Reliability

Nils Myszkowski, PhD

Outline

What is reliability?

Estimating reliability

Predicting reliability

Aims of this session

- At the end of this session, you should be able to:
  - Know the different forms of reliability
  - Estimate the reliability of a test using different research designs and statistical techniques
  - Understand that tests characteristics play a role in what type and what degree of reliability can reasonably be expected

Reliability

- Ensuring and examining reliability is a very important step in the process of building robust psychological tests.
- In this course, we will review ways to ensure and improve reliability.
- We will also review ways to examine the reliability of tests, based on gathered data.

Outline

What is reliability?

Estimating reliability

Predicting reliability

What is reliability?

Conceptually

Statistically

Psychometrical qualities

- The main metrological qualities of a measure are:
  - Discrimination power
    - Ability to reveal differences between different statistical units (persons)
    - $\approx$ Variability between units
    - *Ex : If you weight 100 different individuals, the scale indicates different weights*

Psychometrical qualities

- The main metrological qualities of a measure are:
  - Discrimination power
    - Ability to reveal differences between different statistical units (persons)
    - $\approx$ Variability between units
    - *Ex : If you weight 100 different individuals, the scale indicates different weights*
  - **Reliability**
    - Ability to **not** reveal differences within one statistical unit (person)
    - $\approx$ Consistency within units
    - *Ex : If individuals weight themselves twice, the scale indicates the same weight*

Psychometrical qualities

- The main metrological qualities of a measure are:
  - Discrimination power
    - Ability to reveal differences between different statistical units (persons)
    - ≈ Variability between units
    - *Ex : If you weight 100 different individuals, the scale indicates different weights*
  - **Reliability**
    - Ability to **not** reveal differences within one statistical unit (person)
    - ≈ Consistency within units
    - *Ex : If individuals weight themselves twice, the scale indicates the same weight*
  - Validity
    - Ability to measure what what was intended
    - ≈ Adequacy with "the real world"
    - *Ex: The scale attributes individuals their real weight.*

Reliability

- Reliability is the ability to **not** reveal differences within one individual.
- In other words, the score produced by a measure should ideally be as stable as the construct.
- The construct is of course most of the time considered theoretically rather stable.

True score theory

- Between two occasions of measurement of the construct in the same unit (individual), we consider that what is stable is the "true score" of the individual. The part that is unstable and that made the two measures differ is considered an random error of measurement.

$$Observed\ score = \overbrace{True\ score}^{\text{constant}} + \underbrace{Error}_{\text{random}}$$

$$x = \tau + \epsilon$$

True score theory

- Between two occasions of measurement of the construct in the same unit (individual), we consider that what is stable is the "true score" of the individual. The part that is unstable and that made the two measures differ is considered an random error of measurement.

$$Observed\ score = \overbrace{True\ score}^{\text{constant}} + \underbrace{Error}_{\text{random}}$$

$$x = \tau + \epsilon$$

True score theory

- Therefore, the more stable the observed score, the "truer" it is, and the less full of random error it is.
- This theory is also known as "Classical Test Theory" (CTT), as opposed to Modern Test Theory or Item-Response Theory (IRT).

True score theory

- To better understand True Score Theory, we have to think of a measurement of a construct as **one** measurement of it **among other** possible measurements.
- If we averaged an infinite number of measurements, the random errors would cancel themselves and we would have a perfect estimate of the true score...

True score theory

- To better understand True Score Theory, we have to think of a measurement of a construct as **one** measurement of it **among other** possible measurements.
- If we averaged an infinite number of measurements, the random errors would cancel themselves and we would have a perfect estimate of the true score...
- ...but, in practice, it doesn't happen : We have a limited number of measurements within an instrument, we use the instrument generally only once, with one judge/examiner, etc.
- Checking reliability is about making sure that we don't have too much "noise" in the signal due to these limits.

Factors of stability

- Conceptually, there is one element that contributes to the stability of the measure.
- What contributes to the stability of the measure is the (assumed) stability of the construct.
- So, a good question to ask ourselves is the extent to which we believe that to be true...
  - Is the construct itself very stable?
  - If the construct is not so stable, this will be reflected in its measure.

Factors of instability

- What contributes to the instability of the results (and what standardization aims at reducing) are features of the individual or the situation that can affect test scores but have nothing to do with the attribute being measured:
  - Temporary general individual characteristics: health, fatigue, motivation, emotional strain
  - Temporary test specific individual characteristics: comprehension of the specific test task, specific tricks or techniques of dealing with the particular test materials, fluctuations of memory, attention or accuracy
  - Testing situation characteristics: freedom from distractions, clarity of instructions, interaction with characteristics of examiner
  - Luck in selection of answers by guessing, momentary distractions
  - Previous exposure to the items

Random error and standardization

- Error is simply caused by all these multiple sources of non-trait (or non-construct, or non-true score) variability (different time, different scorers, etc.).
- Therefore, better reliability is notably achieved through reducing this non-construct variability through **standardization**.
- Implementing standards in administration and scoring reduces measurement error. Hence the importance of standardization.

What is reliability?

Conceptually

Statistically

True Score Theory

- We can "translate" the postulate $x = \tau + \epsilon$ in variances:

$$\sigma_x^2 = \sigma_\tau^2 + \sigma_e^2$$

Reliability

- From there, a general definition of reliability $\rho_{xx'}$ is given as the ratio of "true" variability in the "observed" variability.

$$\begin{aligned}\rho_{xx'} &= \frac{\sigma_\tau^2}{\sigma_x^2} \\ &= 1 - \frac{\sigma_e^2}{\sigma_x^2}\end{aligned}$$

Reliability

- From there, a general definition of reliability $\rho_{xx'}$ is given as the ratio of "true" variability in the "observed" variability.

$$
\begin{aligned}
\rho_{xx'} &= \frac{\sigma_\tau^2}{\sigma_x^2} \\
&= 1 - \frac{\sigma_e^2}{\sigma_x^2}
\end{aligned}
$$

- Unfortunately, as you can guess, in practice, there is no way to observe or calculate the true score, so a variety of methods are used to estimate reliability (we'll use $r_{xx'}$ to denote them).

Standard Error of Measurement

- From reliability (or, practically, from reliability estimates), we can estimate the expected variability of the observed scores of an individual for an infinite number of measures, known as the Standard Error of Measurement (SE_M).
- A simple way to conceptualize SE_M is to think of it as the typical amount of error in a score.

Standard Error of Measurement

- The Standard Error of Measurement SE_M is calculated through reliability $\rho_{xx'}$ and population standard deviation σ_x as:

$$SE_M = \sigma_x \sqrt{1 - \rho_{xx'}}$$

- In practice, we replace population standard deviation by its sample estimate s and reliability $\rho_{xx'}$ by its sample estimate $r_{xx'}$.

An important note

- It is here important to note an important limit of CTT: Reliability (and thus SE_M) do not depend on the true score.
- In other terms, a measure is not expected to have more or less error as a function of an individual's true score. The error is randomly distributed but not conditional upon τ.
- It's a strong (and sometimes wrong) conceptualization, that is one of the reasons for psychometricians turning to Item-Response Theory (IRT), for which items (and measures) provide more or less information and are more or less reliable for specific latent levels.

What is reliability?

Statistically

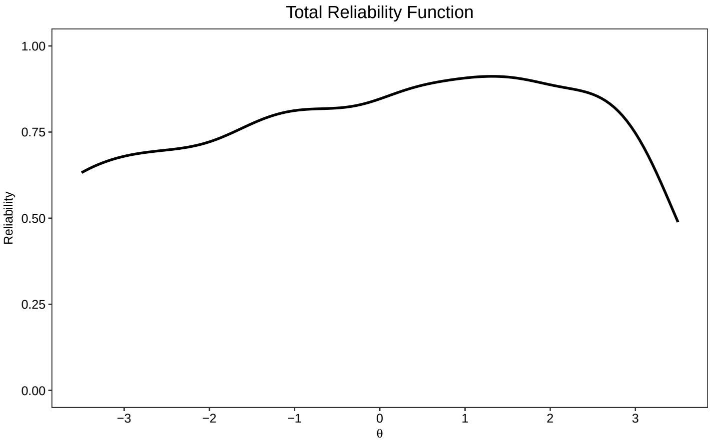


Figure: In contrast, in Modern Test Theory, reliability is conditional upon θ_i...

What is reliability?

Statistically

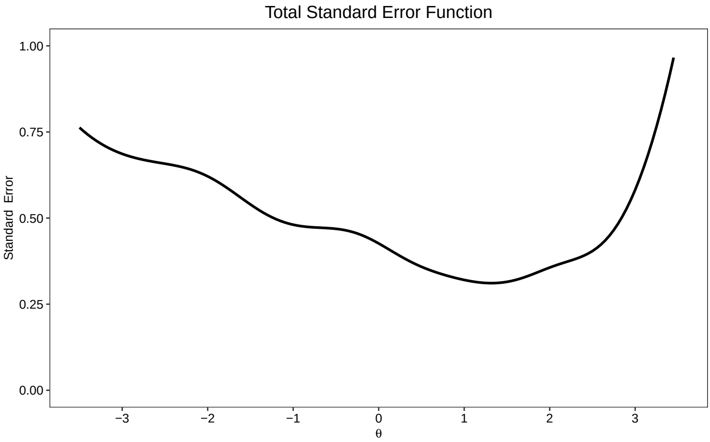


Figure: and so is the Standard Error of Measurement

Outline

What is reliability?

Estimating reliability

Predicting reliability

Stability

- Back to the true score formula:

$$Observed\ score = \overbrace{True\ score}^{\text{constant}} + \underbrace{Error}_{\text{random}}$$

- Unfortunately, there is no way to directly observe or calculate the true score, so a variety of methods are used to estimate the reliability of a test.
- Better reliability is achieved:
  - When the measurement error is minimal.
  - Better reliability is achieved when the observed score is close to the true score, which is constant.
  - As a consequence, better reliability is achieved when the observed score is stable.

Estimating reliability

Types of reliability

Statistics

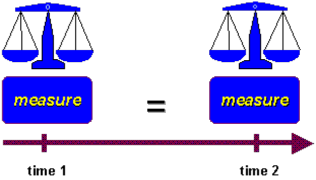
Internal consistency

Research design considerations

- Therefore, to check that the observed score is stable, our research design has to allow us to have at least two occasions of measurement.
- For this reason, psychometric evaluation research designs generally include evaluating reliability through verifying stability notably across **different sources of non-trait or non-true score variability**:
  - Time
  - Judges/scorers/raters
  - Test-forms

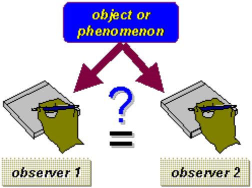
Test-retest reliability

- Typical psychometric research designs involve examining reliability across **time**.
  - Test-retest procedure
  - Used to assess the consistency of a measure from one time to another.



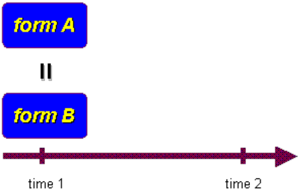
Inter-rater reliability

- Typical psychometric research designs involve examining reliability across **raters** (when it's relevant and especially when the administration and the scoring could be subjective at some point).
  - Multiple judges/scorers/raters procedure
  - Used to assess the degree to which different raters/observers give consistent estimates of the same phenomenon.



Parallel-Forms reliability

- Typical psychometric research designs involve examining reliability across **parallel forms** (when there are different parallel forms of a test).
  - Parallel forms procedure
  - Used to assess the consistency of the results of two tests constructed in the same way from the same content domain.



Statistics

- To assess reliability, the extent to which two series of observed scores in the same sample are consistent, we use correlational analyses.
- A positive strong correlations are expected (>.70) between the two series of scores.
  - Test-retest correlation
  - Inter-rater correlation
  - Parallel-form correlation

Estimating reliability

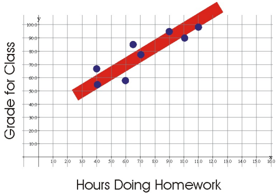
Types of reliability

Statistics

Internal consistency

Correlations

- Correlations indicate the strength and the direction of a linear relationship between two numerical variables (in our case, two scores)
- 2 complimentary techniques:
  - Correlation coefficients: r = $\frac{cov(X,Y)}{\sqrt{var(X)}\sqrt{var(Y)}}$ (formula is not to be known for exams)
  - Correlation diagrams:



Strength of correlations

- To quantify strength, we look at the **absolute value** of r, or at how close to the regression line the points are.
- **Null** or negligible correlation $\Longleftrightarrow 0 \leq |r| < .10 \Longleftrightarrow$ points are not aligned at all
- **Weak** correlation $\Longleftrightarrow .10 \leq |r| < .30$
- **Moderate** correlation $\Longleftrightarrow .30 \leq |r| < .50 \Longleftrightarrow$ points are not aligned at all
- **Strong** correlation $\Longleftrightarrow .50 \leq |r| < 1 \Longleftrightarrow$
- **Perfect** correlation $|r| <= 1 \Longleftrightarrow$ points are perfectly aligned.

Direction

- The direction is indicated by the **sign** of r, or by the slope of the regression line.
- **Negative** correlation $\Longleftrightarrow$ $r < 0$ $\Longleftrightarrow$ Slope is negative (descending line)
- **Positive** correlation $\Longleftrightarrow$ $r > 0$ $\Longleftrightarrow$ Slope is positive (ascending line)

Test–Retest Reliability

Test–retest correlation coefficients and average measure Intraclass Correlation Coefficients (ICC [2,k]) were computed for each scale and subscale. The RSMS (r = .85; ICC = .53), the CAS (r = .82; ICC = .49), AMSP (r = .86; ICC = .61), SEBO (r = .80; ICC = .59), ATSCI (r = .84; ICC = .53), and CSV (r = .81; ICC = .66) showed acceptable temporal stability (Landis & Koch, 1977).

Figure: An example reporting of test-retest reliability (from Myszkowski et al., 2014)

Estimating reliability

- Types of reliability
- Statistics
- Internal consistency

Items and scales

- Items are **different** measures of **one** construct (or one facet).
- Therefore, the responses of one individual to these items should have a certain degree of coherence/consistency.
- In other terms, because each item is itself a measure, we should also check reliability *inside* the test.
- This degree of coherence is called **internal consistency**. Because it is related to the stability of the score itself, it is sometimes referred to as **scale score reliability**

Internal consistency

- Internal consistency is the extent to which the responses to the items of a scale are consistent with each other.
- The degree of internal consistency is most of the time quantified by a statistical index called Cronbach's α .

$$\alpha = \frac{K}{K-1}\left(1 - \frac{\sum_{i=1}^{p} \sigma_{y_i}^2}{\sigma_x^2}\right)$$

  - K is the number of items
  - σ_{y_i} is the variance of item i
  - σ_x is the variance of the total score
- This formula is *not* to be known for the exams or assignments.

Alpha values

- The better the internal consistency, the higher the value of the α (even though α is considered a lower bound to internal consistency).
- A general rule of thumb (the most widely used in classical test theory) is that α values should be **higher than .70** (at least).

Estimating reliability

Internal consistency

Scale Score Reliability

Both scales showed good scale score reliability, as the Cronbach's alpha for the RSMS was .82 (95% CI [.80, .84]) and the Cronbach's alpha for the CAS was .81 (95% CI [.79, .83]). Concerning the subscales of the RSMS and the CAS, Cronbach's alpha were .77 (95% CI [.74, .80]) for the AMSP, .82 for the SEBO (95% CI [.80, .84]), .79 for the ATSCI (95% CI [.77, .81]) and .80 for the CSV (95% CI [.78, .82]).

Figure: An example reporting of internal consistency (from Myszkowski et al., 2014)

Estimating reliability
Internal consistency

Table 3. Reliability of the original and French versions of the TIPI

Version	Dimension	Cronbach's α	Inter-item correlation	$R_{test-retest}$ 3 weeks	$R_{test-retest}$ 6 weeks
Original version	Extroversion	.68			.77
	Agreeableness	.40			.71
	Conscientiousness	.50			.76
	Emotional stability	.73			.70
	Openness	.45			.62
French version	Extroversion	.69	.52	.78	.82
	Agreeableness	.22	.13	.62	.68
	Conscientiousness	.57	.40	.58	.72
	Emotional stability	.61	.44	.70	.76
	Openness	.39	.23	.69	.68

Figure: An example reporting of both internal consistency and test retest reliability (from Storme, Tavani & Myszkowski, 2016)

Alpha values

- However, if α is too high it may suggest that some items are redundant as they are testing the same question but in a different guise. A maximum α value of 0.90 has been recommended.
- An important thing to know is that Cronbach's α values are typically higher when the number of items is high.
- Therefore, a low value of α could be due to a low number of questions, poor inter-relatedness between items and/or heterogeneous constructs ("weak unidimensionality").

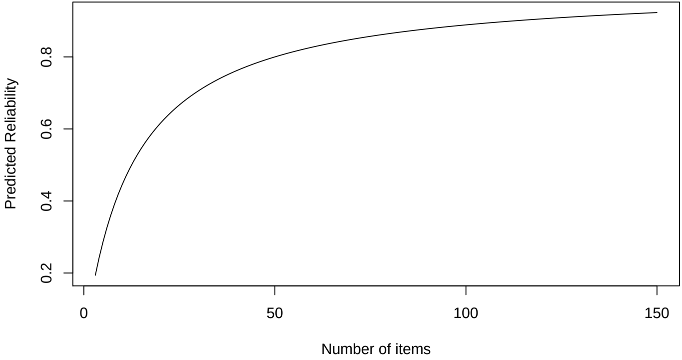


Figure: Predicted reliabilities (based on an initial reliability of $.80$ for 50 items)

Estimating reliability

Internal consistency

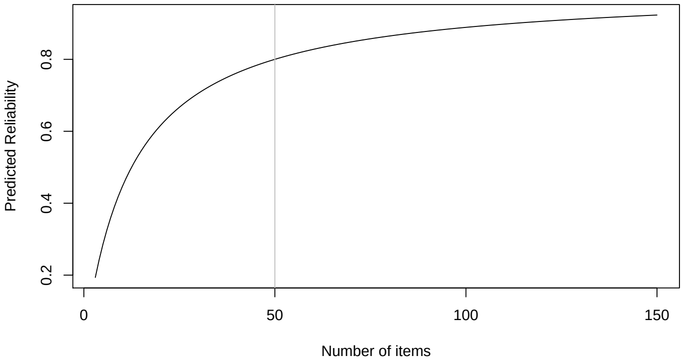


Figure: Predicted reliabilities (based on an initial reliability of $.80$ for 50 items)

Estimating reliability
Internal consistency

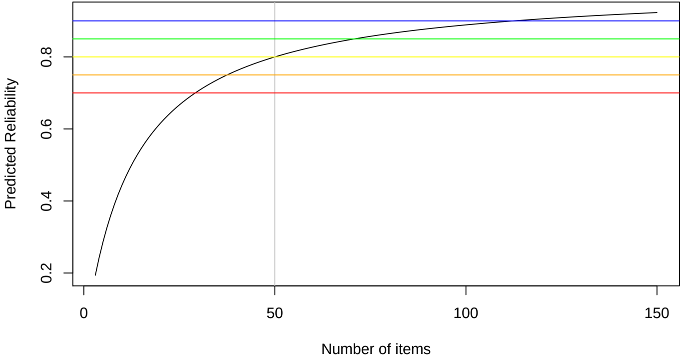


Figure: Predicted reliabilities (based on an initial reliability of $.80$ for 50 items)

Item reduction processes

- It is frequent to eliminate items on the basis of the degree to which they lower the α.
- To do so, we examine the "α **if deleted**" of each item, which most statistical software provide.
- We eliminate the items that lower the α (they are the ones with an alpha if deleted that is higher than the alpha of the entire scale). By deleting items one by one, we perfomr a stepwise backwards item selection.
- Note: Other strategies for item deletion exist, and are not necessarily based on α.

Exercise

- In this case, should an item be deleted? Which one and why?

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.805	.796	9

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Qu1	24.20	45.029	.633	.588	.767
Qu2	23.93	47.352	.520	.651	.783
Qu3	24.07	46.638	.654	.899	.767
Qu4	23.40	47.114	.551	.823	.779
Qu5	23.60	51.257	.389	.573	.799
Qu6	24.47	50.695	.372	.693	.802
Qu7	24.07	45.210	.615	.777	.770
Qu8	24.20	56.457	.128	.791	.823
Qu9	24.07	45.210	.589	.610	.774

Exercise

- In this case, should an item be deleted? Which one and why?

Reliability Statistics

Cronbach's Alpha	N of Items
.867	5

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
data1	29.50	41.538	.603	.862
data2	29.80	37.908	.691	.842
data3	30.00	41.026	.812	.812
data4	30.00	42.051	.757	.825
data5	30.20	43.446	.631	.853

A final note on α

- Cronbach's α is by very far the most reported estimate of reliability since its inception in the 1951.
  - It's easy to compute, and available in all statistical packages
- However, it is very criticized by psychometricians for not being an accurate estimate of reliability, nor an estimate of "internal consistency", because it makes unrealistic assumptions (especially τ-equivalence).
- More recommended (Zinbarg et al., 2005) but less known is McDonald's ω, which estimates the general factor saturation of a test.
- Another possibility is to compute reliability from information in Item-Response Theory Analysis (Raju et al., 2007).

Outline

What is reliability?

Estimating reliability

Predicting reliability

Spearman-Brown Reliability Prediction

- Reliability can be predicted for different test-lengths using the **Spearman–Brown prediction formula**.

$$Predicted\ \rho_{xx'} = \frac{n\rho_{xx'}}{1 + (n-1)\rho_{xx'}}$$

- n is the number of replications of the original test length (e.g. $n = 2$ means twice the original test length, $n = .5$ means half of it)
- $\rho_{xx'}$ is the test reliability

Example

- If we have a reliability of $.6$ for 10 items and are thinking of adding 5 items ($n = 1.5$).

$$
\begin{aligned}
Predicted\ \rho_{xx'} &= \frac{n\rho_{xx'}}{1 + (n-1)\rho_{xx'}} \\
&= \frac{.9}{1.3} \approx .69
\end{aligned}
$$

- So we predict a reliability of $.69$.

Spearman-Brown Reliability Prediction

- The same formula can be used to solve for item length for a desired $Predicted\ \rho_{xx'}$.

$$n = \frac{Predicted\ \rho_{xx'}(1 - \rho_{xx'})}{\rho_{xx'}(1 - Predicted\ \rho_{xx'})}$$

Spearman-Brown Reliability Prediction

- Note that, even though you may encounter it in the literature, this formula is also heavily criticized in the literature (e.g. Borsboom & Mellenbergh, 2002).
- Interestingly, it is heavily used in psychometric meta-analysis, especially in Personnel Psychology, which is probably due to the influence of Hunter and Schimdt on the developement of meta-analysis.

Example

- If we have a reliability of $.9$ for 50 items and want to know how many we should keep to have a reliability of $.80$.

$$n = \frac{Predicted\, \rho_{xx'}(1 - \rho_{xx'})}{\rho_{xx'}(1 - Predicted\, \rho_{xx'})}$$
$$= \frac{.8(1 - .9)}{.9(1 - .8)} = \frac{.08}{.18} \approx .44$$

- So we can keep $.44$ times the original test, so we should keep $.44 \times 50 = 22$ items.