement on **a single occasion**.

Cronbach's Alpha

- **Cronbach's alpha coefficient** is a common way to indicate internal consistency of a construct. It is given as:

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k} \sigma_i^2}{\sigma_T^2}\right)$$

$k =$ number of items
$\sigma_i^2 =$ variance for i th item score
$\sigma_T^2 =$ variance for total score

- Ranges 0-1.
  → When α=1, the items are all identical and perfectly correlated to each other, i.e measure the same thing.
  → When α=0, the items are all independent and none related to each other, i.e do not measure the same thing.
- Satisfactory 0.7-0.8. Clinical use >0.9 (Bland & Altman, 1997).

References and recommended reading

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

Articles for class activities

Beaton, D. E., Bombardier, C., Guillemin, F., & Ferraz, M. B. (2000). Guidelines for the process of cross-cultural adaptation of self-report measures. *Spine*, *25*(24), 3186-3191.

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.

Polit, D. F., Beck, C. T., & Owen, S. V. (2007). Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. *Research in nursing & health*, *30*(4), 459-467.

Yusoff, M. S. B. (2019). ABC of content validation and content validity index calculation. Education in Medicine Journal, 11(2), 49-54.

Wild, D., Grove, A., Martin, M., Eremenco, S., McElroy, S., Verjee-Lorenz, A., & Erikson, P. (2005). Principles of good practice for the translation and cultural adaptation process for patient-reported outcomes (PRO) measures: report of the ISPOR Task Force for Translation and Cultural Adaptation. *Value in health*, *8*(2), 94-104.