Study guide

PSY721 – Tests and Measurements

Nils Myszkowski

12/5/22

Table of contents

> ⚠ Warning
>
> This covers aspects of the class considered important and not self-evident, but it should *not* be considered as the only document to review.

> **Note**
>
> This document is updated regularly. Check back soon!

1 Introduction

1.1 Psychometric terminology

- The psychological attribute being measured is referred to as the **construct**.
- To increase accuracy, dimensions are measured through responses on multiple indicators, called **items**.
- When persons are administered a test, it yields **item responses**.
- If the measurement tool is intended to measure multiple **facets** of the construct, it may be called an **inventory**, and the measure of one of the facets is often called a **subscale**.
- The properties of psychological tests are investigated through the discipline of **psychometrics**, and these properties are generally referred to as **psychometric** (or **metrological**) **qualities**.

1.2 Fundamental statistics for this class

1.2.1 Levels of measurement

- **Nominal/categorical**
  - Special case: **Dichotomous/binary** (2 categories)
  - e.g., multiple choice
- **Ordinal** (response categories that can be ordered)
  - e.g., Likert scales
- **Interval** (the interval between responses is useful information)
  - e.g., an extroversion scale's sum scores
- **Ratio** (the value 0 is meaningful, in that it implies the absence of the quantity of interest)
  - e.g., response time in seconds

1.2.2 Types of statistics

- Univariate vs. bivariate vs. multivariate
  - **Univariate statistics** study one variable at a time (e.g., the mean score of item 1).
  - **Bivariate statistics** study the relation between 2 variables (e.g., the correlation between item 1 and item 2).
  - **Multivariate statistics** study the overall relations between more than 2 variables (e.g., a factor analysis of a 10-item instrument).
- Descriptive vs. inferential
  - **Descriptive statistics** are used to describe phenomena observed in the **sample**.
  - **Inferential statistics** are used to draw conclusions in the **population**.

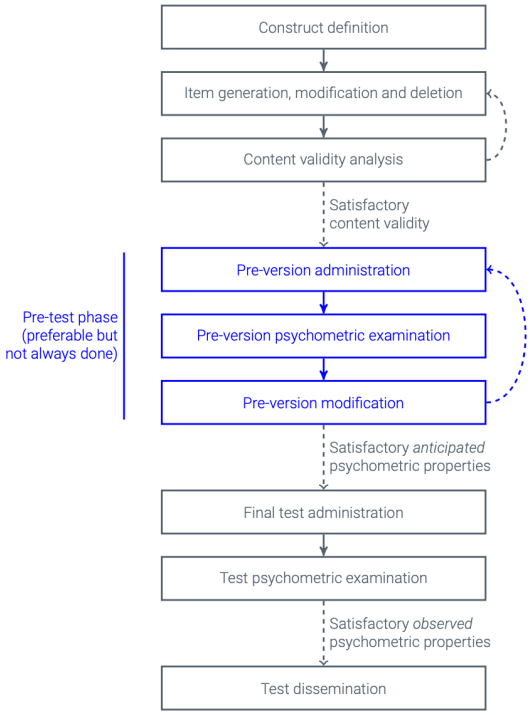
1.2.3 Note on correlations

Correlations are largely used in psychometrics, especially to study reliability and validity.

A correlation coefficient r is a statistic that represents the strength and direction of the relation between 2 numeric variables (e.g., 2 item scores).

Correlations range between -1 (perfect negative) and $+1$ (perfect positive):

- Strength
  - Negligible/Null : $|r| < .10$
  - Weak/Small : $.10 < |r| < .30$
  - Moderate : $.30 < |r| < .50$
  - Strong/Large : $.50 < |r|$
  - The stronger the correlation, the more aligned the points.
- Direction
  - Positive : $r > 0$
    - When one variable increases, the other tends to increases
  - Negative : $r < 0$
    - When one variable increases, the other tends to decrease



2 Test construction process

3 Item construction

3.1 Response formats

Typical response formats:

- Likert scales (ordinal response)
- Multiple Choice Questionnaire (transformed to binary usually)
- True/false (binary response)
- Open-ended questions (usually uses raters, which turn measures into binary or ordinal scales)
- Forced-choice (categorical response)
- Time to solve a problem (continuous response)
- Visual analog scales (continuous response)

To use average or sum scoring, all the items should have the same response format.

3.2 How to build items

Items are in general built so that item responses are manifestations of the construct (i.e., are caused by the construct, see causal theory of measurement).

For example, it is because a person is extroverted (construct) that they will answer "yes" (item response) to "Do you like going out with friends?" (item).

Two main approaches (which can be mixed) to item construction:

- **Deductive approach**: Items are designed from theory/previous research.
- **Inductive approach**: Items are designed from asking a sample of subject matter experts various questions about how the construct manifests (e.g., to measure clinical depression, we could ask a panel of clinical psychologists how they detect depressive symptoms in their clients).

In general, at this stage we want to maximize content validity, primarily (see section on validity and content validity).

3.3 Response biases

Common biases in psychological testing:

- **Social desirability bias**: Tendency of respondents to respond in order to present themselves more or less favorably.
  - Common problem of self-report measures
  - Comprised of **self-deception** (biased opinion of oneself) and **other-deception** ("faking")
  - Context specific (e.g., higher in recruitment contexts) and construct specific (e.g., some attributes are more desirable than others).
  - Solutions notably include using neutral items, using forced-choice response formats, or lie scales)
- **Comprehension biases** (complexity, double-barreled items, ambiguity, etc.)
- Response format biases
  - **Central tendency vs. extreme response bias** (tendency to prefer central or extreme responses)
  - **Acquiescence bias** (solutions include notably using negatively worded items)
- **Emotional biases** (negative or positive emotions can induce bias)
- **Applicability bias** (items need to be equally applicable and appropriate to any individual of the aimed population)

- **Cultural biases** (may lead to discriminatory practices)
- **Stereotype bias** (when an individual is put in a testing situation that may involve negative stereotypes about their performance in similar tests, they tend to underperform).
- **Non-response bias** (when the individuals that do not respond a specific item or a test have different characteristics that individuals who respond).

3.4 Scoring biases

Biases may be induced by the use of raters:

- **Order biases** (biased induced by the order of rating responses, e.g., contrast bias)
- **Consistency bias** (raters tend to seek a consistent view of respondents)
- **Expectancy bias** (raters tend to use expectations about a respondent's performance in their judgement)

4 Test theory

4.1 Causal theory of measurement

The Causal Theory of Measurement (CTM) is a core underlying assumption of Classical Test Theory (as well as Modern/Item-Response Theory).

According to this theory, **observed item responses are caused by unobserved (i.e., latent) attributes** (called the true score in Classical Test Theory, called a latent variable in Modern/Item-Response Theory).

4.2 Classical test theory (CTT)

4.2.1 General assumptions

CTT assumes that (observed) item scores X are comprised of a true score T (an error-free score, or expected score) and a random error E:

$$X = T + E$$

The error E is assumed to a random draw from a Gaussian (Normal) distribution, of mean 0 and error variance σ^2 (the error variance is fixed for a given item, but for certain models it is also fixed across items). An implication of this assumption is that items are distributed according to Gaussian (Normal) distributions.

$$E \sim \text{Normal}(0, \sigma^2)$$

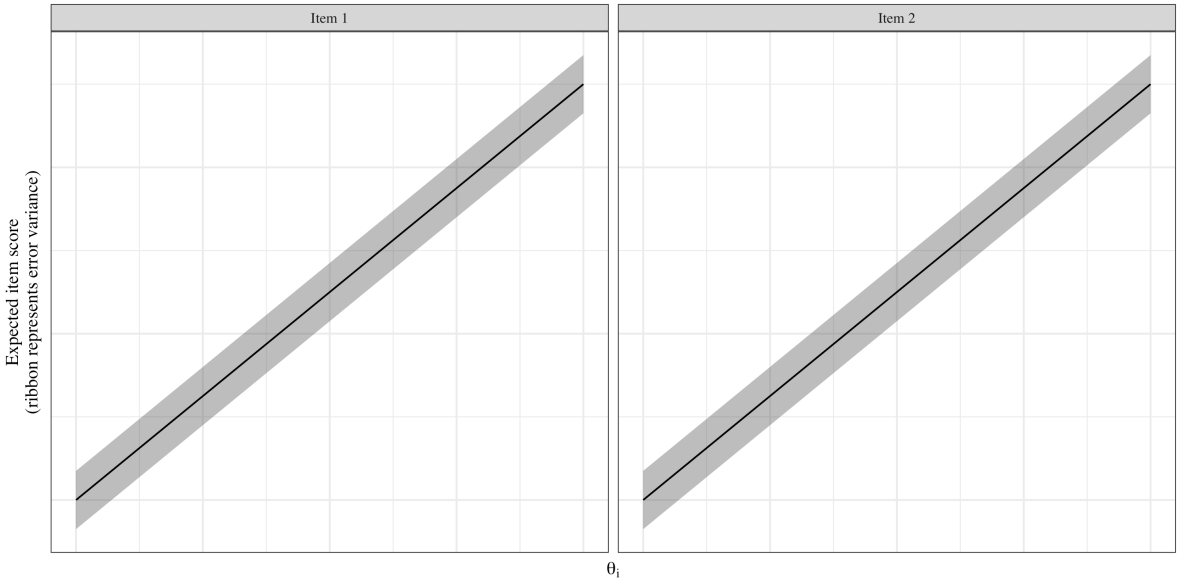
The true score T corresponds to what a person's score would be without any measurement error. In general, we approximate true scores in CTT by averaging many observed scores (e.g., averaging scores across items). Since scales of measurement are arbitrary in psychology, sum scores may also be used.

To note, the equation presented implies a linear relation between T and X : The item response changes linearly with the true score. Modern (Item-Response) Theory allows non-linear relations, as well as non-Gaussian items.

4.2.2 Models

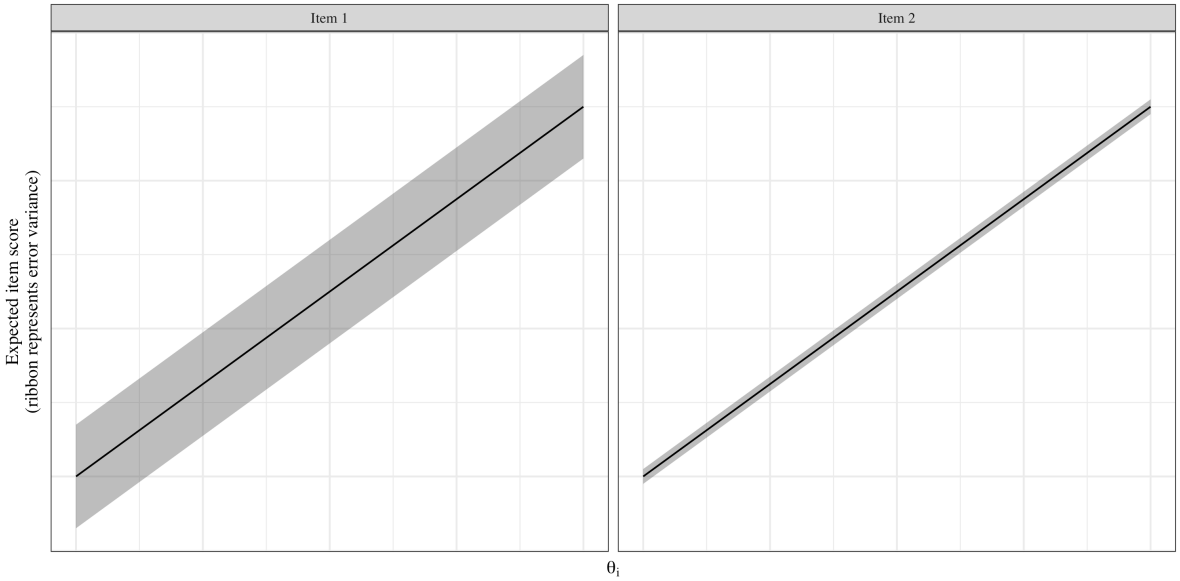
There are several more or less constrained variants of CTT (Lord, Novick, and Birnbaum 1968). What changes between them is whether the error variance and the true score changes by item. Throughout here, the person is represented with i, and the item with j.

- **Parallel CTT** (most constrained)
  - Items are completely interchangeable (nothing differs by item).
  - Average scores can be used as estimators of the true score.
  - Reliability can be predicted by the number of items (see Spearman-Brown reliability prophecy formula).

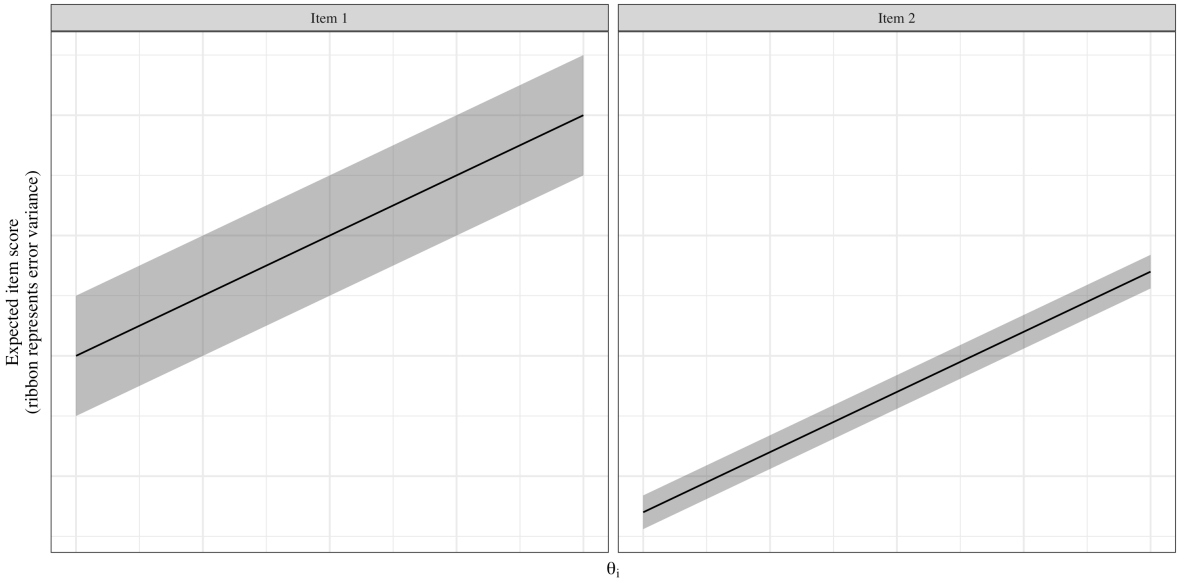


- **Tau-equivalent CTT**

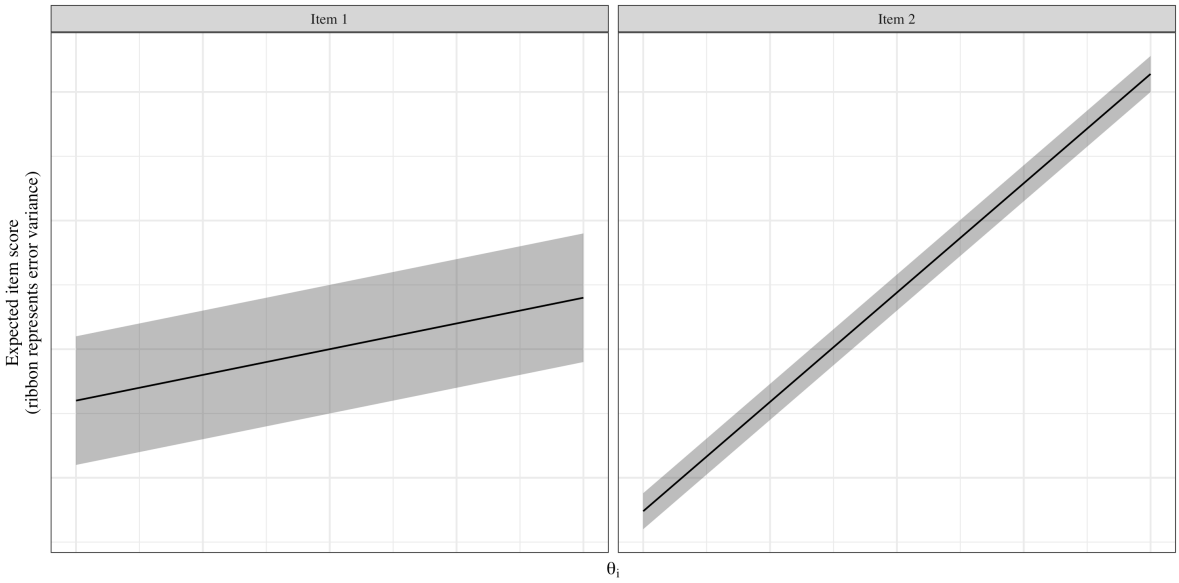
- Items differ in their error variance σ_j^2 : $E \sim \text{Normal}(0, \sigma_j^2)$. The true score remains the same across items.
- Average scores can be used as estimators of the true score.
- Reliability cannot anymore be predicted from the number of items.



- **Essentially tau-equivalent CTT**
  - Items differ in their error variance and true score. The true score is given as a person true score and an item location parameter: $T_{ij} = T_i + b_j$.
  - Compared with the previous model, two sets of items have different expected scores (i.e., a set of items may be more difficult than another).
  - Average scores are not estimators of the true score, but, given a set of items, they are perfectly correlated to true scores (so average score can still be used).



- **Congeneric CTT** (least constrained)
  - Items differ in their error variance and true score. The true score is given as a linear function of the person true score, item location parameter:, and item slope/discrimination/loading parameter $T_{ij} = a_j T_i + b_j$.
  - Average scores are not anymore perfectly correlated with the true score. Other methods of scoring (i.e., factor scoring) are advised (McNeish and Wolf 2020).



4.3 Other approaches

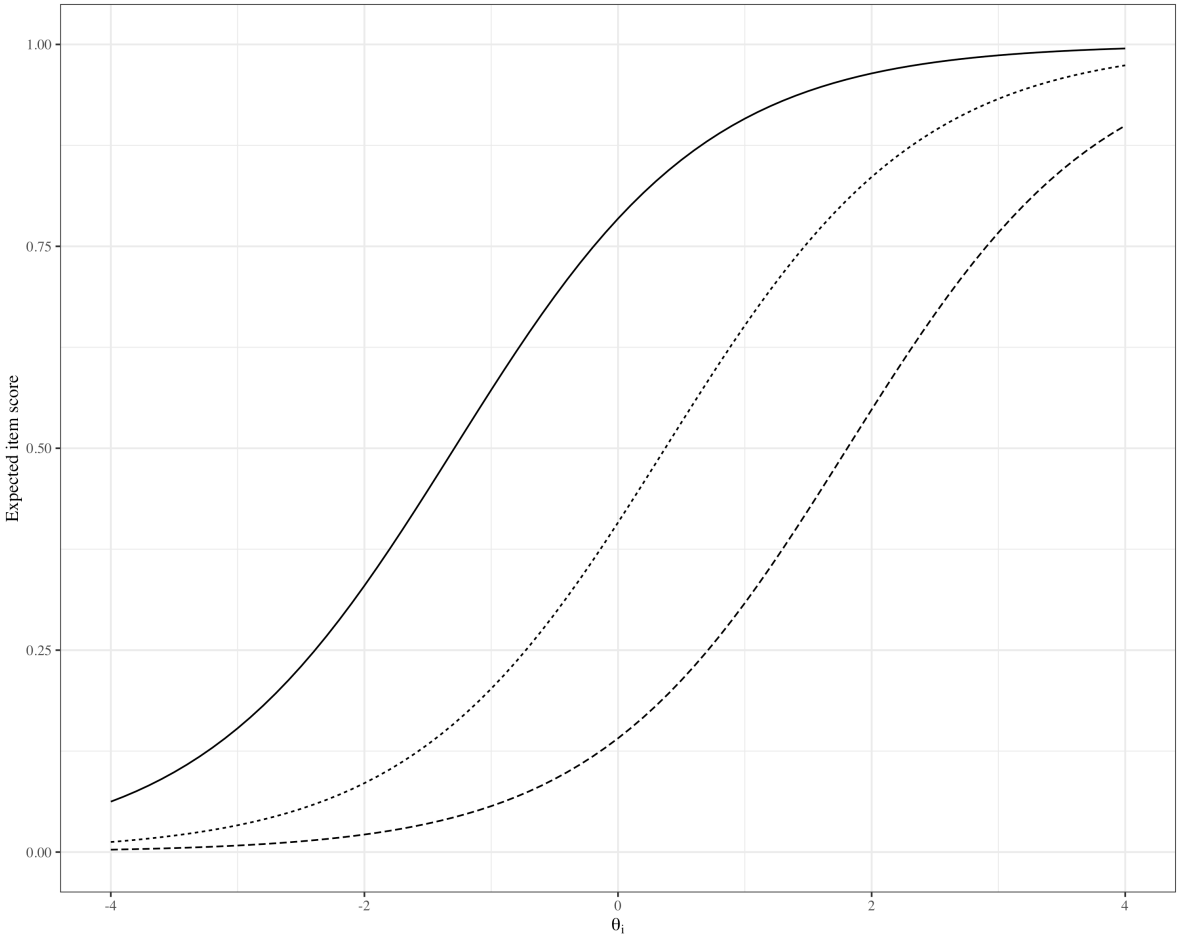
4.3.1 Item response theory

Item-response theory (IRT, i.e., modern test theory) is the most common alternative framework to CTT.

Like CTT, IRT describes item responses as a mathematical function of person attributes and item parameters(Ayala 2013). Unlike CTT, IRT allows this function to be non-linear and the item distribution to not be assumed Normal.

IRT may be seen as a generalization of CTT models (Mellenbergh 1994).

A typical example of IRT models are logistic models (see below), which are notably used for binary item responses (e.g., pass-fail) and ordinal models (e.g., Likert scales). For these the probability of responses (i.e., expected item score) is given as a function of the latent attribute (noted θ traditionally in IRT).



4.3.2 Behavior domain theory

According to behavior domain theory, item are not caused by psychological attributes, but they form them. Behavior domain theory uses models that are called formative models.

4.3.3 Network psychometric models

In network psychometric models, there are no latent variables causing item scores, and the relations between items are directly investigated.

5 Discrimination power

In testing, we measure attributes that are supposed to vary across individuals (e.g., extroversion). Thus, tests (and their items) are expected to yield different scores for different individuals. The extent to which a test (or an item) does so it called discrimination (or discriminating) power.

In general, discriminating power is studied by using univariate statistics (i.e., studying the distribution) of the items and the scores. These typically include:

- Graphical methods
  - Frequency histograms/Frequency (bar) plots
- Location (i.e., central tendency) indices
  - Mean (or pass rate for binary items)
  - Median
  - Mode
- Dispersion indices
  - Range
  - Standard deviation/variance

Good discrimination power is achieved when a test and its items produce variable scores, which often implies that the dispersion indices are large (wide range of scores, large standard deviation), and implies that the locations are not too close to the lower or upper bounds (which signals items that are too easy or difficult).

6 Reliability

A measure has good reliability if it attributes consistent scores to a given person. In other words, reliability is the degree to which a measure is free from random error.

6.1 Notation

Reliability is frequently noted as $\rho_{xx'}$.

However, reliability is generally unknown and estimated from a dataset. Estimators of reliability are often written as $r_{xx'}$.

6.2 Conceptual definition

In Classical Test Theory, reliability is viewed as the ratio of the true score variance σ_T^2 over the observed score variance σ_X^2:

$$\rho_{xx'} = \frac{\sigma_T^2}{\sigma_X^2}$$

In practice, it cannot be calculated directly, as σ_T^2 is unknown. We use instead estimators of reliability. In general, they are based on correlations between repeated measurements.

6.3 Relation to standard error of measurement

While reliability is generally presented as a feature of a test in general, the accuracy of a person's score is generally presented through the **standard error of measurement** (or through confidence intervals). The standard error of measurement is directly related to reliability: The higher the reliability, the smaller the standard error of measurement (and the more narrow the confidence intervals).

6.4 Test-retest reliability

In **test-retest reliability** (i.e., time stability), we verify that a person obtains scores that are consistent when they are measured several times.

The typical design to study it consists in having the same sample of respondents take the instrument twice (or more, but usually twice), with a time interval (usually 1 to 4 weeks).

Good test-retest reliability is achieved if a (very) strong high correlation (generally $r > .70$) is obtained between the time points.

6.5 Parallel forms reliability

Some tests exist in multiple similar ("parallel") forms (in general, to avoid test contamination and biases induced from training).

Parallel forms reliability is studied by having the same sample of respondents take the two (or more) parallel forms.

Good parallel forms reliability is achieved if a (very) strong high correlation (generally $r > .70$) is obtained between the scores of the forms.

6.6 Inter-rater reliability

Some tests require the intervention of raters, which may induce some random error.

Inter-rater reliability is studied by having the same sample of responses judged by a group of raters (2 or more).

Good inter-rater reliability is achieved if a (very) strong high correlation (generally $r > .70$) is obtained between the scores of the raters.

6.7 Internal reliability

Within a test, each item may be considered an indicator of the construct in of itself. Thus, there is some random error attached to each item.

Internal reliability is studied by administering the test in a sample of respondents. Internal reliability indices capture, using different approaches, the extent to which items produce consistent scores.

The most used measure of internal reliability is **Cronbach's** α (Cronbach 1951). In general, values above .70 are considered sufficient (the maximum is 1).

In general, α tends to increase with the number of items, and to decrease if the instrument is not unidimensional. It assumes the essentially tau-equivalent model. A common alternative to it is McDonald's omega (ω) (McDonald 1999) which can be computed from different models (e.g., congeneric).

Internal reliability indices are often used as decision rules to discard items. For example, the "alpha if deleted" method consists in discarding items without which Cronbach's α would be higher.

6.8 Predicting reliability

Once we have an estimate for reliability with a given number of items, assuming that all items are completely interchangeable (i.e., parallel CTT model), we can use the **Spearman-Brown** formula (Spearman 1910; Brown 1910) to compute reliability for any number of items.

Note: Assuming all items being interchangeable, reliability increases with the number of items.

6.9 Reliability in item-response theory

In Item-Response Theory, reliability is not constant for a group of respondents, but varies by person. Thus, reliability (as well as standard errors of measurement and confidence intervals) is generally studied as a function of the person's latent attribute (although group-level reliability measures can also be computed).

7 Validity

7.1 General definition

Validity is generally defined as the extent to a test measures what it purports to measure.

Test validity can also be seen as the extent to which there is sufficient evidence to interpret the results of a test for its intended purpose (Messick 1995).

7.2 Content validity

During the test process, we want to ensure that we build items that capture the construct of interest.

For this, we often use Subject Matter Experts (SMEs) to rate items. A typical procedure consists in asking a panel of SMEs to rate items on a 3 point scale (item is not useful/useful but not essential/essential), and to calculate for each item a **Content Validity Ratio** (CVR).

In general, CVRs at least above 0 and as close to 1 as possible are desirable. CVRs can be significance tested, and we may choose to retain items with CVRs that are significantly above 0 .

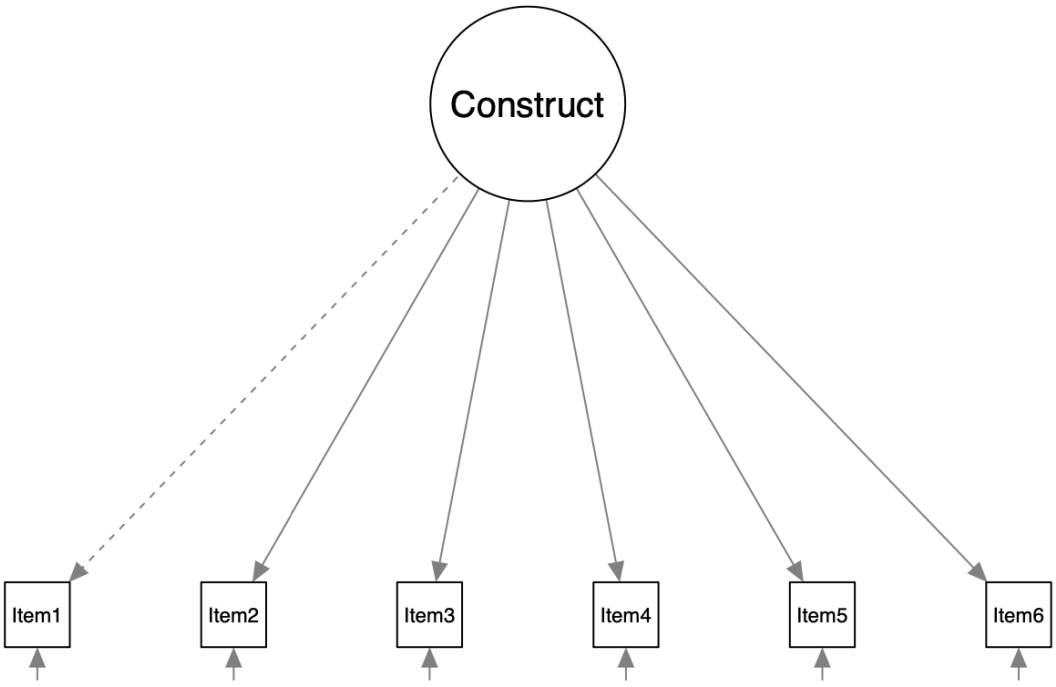
7.3 Structural validity

Structural validity (i.e., factorial validity) is the capacity of a test to produce items scored whose structure is in adequation with the theoretical structure of the test (e.g., if a test is supposed unidimensional, the empirical structure of the test should be unidimensional).

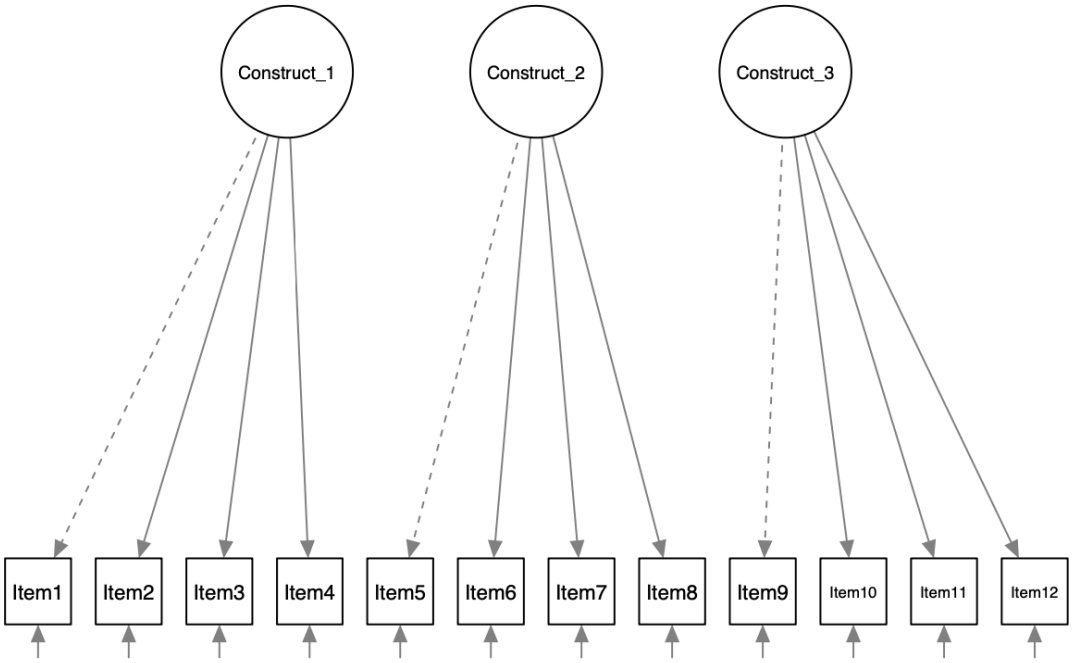
7.3.1 Overview of common structures

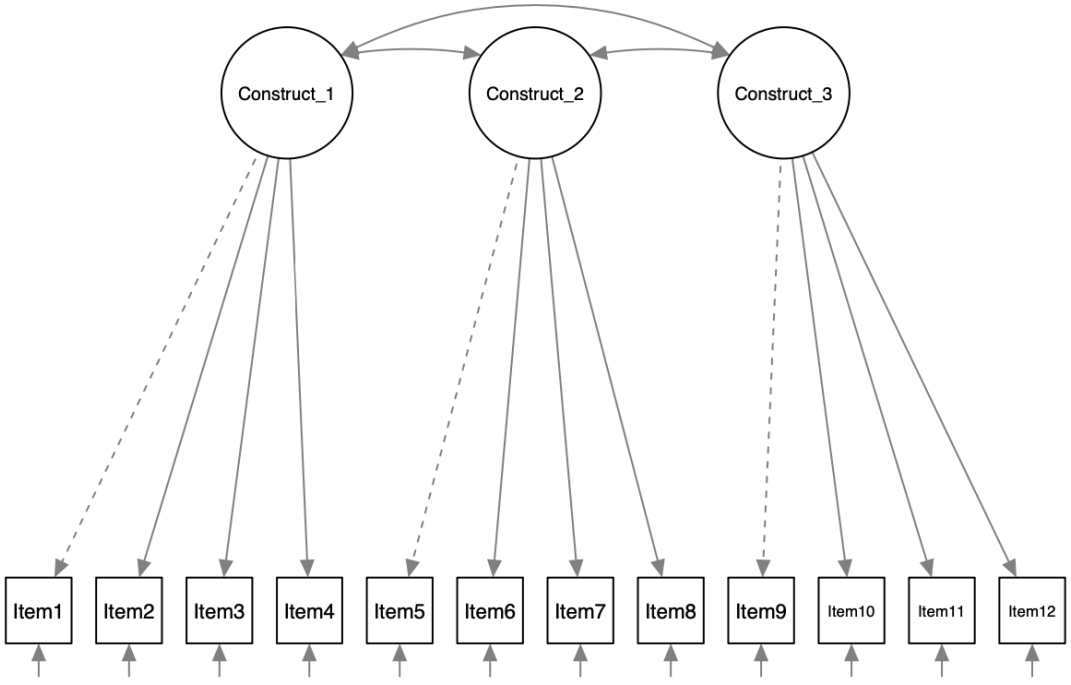
The name of the structures is generally given the number of latent attributes. They are often referred to as factors in this context.

- **Unidimensional structure**
  - one factor explains all item scores

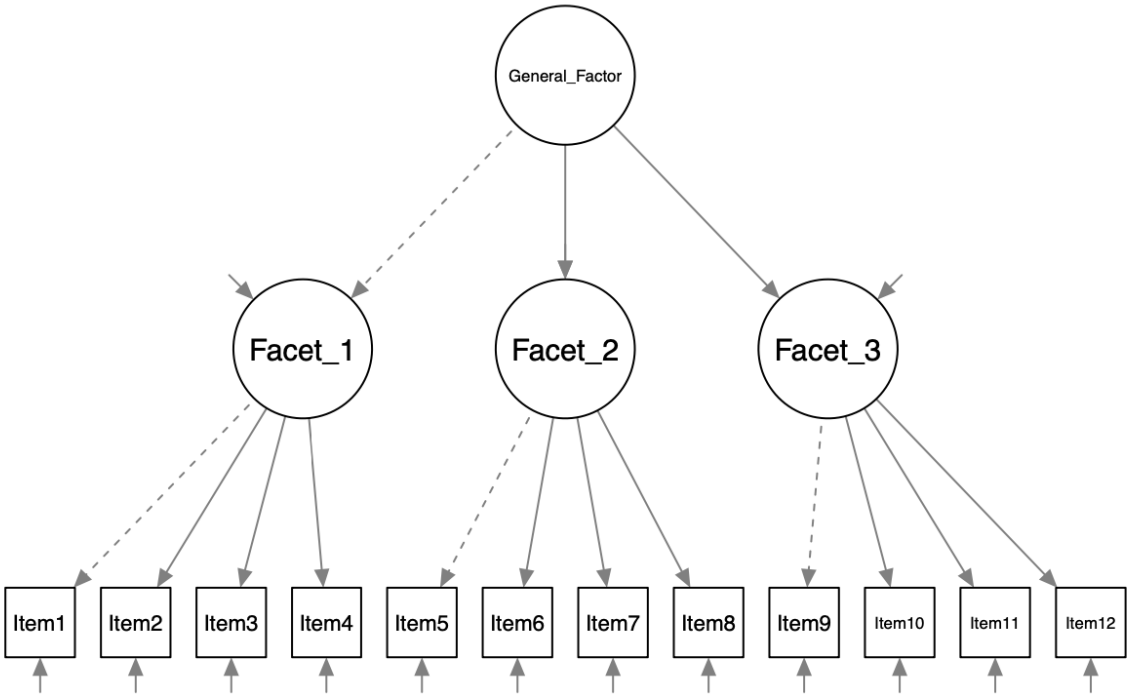


- **Multidimensional structure**
  - different factors explain different item scores (i.e., each item score is explained by one of the factors)
  - the factors may be **correlated** or **independent**

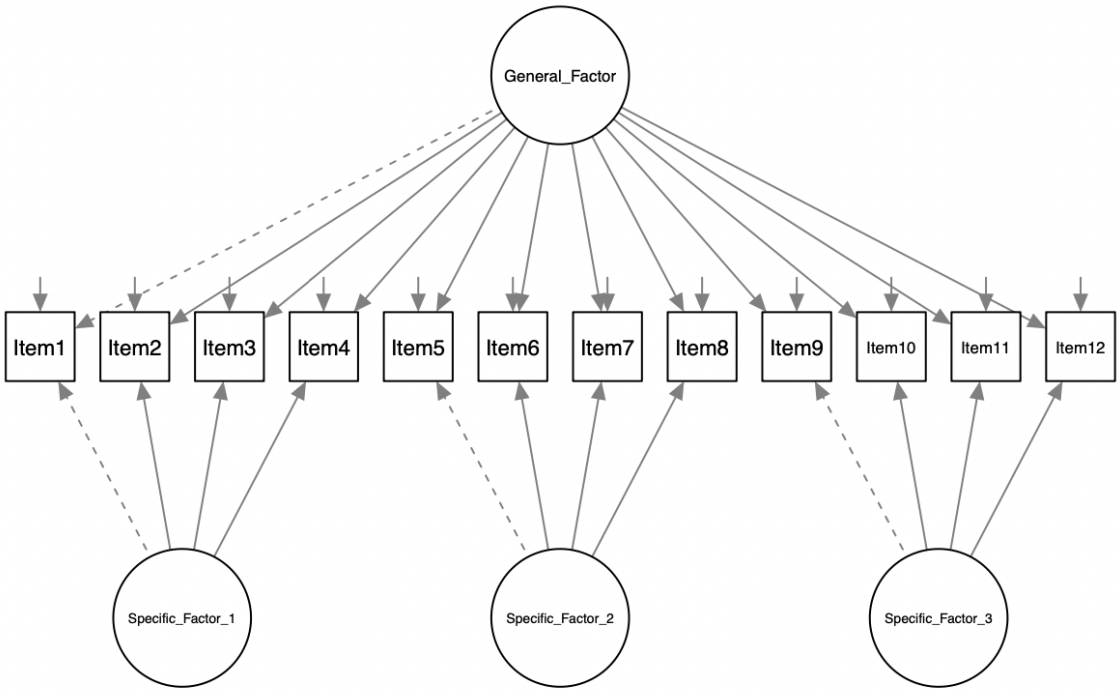

Construct_1
Construct_2
Construct_3
Item1
Item2
Item3
Item4
Item5
Item6
Item7
Item8
Item9
Item10
Item11
Item12


Construct_1
Construct_2
Construct_3
Item1
Item2
Item3
Item4
Item5
Item6
Item7
Item8
Item9
Item10
Item11
Item12

- **Hierarchical structure**
  - multidimensional structure, where the factors are not directly correlated but explained by a general (second-order) factor



- **Bifactor structure**
  - each item score is explained by a general factor and one of several specific factors



7.3.2 Correlation matrices

One approach to studying structural validity is to inspect the correlations between items, which is often done using a **correlation matrix**, where items that are expected to be clustered together (i.e., items that are supposed to be in the same scale/facet) are grouped together.

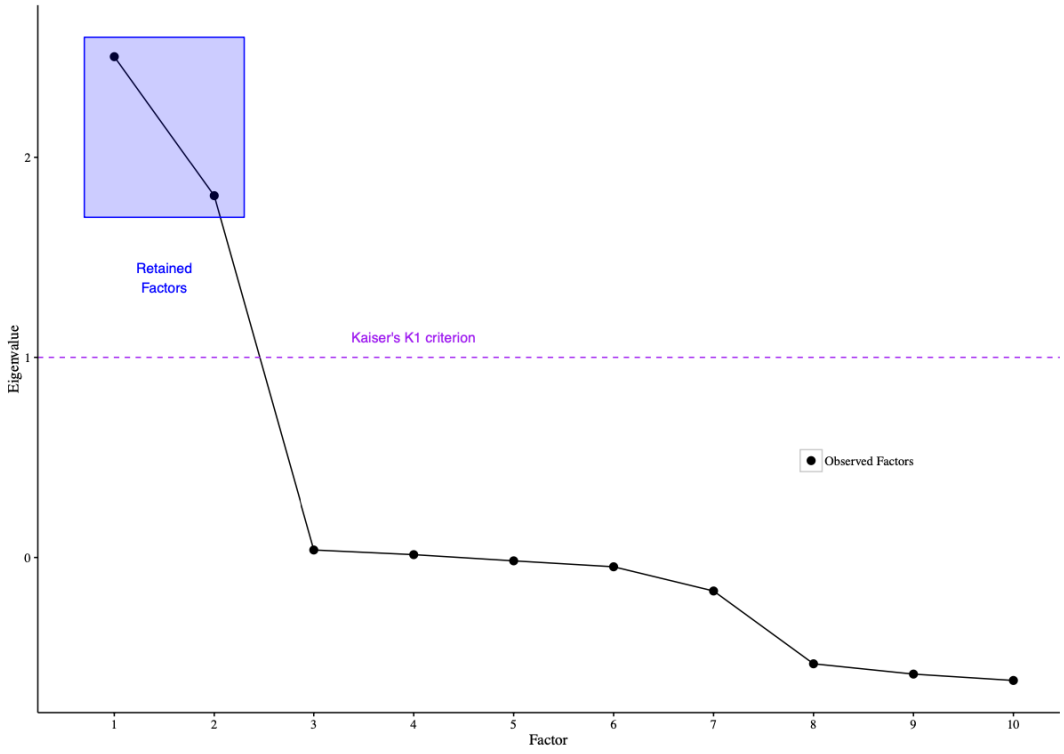
Correlation Matrix

		cog1	cog2	cog3	cog4	cog5	aff1	aff2	aff3	aff4	aff5
Correlation	cog1	1.000	.571	.666	.149	.692	.041	.088	.097	.051	.036
	cog2	.571	1.000	.483	.124	.545	.073	.080	.116	.070	-.066
	cog3	.666	.483	1.000	.127	.606	.021	.017	.116	-.006	-.068
	cog4	.149	.124	.127	1.000	.162	-.030	-.045	.005	.026	-.035
	cog5	.692	.545	.606	.162	1.000	.053	.023	.103	.110	-.021
	aff1	.041	.073	.021	-.030	.053	1.000	.578	.632	.587	.242
	aff2	.088	.080	.017	-.045	.023	.578	1.000	.584	.534	.167
	aff3	.097	.116	.116	.005	.103	.632	.584	1.000	.620	.218
	aff4	.051	.070	-.006	.026	.110	.587	.534	.620	1.000	.257
	aff5	.036	-.066	-.068	-.035	-.021	.242	.167	.218	.257	1.000

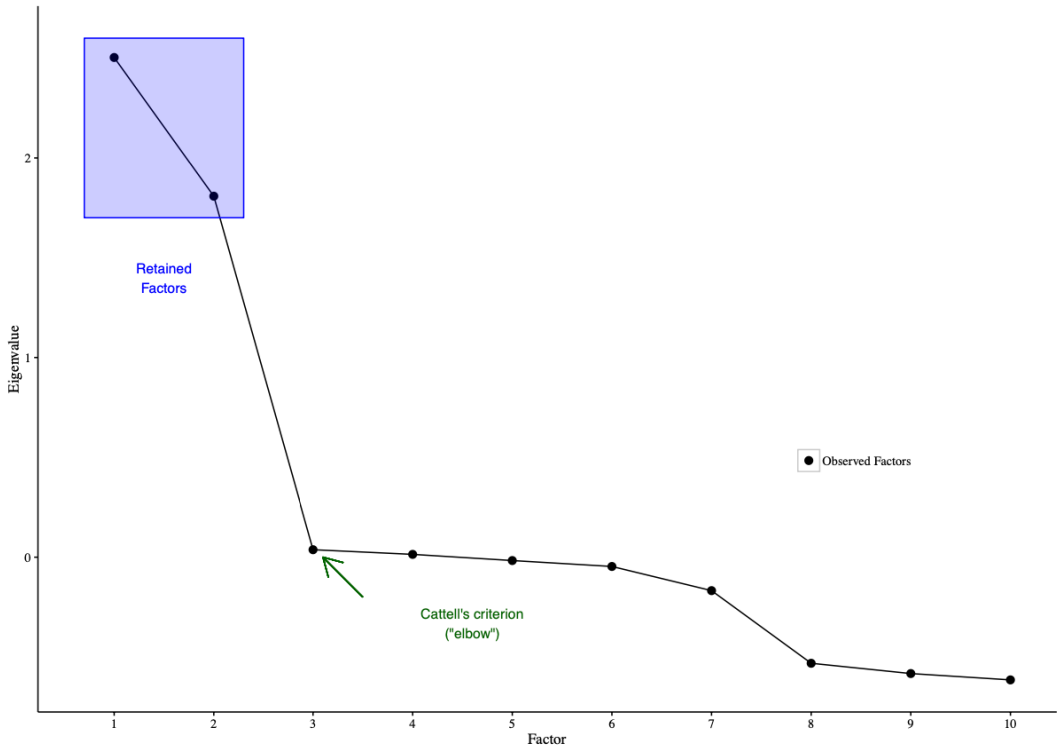
7.3.3 Exploratory factor analysis (EFA)

Exploratory Factor Analysis is an exploratory method that consists in identifying and interpreting a set of factors that parsimoniously explain the relations between the items.

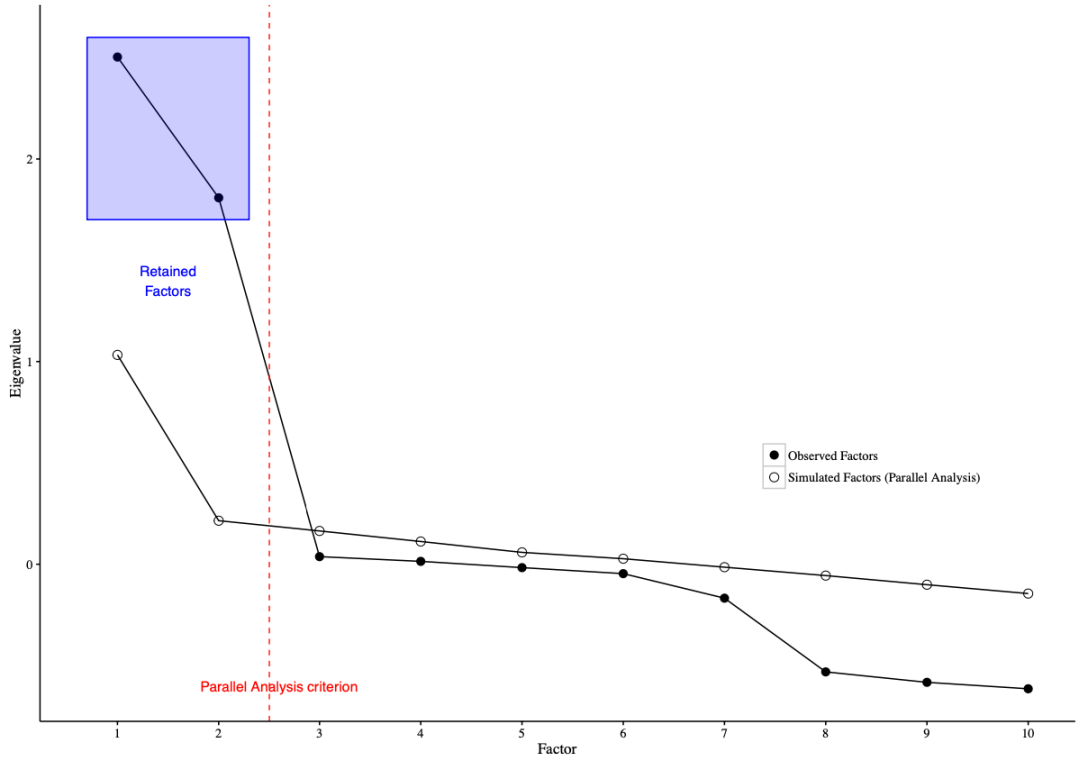
- Methods for estimating the factors (extraction methods)
  - **Maximum Likelihood Estimation** (preferred for continuous/ordinal with more than 4 response categories)
  - **Principal Axis Factoring** (preferred for binary/categorical/ordinal with 4 or less response categories)
  - **Principal Components Analysis** (outdated for psychometrics investigations)
- Methods for obtaining a parsimonious number of factors
  - Kaiser's **K1** criterion (removes factors with eigenvalues below 1)



  - Cattell's **scree test** (graphical technique that involves sorting factors by eigenvalues, and dropping all factors after a sharp drop in eigenvalues is observed) (Cattell 1966)



– Horn's **parallel analysis** (consists in simulating factors and retaining factors whose eigenvalues are higher than the simulated factors, or a certain percentile of their distribution) (Horn 1965)



  - **Interpretability** (rejects factors that are not clearly interpretable — see below)

- Rotation methods (facilitate interpretation)
  - **Orthogonal rotations** (e.g., **varimax** rotation) assume factors to be uncorrelated, should be used if there is a strong rational for hypothesizing independent factors
  - **Oblique rotations** (e.g., **oblimin** rotation) do *not* assume factors to be uncorrelated, should be used if the relations between factors are presumed unknown or if we hypothesize correlated factors

- Interpretation
  - **Factor loadings** represent the strength of the relation between items and factors. They are scaled similarly as a correlation coefficient.
  - In general, we want to look at tables that show loadings by item and factors. If an item has a strong loading ($> .50$ in absolute value) with a factor, we interpret this as the item being well represented by this factor.

– By looking at the content of the items and which factor they are the best explained by, we can (in most cases) interpret the factors conceptually (i.e., understand what attribute they correspond to).

– Items that do not have strong loadings on any retained factor are often discarded.

– Items that have strong loadings on a factor that they are not expected to be related to (we often referred to these as "cross-loadings" are also often discarded.

Pattern Matrix[a]

	Factor	
	1	2
cog1	.864	.025
cog2	.657	.058
cog3	.762	-.006
cog4	.183	-.027
cog5	.801	.035
aff1	-.001	.787
aff2	.018	.718
aff3	.074	.805
aff4	.019	.756
aff5	-.039	.292

Extraction Method: Maximum Likelihood. Rotation Method: Oblimin with Kaiser Normalization.

a. Rotation converged in 3 iterations.

Figure 1: Factor loadings (here, loadings indicate that factor 1 represents "cog" and factor 2 represents "aff").

7.3.4 Confirmatory factor analysis (CFA)

Confirmatory factor analysis consists in specifying a priori a factor structure (or several alternate structures), and investigating whether it explains well the relations between the items.

CFA is generally performed using Structural Equation Modeling (SEM), which is more extensively covered in later classes.

CFA notably produces:

- **Factor loadings** for all items, which are similarly interpreted as in EFA, and can be used to retain or discard items
- **Fit indices**
  - The χ^2 test should be non-significant (i.e., p should be larger than .05.
  - CFI, TLI should be above .90 (the higher the better)
  - RMSEA, SRMR should be below .08 (the lower the better)

7.3.5 Measurement invariance

Factor structures remain general assumptions about how a test works in a population. These assumptions may be more or less appropriate to different persons. The assumption that the same model (with the same parameters) represents accurately the functioning of a test for an entire population is referred to as the assumption of measurement invariance.

Structures and their parameters may vary by individual (e.g., it may vary by a demographic variable), and the study of these variations if generally referred to as the study of measurement invariance (differential item functioning in Item Response Theory).

7.4 External validity

In general, when instruments are developed, we have expectations regarding what construct is measured and what the scores should be (and should not) related to. External validity is the extent to which an instrument yields scores that are related to other measures in a manner that is conceptually consistent.

The measures that are used for comparison with our instrument are generally referred to as validity criteria.

In general, these comparisons are made using correlation coefficients.

7.4.1 Concurrent validity

In concurrent validity analysis, the instrument and the criterion are administered at the same time.

- **Convergent validity** verifies that an instrument and a criterion that should be related are indeed empirically correlated (e.g., the instrument you are building for depression is positively correlated with a measure of emotional exhaustion).
- **Divergent** (or **discriminant**) **validity** verifies that an instrument and a criterion that should be related are indeed empirically correlated (e.g., the instrument you are building for depression is not correlated with a measure of openness)

7.4.2 Predictive validity

In **predictive validity** analysis, the instrument is administered before the criterion. In general, analyses of predictive validity are used when the instrument is used in a prediction context (e.g., to predict health outcomes, educational outcomes).

7.4.3 Diagnostic validity

Diagnostic validity is the ability of a test to perform well of accurately categorizing individuals in clinical vs. non-clinical groups. In general we study both a test and its threshold for significance.

Diagnostic validity is maximized when we have both high sensitivity and specificity:

- **Sensitivity** is the ability of the test to identify correctly those who do have the disorder. (maximizing true positives)
- **Specificity** is the ability of the test to identify correctly those who do not have the disorder. (maximizing true negatives)

Types of error:

- **False positive** (i.e. **type I error)**: Concluding that an individual has the disorder when the individual does not have the disorder.
- **False negative** (i.e. **type II error**): Concluding that an individual does not have the disorder when the individual has the disorder.

Test	Disorder	No disorder
Positive	True positive	False positive (type I error)
Negative	False negative (type II error)	True negative

In general, moving the threshold yields a trade-off between sensitivity and specificity (i.e., maximizing sensitivity tends to minimize specificity, and vice-versa).

Depending on the context, we may favor specificity or sensitivity, or aim for a balance.

8 Practical use of tests

8.1 Test administration

Test administration procedures notably aim to reduce biases.

We generally consider preparing:

- The test environment (ventilation, space, noise, disruptions, material, etc.)
- The test taker (avoid emotional biases, stereotype threat bias, social desirability bias, put effort into engaging interest and cooperation, avoid test sophistication bias by providing explicit instructions)
- The rater/test scorer (provide clear instructions for rating)

8.2 Test scoring

8.2.1 Norm-referenced vs. criterion referenced tests

Scores are usually interpreted:

- By comparing the person's responses with the responses of others (e.g., compare with reference group means, compare using percentile rank).
  - In these cases the test is called a **norm-referenced test**.
  - Example: SAT,GRE,WAIS
- By using a fixed threshold decided in advance (e.g., a person has to succeed 70% of the items to pass).
  - In these cases the test is called a **criterion-referenced test.**
  - Example: Driving licence examinations, most (non "curved") examination tests.

8.2.2 Standard scores

Norm-referenced tests are often interpreted by transforming scores into standard scores. This allows easier interpretation and comparisons across tests.

The most basic format for standard scores is the z score. z scores are made so that they have in the reference group a mean of 0 and standard deviation of 1. Therefore, scores represent directly the distance to the mean (e.g., a z score of $-.04$ means that the individual's score is .04 standard deviations lower than the mean of their reference group).

Once a score has been z-transformed, it may be attributed a different mean and standard deviation. Common examples include:

- IQ scores have a mean of 100 and standard deviation of 15
- T scores have a mean of 50 and standard deviation of 10
- CEEB scores (e.g., SAT) have a mean of 500 and standard deviation of 100

They present the same advantages of z scores, but tend to facilitate communication of results to stakeholders.

8.3 Score uncertainty

Since all psychological measures are imperfect, it is important to measure and communicate the uncertainty about scores.

Common approaches include:

- Using the standard error of measurement
  - Represents the typical distance between a person's observed score and their true score
  - Decreases as reliability increases
- Using confidence intervals
  - Calculated so that it is estimated that, if the person were measured many times, 95% (or 99%, or 90%) of their scores would fall within the interval.
  - Are narrower as reliability increases.

8.4 Alternate scoring strategies

An alternate to sum/average scoring (and standardizing) scores consists in using latent variable models that can produce estimates for the person attributes.

This procedure is generally referred to as **factor scoring** (or using **factor scores**).

Common models that can produce factor scores:

- Exploratory factor analysis
- Confirmatory factor analysis
- Item-response theory models

The models used generally yield scores that are already standardized (on a z score scale). In general estimates of uncertainty (e.g., standard error of measurement) can also be obtained.

Note that the quality of the scores depends on the quality of the model, as well as of the data it was estimated on.

9 Ethics

The ethics of psychological testing for psychologists are described in the **Ethical Principles of Psychologists and Code of Conduct** (Standard 9) (http://www.apa.org/ethics/code/).

Key principles:

- Psychologists shall use valid tools that are not outdated.
- Information about psychometric qualities should be provided by test developers.
- The choice of tests should be justified, notably by psychometric qualities.
- Psychologists remain responsible for the quality of their instruments and their interpretation.
- Psychologists take into account the limitations of test results in their practice and decisions.
- Psychologists have to obtain informed consent to use tests.
- Psychologists must inform test takers in understandable language.
- Unless there is a lawful good reason (protecting people), test data must be provided to the client.
- Psychologists must give and explain results, unless it is explained in advance and justified by the purpose of the test.
- Psychologists should respect the integrity and security of test materials.
- Psychologists are responsible for the proper use of tests if not done by them.

References

Ayala, R. J. de. 2013. *The IRT Tradition and Its Applications*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199934874.013.0008.

Brown, William. 1910. "Some Experimental Results in the Correlation of Mental Abilities." *British Journal of Psychology, 1904-1920* 3 (3): 296–322. https://doi.org/10.1111/j.2044-8295.1910.tb00207.x.

Cattell, Raymond B. 1966. "The Scree Test for the Number of Factors." *Multivariate Behavioral Research* 1 (2): 245–76. https://doi.org/10.1207/s15327906mbr0102_10.

Cronbach, Lee J. 1951. "Coefficient Alpha and the Internal Structure of Tests." *Psychometrika* 16 (3): 297–334. https://doi.org/10.1007/BF02310555.

Horn, John L. 1965. "A Rationale and Test for the Number of Factors in Factor Analysis." *Psychometrika* 30 (2): 179–85. https://doi.org/10.1007/BF02289447.

Lord, F. M., M. R. Novick, and A. Birnbaum. 1968. *Statistical Theories of Mental Test Scores*. Statistical Theories of Mental Test Scores. Oxford, England: Addison-Wesley.

McDonald, Roderick P. 1999. *: A Unified Treatment*. New York: Psychology Press. https://doi.org/10.4324/9781410601087.

McNeish, Daniel, and Melissa Gordon Wolf. 2020. "Thinking Twice about Sum Scores." *Behavior Research Methods* 52 (6): 2287–2305. https://doi.org/10.3758/s13428-020-01398-0.

Mellenbergh, Gideon J. 1994. "Generalized Linear Item Response Theory." *Psychological Bulletin* 115 (2): 300–307. https://doi.org/10.1037/0033-2909.115.2.300.

Messick, Samuel. 1995. "Validity of Psychological Assessment: Validation of Inferences from Persons' Responses and Performances as Scientific Inquiry into Score Meaning." *American Psychologist* 50: 741–49. https://doi.org/10.1037/0003-066X.50.9.741.

Spearman, Charles. 1910. "Correlation Calculated from Faulty Data." *British Journal of Psychology* 3 (3): 271.