Reliability and Validity of a Measure

Advanced Master in Agricultural Economics and Policy

Giovanbattista Califano

University of Naples Federico II, giovanbattista.califano@unina.it

Attitudinal Scales for the Study
of Consumer Preferences
April 24, 2026

Roadmap

20/04	Hello, Psychometrics!	✓
23/04	The Questionnaire	✓
24/04	Reliability and Validity of a Measure	◯
27/04	Latent Variables: Reflective or Formative?	◯
30/04	A Bit of SEM	◯
07/05	Stata Stata Stata Stata Stata Stata Stata	◯

Recommended readings:

- Chapters 4 and 7 from Jhangiani et al.2019)
- Chapters 7, 10 and 14 from Olivero and Russo2022)
- Chapter 12 from Mehmetoglu and Jakobsen2022)

Jhangiani, R. S., Chiang, I. A., Cuttler, C., and Leighton, D. C. (2019). *Research Methods in Psychology*. Kwantlen Polytechnic University, 4th edition.

Mehmetoglu, M. and Jakobsen, T. G. (2022). *Applied Statistics Using Stata: A Guide for the Social Sciences*. SAGE Publications Ltd, 2nd edition.

Olivero, N. and Russo, V. (2022). *Psicologia dei Consumi*. McGraw-Hill Education, 3rd edition.

Before we begin...

Previously on...

What do we mean by measurement?

The *systematic* assignment of a score to entities (individuals, objects, events...), such that the score represents a characteristic of those entities.

Previously on...

What do we mean by measurement?

The *systematic* assignment of a score to entities (individuals, objects, events...), such that the score represents a characteristic of those entities.

- But how do we know that score **truly** represents the characteristic we want to study?
- Researchers do not take this for granted...

Example

Califano's intelligence test

Califano's test was developed to measure intelligence quotient. According to the test, a person's IQ is given by the number of letters in their name (n) multiplied by 20:

$$IQ = n \cdot 20$$

Example

Califano's intelligence test

Califano's test was developed to measure intelligence quotient. According to the test, a person's IQ is given by the number of letters in their name (n) multiplied by 20:

$$IQ = n \cdot 20$$

- This test is more **reliable** than most existing tests for measuring IQ :)

Example

Califano's intelligence test

Califano's test was developed to measure intelligence quotient. According to the test, a person's IQ is given by the number of letters in their name (n) multiplied by 20:

$$IQ = n \cdot 20$$

- This test is more **reliable** than most existing tests for measuring IQ :)
- It might be quite **invalid** :(

Moving on to...

Reliability

Validity

Stata

Reliability of a measure

Reliability

refers to the degree of **consistency** of a measure. Psychometrics focuses, in particular, on three types of reliability:

1. Over time (*test-retest*)
2. Across researchers (*inter-rater*)
3. Across items (*internal*)

Reliability over time

Test-Retest

If the measured construct is hypothesised to be consistent over time, the scores should be as well. Reliability over time consists in studying the correlation between two measurement distributions obtained by administering the same test twice to the same group of subjects after a given time interval.

Example

If a person is very intelligent today, they should be so a week from now too...

Reliability across researchers

Inter-Rater

If different researchers use the same system to assign scores, these scores should be positively correlated with one another.

Reliability across items

Internal

Internal reliability is generally the most relevant when using self-report instruments, and refers to the consistency of respondents' answers to a multi-item measure.

Reliability across items

Food Technology Neophobia Scale (Cox & Evans, 2008)

The FTNS is a popular instrument for measuring neophobia towards food technologies and is based on the Likert scale.

FTNS 1: I don't need to eat food made using new technology because the food I already eat is fine

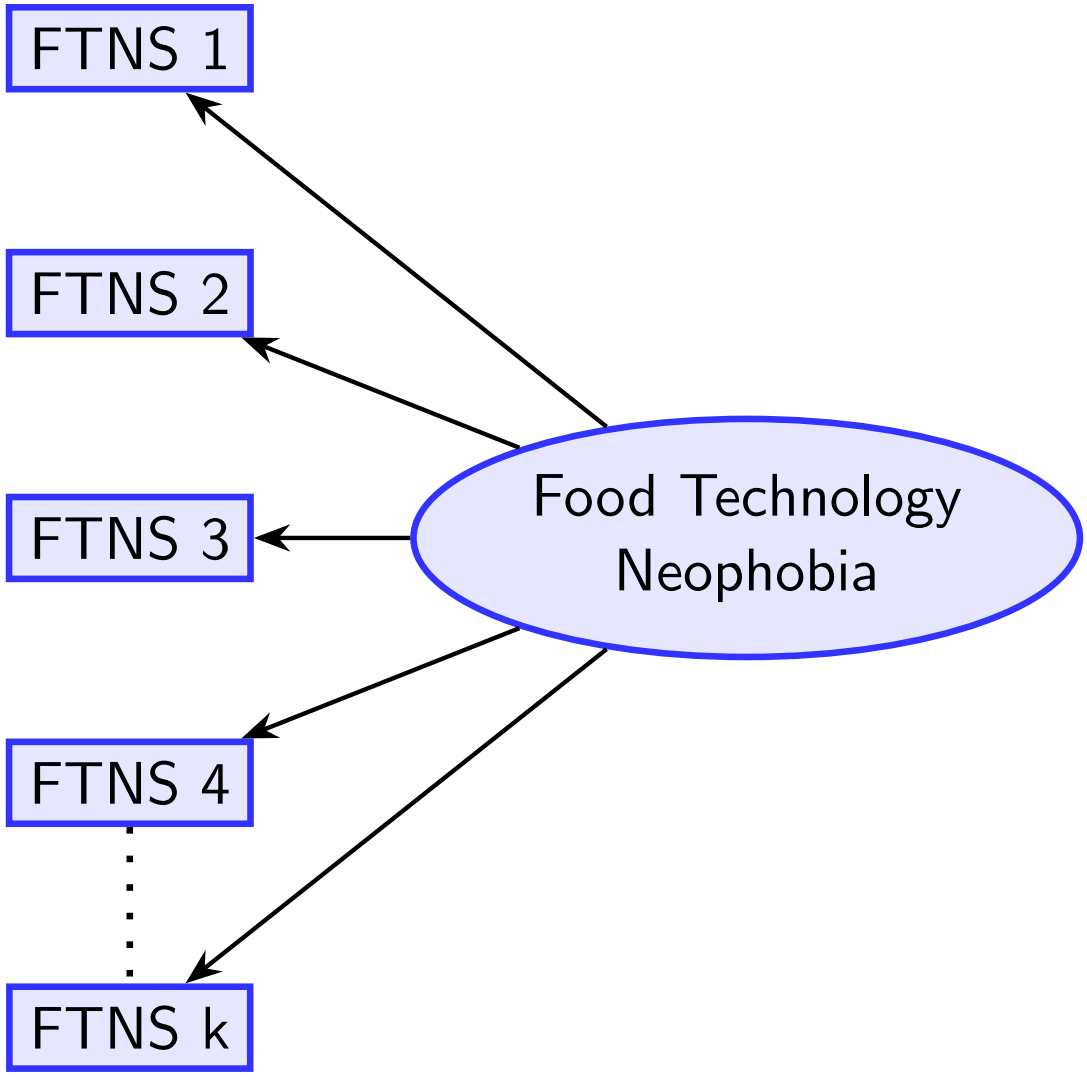
FTNS 2: New food technologies are something I am uncertain about

FTNS 3: New food products are not healthier than traditional foods

(...)

Strongly disagree	*Disagree*	*Neither agree nor disagree*	*Agree*	*Strongly agree*
○	○	○	○	○

Reliability across items



Reliability across items

The internal consistency of a scale can be measured in various ways, again based on the correlation index.

The most intuitive approach is *split-half* correlation:

1. The scale items are divided into two blocks (e.g. odd and even);
2. Scores are aggregated for the two blocks, by sum or mean;
3. The correlation index is calculated between the aggregated scores of the two blocks.

A split-half correlation coefficient $r > .80$ is generally considered an indicator of good internal consistency.

Reliability across items

A more sophisticated and widely used approach is *Cronbach's alpha*, defined as:

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k} \sigma_{Y_i}^2}{\sigma_X^2}\right)$$

where:

- k is the number of items composing the scale,
- $\sigma_{Y_i}^2$ is the variance of item i,
- σ_X^2 is the variance of the total score (given by the sum of all items).

In most cases, $\alpha > .70$ *is considered satisfactory.*

Moving on to...

Reliability

Validity

Stata

Validity of a measure

We have seen how **Califano's test** for IQ might enjoy a good degree of *reliability* (e.g. test-retest). What can we say about its *validity*?

Validity of a measure

Validity tells us how well the scores of a measure represent the characteristic they are supposed to represent.

Validity is also divided into several types:

- Face validity
- Content validity
- Criterion validity
- Discriminant validity

Face and content validity

These are the two types of validity that are not established by quantitative methods. **Face validity** refers to common sense: it takes little to say that the number of letters in a name has nothing to do with intelligence.

Content validity, on the other hand, refers to how well the measure *covers* the construct of interest (e.g. the risk perception and perceived uselessness of new food technologies components in the FTNS).

Criterion validity

A **criterion** can be any variable we think should be correlated with the construct we are trying to measure:

- Concurrent validity: when the criterion is measured at the same time as the construct;
- Predictive validity: when the criterion is measured after the construct;
- Convergent validity: when the criterion is an alternative measure of the same construct.

Discriminant validity

This is the degree to which the scores of a measure are *not* correlated with measures of conceptually distinct variables.

Criterion and discriminant validity

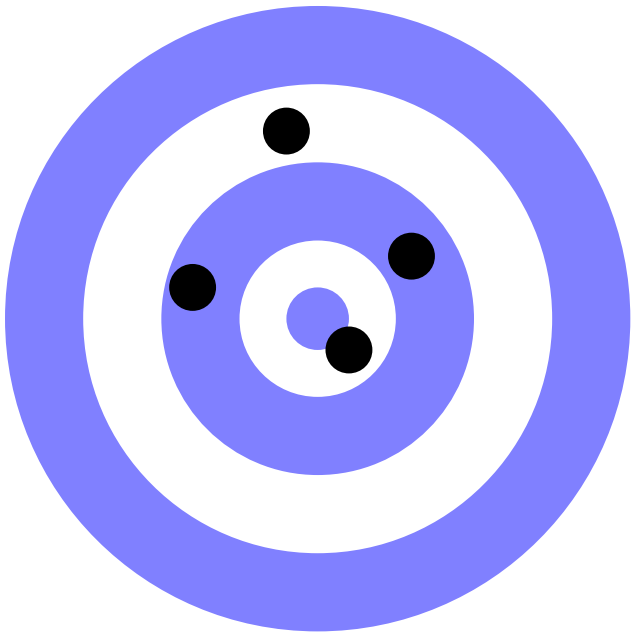
An example

Suppose we want to test the criterion and discriminant validity of a scale measuring self-efficacy in tackling the master's programme.

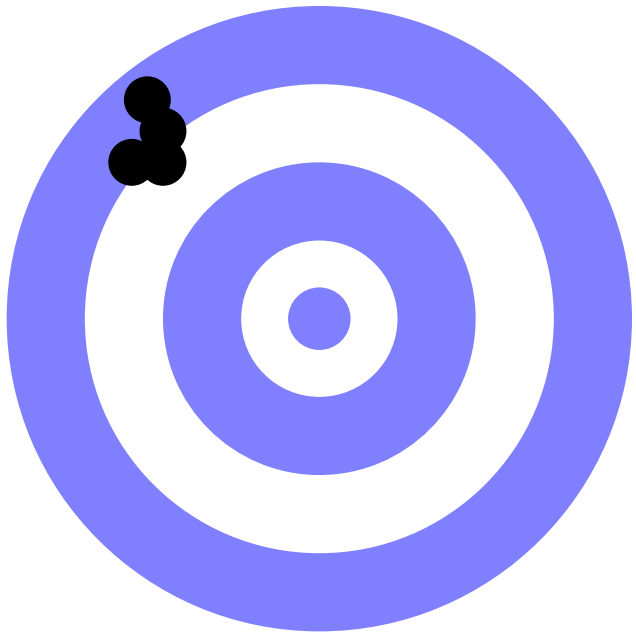
- **Concurrent validity**: self-efficacy in tackling the master's should be positively correlated with attitudes towards Stata;
- **Predictive validity**: self-efficacy in tackling the master's should be able to predict the average grade obtained during the master's;
- **Convergent validity**: if we had another scale measuring the same self-efficacy, the two scales should show a positive correlation with each other;
- **Discriminant validity**: self-efficacy refers to confidence in one's ability to tackle specific tasks or situations, while self-esteem is a general evaluation of oneself. The two measures should therefore not be too highly correlated.

Summary

Reliability $\equiv$ Precision
Validity $\equiv$ Accuracy



Accurate, Not precise



Not accurate, Precise

Summary

- Measuring a psychological construct requires four fundamental steps:
  1. **Conceptual definition**: clarifying what is meant by the construct;
  2. **Operational definition**: translating the construct into something observable and measurable;
  3. **Implementation of the measure**: choosing or building the appropriate instrument;
  4. **Evaluation of the measure**: testing the instrument to verify its validity and reliability.

Summary

1. **Conceptual definition**
   - A clear definition of the construct is needed.
   - Consulting the scientific literature is essential.
2. **Operational definition**
   - Specifies *how* the construct will be measured.
   - Multiple operational definitions of the same construct are possible.

Summary

3. Implementing the measure

- Using an existing instrument: faster, already validated, comparable with other studies.
- Creating a new measure: only if *necessary*; pay attention to simplicity, clarity and number of items.

4. Evaluating the measure

- Preliminary testing with a few participants.
- Check: clarity of instructions, timing, comprehensibility.
- Collect feedback to correct any issues before data collection.

Moving on to...

Reliability

Validity

Stata

clear all ;)

```
.  webuse set https://califano.xyz/data
.  webuse masteristi
```