Advances in Psychometric Theory and Measurement for Psychological Sciences

With Tutorials in R

Rafael Valdece Sousa Bastos

2024-04-30

Table of contents

Overview 6

Suggest Edits and Improvements . . . 6

Support my work . . . 7

APA Citation . . . 7

BibTeX . . . 7

About the Author . . . 7

1 The Concept of Validity of Psychological Tests 9

1.1 Introduction . . . 9

1.2 A Brief Note on the History of Validity . . . 10

1.3 What is Validity Then? . . . 11

1.4 Sources of Validity . . . 12

1.4.1 Evidence of Content-Based Validity . . . 12

1.4.2 Evidence Vased on Response Processes . . . 12

1.4.3 Evidence Based on Internal Structure . . . 13

1.4.4 Evidence Based on its Relationships With External Variables . . . 13

1.4.5 Evidence Based on the Consequences of Testing . . . 13

1.5 Validity Crisis: How Validity is Done in Practice . . . 14

1.6 How to Move Forward? . . . 15

1.7 References . . . 16

2 Classical Test Theory 18

2.1 Overview of Classical Test Theory . . . 18

2.2 Reliability in CTT . . . 19

2.3 References . . . 20

3 Theory of Latent Variables in Psychometrics 21

3.1 The Importance of Psychometrics . . . 21

3.2 Misunderstandings in Psychometric Practice . . . 21

3.3 Obstacles to the Psychometric Revolution . . . 22

3.4 What are Constructs and Latent Variables . . . 22

3.5 Ways to Represent a Construct . . . 23

3.6 The Platonic Relationship of Cause and Effect in Psychological Measures . . . 24

3.7 References . . . 26

4 Measurement Theory: Why it is Possible to Measure Psychological Phenomena 29
4.1 The Objective of Psychometrics 30
4.2 Dr. Jekyll and Mr. Hyde: Measurement and Validity 30
4.2.1 History of Psychological Measurement 31
4.3 How to Move Forward: Defining Quantity 34
4.3.1 Additive Conjoint Measurement 36
4.4 The Tale of Taxometric Analysis 41
4.5 Does Rasch Modeling Entail Measurement? 42
4.6 The Direct Test of Conjoint Measurement Axioms 43
4.7 References 43

5 Exploratory Factor Analysis 47
5.1 Why is it Called Exploratory? 48
5.2 EFA Step-by-step 48
5.2.1 Data adequacy 48
5.2.2 Factor Retention 50
5.2.3 Factor Extraction 51
5.2.4 Factor Rotation 53
5.3 How to Run an Exploratory Factor Analysis in R 53
5.3.1 Data Adequacy in R 53
5.3.2 Parallel Analysis in R 55
5.3.3 Factor Extraction in R 56
5.4 How to report a Factor Analysis 59
5.5 References 59

6 Confirmatory Factor Analysis 61
6.1 What is it and When Do We Apply Confirmatory Factor Analysis? 61
6.2 Model Specification 61
6.3 Differences Between Saturated/Unrestricted Model and Restricted Model . . . 62
6.4 Model Identification 65
6.5 Fit Indices 66
6.5.1 Fit Indices Commonly Used in Factor Analysis 66
6.5.2 Criticisms of Factor Analysis 67
6.6 How to Run a Confirmatory Factor Analysis in R. 67
6.7 References 77

7 Exploratory Graph Analysis 80
7.1 About the Method 80
7.2 Unique Variable Analysis (UVA) 81
7.3 Exploratory Graph Analysis 82
7.3.1 Partial Correlation 82
7.3.2 What is EBICglasso 82
7.3.3 The Clustering Algorithm: An Example with Walktrap 83

7.4 EGA's Stability . . . 83
7.5 How to Run the EGA Step By Step in R . . . 84
7.5.1 UVA in R . . . 84
7.5.2 EGA in R . . . 86
7.5.3 EGA's Stability in R . . . 87
7.6 References . . . 94

8 Social Desirability Bias 96
8.1 Faking: The Good, The Bad, and the Ugly . . . 96
8.2 Faking Good: Social Desirability . . . 97
8.3 How to Represent Social Desirability . . . 97
8.4 Modeling Faking with Classical Test Theory . . . 98
8.5 Some Ways to Control Social Desirability . . . 99
8.5.1 Correlations and Social Desirability Scales . . . 99
8.5.2 Ferrando et al. (2009) . . . 100
8.5.3 Leite and Cooper (2010) . . . 101
8.5.4 Ziegler and Buehner (2009) . . . 101
8.5.5 Peabody quadruplets (1967) . . . 102
8.6 How to Control Social Desirability in R . . . 103
8.6.1 Controlling Desirability with Ferrando et al. (2009) . . . 103
8.6.2 Controlling with MIMIC and Quadruples (Bastos & Valentini, 2023) . . . 108
8.7 References . . . 122

9 Acquiescence Bias 127
9.1 Acquiescence: trait or state? . . . 128
9.2 Problems Related to Acquiescence: Control Through Inverted Items . . . 128
9.3 The Removal of Acquiescence: Statistics and Design . . . 130
9.4 How to Control Acquiescence in R . . . 131
9.4.1 Controlling Acquiescence with Ferrando et al. (2009) . . . 131
9.4.2 Controlling Acquiescence with Random Intercepts . . . 136
9.4.3 Controlling Acquiescence with Random Intercepts Exploratory Graph Analysis . . . 141
9.5 References . . . 147

10 Measurement Invariance/Equivalence 152
10.1 Defining Measurement Invariance/Equivalence . . . 152
10.2 Measurement Invariance for Categorical Variables . . . 154
10.3 Measurement Invariance in R . . . 154
10.3.1 Baseline Model . . . 156
10.3.2 Thresholds Invariance Model . . . 157
10.3.3 Thresholds and Factor Loadings Invariance Model . . . 158
10.4 Data Interpretation . . . 160
10.5 References . . . 160

11 Non-Parametric Item Response Theory 162

11.1 (Non)-Parametric Tests . . . 162

11.2 Mokken Scaling . . . 162

11.3 Scalability Coefficient . . . 164

11.4 Mokken Scaling in R . . . 165

11.4.1 Dimensionality Analysis in R . . . 166

11.4.2 Latent Monotonicity Analysis in R . . . 167

11.4.3 Local Independence in R . . . 168

11.4.4 Non-intersection Analysis of Item Response Functions . . . 168

11.5 References . . . 168

12 How to Validate Psychological Tests: A Step by Step Guide 170

12.1 Define the Objective of the Test . . . 170

12.2 Literature Review . . . 170

12.3 Should I Build or Adapt an Instrument? . . . 171

12.4 Which Model Should I Choose? . . . 172

12.5 Think About How the Participant Will Answer the Questionnaire . . . 172

12.6 Item Construction . . . 173

12.7 Judges' Assessment and Item Writing . . . 175

12.8 Target Population Assessment . . . 176

12.9 Internal Structure Assessment: Training Stage . . . 177

12.10Internal Structure Assessment: Validation Stage . . . 177

12.11Other Evidence of Validity . . . 178

12.12Extra Observations . . . 178

Overview

> Academic discourse and its grammar not only are like a forest that doesn't allow us to distinguish between individual trees but also go a step further, forcing the researcher to cut the trees down in order to understand the forest. **Paul B. Preciado, Countersexual Manifesto**

Psychometrics is an area that is growing rapidly worldwide: while many still measure psychological phenomena without having evidence that the measurement has some quality or validity evidence, others continue on the more difficult, but much more precise and scientific, path. The book ***Advances in Psychometric Theory and Measurement for Psychological Sciences: With Tutorials in R*** is the first book that makes psychometric science and measurement theory more accessible to you, the reader. Covering classical and modern views on psychometrics, I introduce concepts crucial to understanding psychometrics. This book is aimed at students and teachers who wish to use and understand psychometrics in depth. It's a great resource for studying alone or for study groups.

In this book I teach the theory behind the concept of the analysis, but I also teach how to run the analysis in the R programming language. We use R because it is free, accessible, and easy to use. I know you're afraid of using R, but I'll do everything I can to make sure you understand what you're going to do in your analyses.

This book is intended to be an educational resource for everyone. The content is written in an accessible way and it is possible to deduce conclusions, without being shallow or wrong.

This book integrates information from my website and my scientific work , in addition to several other works that are at the cutting edge of knowledge. The purpose of this book being online is to make the information more accessible and easier to update in the future.

Suggest Edits and Improvements

If you find any mistakes or have suggestions for improvement, you can submit an issue on the GitHub page of this open educational resource. You can also download a PDF or epub version (click the download button in the menu on the top left). This work is shared under a CC-BY-NC-SA License.

Support my work

Please help keep this book maintained and free for everyone by supporting me through my Ko-fi page.

APA Citation

You can cite this book as:

Bastos, R. V. S. (2024). *Advances in Psychometric Theory and Measurement for Psychological Sciences: With Tutorials in R*. https://doi.org/10.5281/zenodo.11094831

BibTeX

```
@book{Bastos2024,
  title     = "Advances in Psychometric Theory and Measurement
  for Psychological Sciences: With Tutorials in R",
  author    = "Bastos, Rafael V. S.",
  publisher = {},
  year      = 2024,
  doi       = https://doi.org/10.5281/zenodo.11094831,
  url       = {}
}
```

About the Author

The current maintainer of the book is Rafael Valdece Sousa Bastos, the creator of this educational resource. I hold a Master's degree (M.Sc.) in Psychology, specializing in the construction, validation, and standardization of measurement instruments from São Francisco University. Additionally, I have a Bachelor's in Psychology from the Pontifical Catholic University of Rio de Janeiro (PUC-RIO). My expertise spans across various domains within psychology, with a primary focus on the following areas:

- Psychometrics
- Statistics
- Meta-Science
- Quantitative Methods
- Measurement Theory

- Construction of Instruments

Figure 1: Photo of the author Rafael Valdece Sousa Bastos

I dedicate this work to Phoebe, my lifelong love and partner.

1 The Concept of Validity of Psychological Tests

What does it mean to say that a test has evidence of validity? Where do the concepts of validity in psychology come from? How can we seek evidence of validity for our instruments? In this chapter, we will answer these questions.

1.1 Introduction

In physics, we usually have an instrument that physically exists and measures physical properties. For example, an instrument that measures length uses this property (i.e., length) to measure the length of another object. Therefore, there is no need to prove that this property is congruent with the same property of the object being measured.

However, there are some cases where this is not so clear. For example, if we are measuring speed using the Doppler effect (Doppler Effect is a physical wave phenomenon that occurs when there is relative approach or distance between a source of waves and an observer), where the approach/distance of the spectral lines of the galaxy's lights is the instrument. In this case, we have the problem of the validity of an instrument, as we need to know whether or not it is true that the distance between the spectral lines is related to speed. To do this, we have to prove it empirically. Validity is common in areas of knowledge that use indirect or derived measures. The same thing that happens with the Doppler effect is very common in behavioral and psychological sciences (for example, psychology and education), especially if we are using the concept of construct (for example, happiness, anxiety or attraction).

From a psychological perspective, we can think of a construct as a characteristic that is inside our heads. These characteristics, like someone's personality, cannot be assessed through direct means. What we do, instead, is measure a person's behaviors, thoughts, emotions, affects, and infer that they come from the same construct (or not).

Of course, we have many ways to measure constructs, a common way is through questionnaires, where people respond each item on a scale of 1 (strongly agree) to 5 (strongly disagree), for example. Let's say we're going to measure self-efficacy in the workplace. We developed the items based on the definition of self-efficacy and then what? How can we know what our test results mean? Is self-efficacy a single phenomenon or can it be divided into different aspects? This is the role of seeking validity.

The need for valid measures seems obvious enough, given that to test theories that relate theoretical constructs (e.g., construct A influences construct B for individuals drawn from population P under conditions C), it is necessary to have valid measures of these constructs. Thus, even successful and replicable tests of a theory may be false if the measures lack construct validity; that is, they do not measure what researchers assume they are measuring (Schimmack, 2021).

1.2 A Brief Note on the History of Validity

A. 1900–1950: The hegemony of content validity

At that time, personality theories were the bomb. Most theories (such as psychoanalytic, gestalt, and phenomenology) generally had little empirical reasoning. In this context, personality trait tests were considered valid to the extent that the content of the test corresponded to the content of the theoretically defined traits.

B. 1950–1970: Prevalence of criterion validity

Behaviorism was very influential for Psychology and, of course, for Psychometrics. The tests were composed with a sample of behaviors that were expected to predict other behaviors or future behaviors. These tests were valid if they accurately predicted behavior in the future (or in another time), becoming the new path of validity (called criterion validity). It didn't matter why the test predicted the behavior, as long as they predicted it, and that was enough for its validity. As we can imagine, there was a shift from theoretical thinking to a focus on statistics. Rather than constructing a test to measure a construct, items were selected from a pool of items that appeared to refer to what they wanted to measure, essentially using statistical analysis to solve their problems.

C. 1970-Today: The rise of construct validity

After an article by Cronbach and Meehl in 1955 on a trinitarian model of validity (content, criterion, and construct), there was a change in the way of thinking about validity. The theory was back in play due to factors such as:

1. The need to develop a theory of personality and intelligence on an empirical basis, using factor analysis.

2. Studies of cognitive processes.
3. Studies of information processes.
4. Dissatisfaction with the results of using the test in education and work situations.
5. The impact of Item Response Theory.

Cronbach and Meehl note that construct validation is necessary

> whenever a test is to be interpreted as a measure of some attribute or quality which is not “operationally defined (p. 282).

This definition makes clear that there are other types of validity (e.g., criterion validity) and that not all measures require construct validity. However, studies of psychological theories that relate constructs require valid measures of these constructs to test psychological theories. Thus, construct validity is the relationship between variation in observed scores on a measure (e.g., scores on a Likert scale) and a latent variable that reflects corresponding variation in a theoretical construct (e.g., Extraversion; i.e., people who feel more energized by social interactions).

However, the problem of construct validaty can be illustrated with the development of IQ tests (Schimmack, 2021). IQ scores can have predictive validity (e.g., graduate school performance) without making any claims about the construct being measured (IQ tests measure whatever they measure, and what they measure predicts important outcomes). However, IQ tests are often treated as measures of intelligence. For IQ tests to be valid measures of intelligence, it is necessary to define the construct of intelligence and demonstrate that observed IQ scores are related to unobserved variation in intelligence. Thus, construct validation requires clear definitions of constructs that are independent of the measures being validated. Without a clear definition of constructs, the meaning of a measure essentially reverts to “whatever the measure is measuring”, as in the old adage “Intelligence is whatever IQ tests are measuring” (Schimmack, 2021).

1.3 What is Validity Then?

The **classic definition** of validity is “when the test measures what it is supposed to measure, what the test measures, and how well it measures” (Baptista & de Villemor-Amaral, 2019). However, the classical definition makes it appear that tests are either valid or not. To change this dichotomous paradigm, the **current definition** of validity is “the degree to which theory and evidence support the interpretation of test results. Thus, for each context/purpose of test use and for each intended interpretation it is necessary that test results have evidence of validity” (Baptista & de Villemor-Amaral, 2019). Now, we can say that each measurement has its own degree of validity. Validity is not a property of the test, but a property of the interpretation of test scores.

1.4 Sources of Validity

As I will explain below, there are different sources of validity. Each of them contributes to the search for the greatest "degree to which theory and evidence support the interpretation of test results". In general, it is always good to look for lots of evidence of validity, always updating this evidence over time.

1.4.1 Evidence of Content-Based Validity

You will collect data regarding the representation of a test's items, investigating whether they are samples of the domain they want to measure. The set of items is judged regarding its scope, with a view to evaluating the proposed construct. In general, it is based on the evaluation of experts, where they evaluate the importance of the items, taking into account their relationship with the aspects to be evaluated. However, it's also important to have the evaluation of the targeted population you will measure. Some statistical tests can be used, such as the percentage of agreement and the Kappa coefficient.

Example: In a paper, Bastos et al. (2022) created a measure of self-perception of prejudice and discrimination for different social groups. The authors used the following procedure to seek content-based validity:

1. Literature review on existing measures of prejudice and discrimination.
2. Self-perceived prejudice is defined as the perception that a person is the victim of negative attitudes towards themselves based on their social group; and self-perceived discrimination as the perception that a person is the victim of negative and unjustified behavior towards themselves based on their social group.
3. Based on these definitions and previous measures, the authors developed new items for other social groups.
4. After creating the items, they sent them to experts (i.e., psychologists and psychometricians) so they could evaluate the items.
5. Based on the proportion of agreement, the authors selected nine items for future analysis.

1.4.2 Evidence Vased on Response Processes

You will collect data on the mental processes involved in performing certain tasks. Normally this is an individual response process, and researchers ask the person being evaluated about the cognitive path used to reach a certain result. As an example, we can see that Noble et al. (2014) sought this type of validity with their study. They found that English language learners (ELL) students had lower scores on high-stakes tests compared to non-English language learners. Based on the interview, they found that

> ELL students' interactions with specific linguistic features of test items often led to alternative interpretations of the items that resulted in incorrect responses.

1.4.3 Evidence Based on Internal Structure

You will collect data on the correlation structure of items assessing the same construct. Statistical tests that are frequently used are Exploratory Factor Analysis (EFA), Confirmatory Factor Analysis (CFA).

As an example, we can use the article by Selau et al., (2020). The authors wanted to measure intellectual disability in children aged 7 to 15. They investigated the internal structure of the scale through EFA and CFA where items are divided into social, conceptual and practical factors that are explained by a higher order factor called adaptive function.

1.4.4 Evidence Based on its Relationships With External Variables

You will collect data on the pattern of correlations between test scores and other variables that measure the same or different constructs. Typically, to obtain this type of validity, researchers use the correlation of test scores with other variables. This type of validity can be:

1. Evidence of the ability of an instrument **to predict the assessed construct**.
2. When we have tests that **measure the same construct**, we expect them to be closely related.
3. When we have tests that **measure related constructs**, we expect them to be moderately related.
4. When we have tests that **measure different constructs**, we expect them to be unrelated.

Beymer et al. (2021) developed a Cost Perceptions of University Students scale. They correlated scale items with students' perceptions and values. They expected (and found) that "costs" were negatively correlated with "expectations" and "value" (you can see the definition of each variable in their article).

1.4.5 Evidence Based on the Consequences of Testing

Examine the intended or unintended social consequences of the use of a test, to verify that its use is providing the desired effects, in accordance with the reason for which it was constructed. Tests have this type of validity if they are being used for the same reason they were created. Although you cannot predict what people will do with an instrument you have developed, the responsibilities of instrument authors need to be discussed.

As an example, we can think of IQ measures. Its purpose is to measure people's intelligence. However, we can see that at times in history IQ was being used to justify racism.

1.5 Validity Crisis: How Validity is Done in Practice

We can see that there are a series of steps to ensure that our measure of psychological characteristics has degrees of validity. By following these procedures, we have more confidence to infer about the relationships between psychological traits and other variables. In practice, people generally look for only three types of validity: content, internal structure, and relationships with other variables. I think there are two reasons why this happens:

1. The difficulty of seeking validity based on the response process and the consequences of the test. Seeking validity based on the response process requires researchers to invest more time and money in interviewing enough participants. Seeking validity based on testing consequences is difficult. Authors are required to think about and predict their use in the recent and distant future, and some consequences may be (almost) impossible to predict.
2. The authors don't think it's their job to pursue these two types of validity, because they both: a) don't think it's their responsibility what people do with their work; b) they think their measurement is incredible and has no flaws, which may be true, but there is a lot to consider before concluding this, and that thing is making sure that some other response bias is not interfering with the results.

Only 12 years ago (in 2012) psychologists became aware that the field of psychology has a replication crisis (Schimmack, 2021). Many published results do not replicate honest replication attempts that allow the data to decide whether a hypothesis is true (Open Science Collaboration, 2015). However, unfortunately, low replicability is not the only problem in psychological science. Schimmack (2021) argues that psychology not only has a replication crisis, but also a validation crisis for psychological instruments.

Cronbach and Meehl make it clear that they were skeptical about the construct validity of many psychological measures.

> For most tests intended to measure constructs, adequate criteria do not exist. This being the case, many such tests have been left unvalidated, or a finespun network of rationalizations has been offered as if it were validation. Rationalization is not construct validation. One who claims that his test reflects a construct cannot maintain his claim in the face of recurrent negative results because these results show that his construct is too loosely defined to yield verifiable inferences (p. 291).

Nothing much has changed in the world of psychological measurement (Schimmack, 2021). For example, the study by Flake et al. (2017), where they reviewed current practices and found that reliability is often the only criterion used to claim construct validity. However, the reliability of a single measure cannot be used to demonstrate construct validity because reliability is only necessary but not sufficient for validity.

Thus, many articles do not provide evidence for construct validity and even if the evidence was sufficient to assert that a measure is valid, it is still unclear how valid a measure is. Another sign that psychology has a validity crisis is that psychologists today still use measures that were developed decades ago (Schimmack, 2010). Measures could be highly valid, it is also likely that they have not been replaced with better measures because quantitative assessments of validity are lacking. For example, Rosenberg's (1965) 10-item self-esteem scale is still the most widely used measure of self-esteem (Bosson et al., 2000; Schimmack, 2021). However, the construct validity of this measure has never been quantified, and it is unclear whether it is more valid than other measures of self-esteem (Schimmack, 2021).

1.6 How to Move Forward?

Although there is general agreement that current practices have serious limitations (Kane, 2017; Maul, 2017), there is no general consensus on the best way to deal with the validation crisis. Some researchers suggest that psychology can do better without quantitative measurement (Maul, 2017), but this is clearly false, and seeks an alternative without empirical foundation for why other methods are better or worse than quantitative science. If psychologists had followed Meehl's advice to quantify validity, psychological science would have made more progress than where we are currently (Schimmack, 2021).

Others believe that the view advocated by Cronbach and Meehl is too ambitious (Kane, 2016, 2017).

> Where the theory is strong enough to support such efforts, I would be in favor of using them, but in most areas of research, the required theory is lacking. (Kane, 2017, p. 81).

This may be true for some areas of psychology, such as educational testing, but it is not true for basic psychological science, where the sole purpose of measures is to test psychological theories. In this context, construct validation is crucial for testing causal theories. The industrial literature shows that it is possible to estimate construct validity even with rudimentary causal theories (Cote & Buckley, 1987), and there are some examples in social and personality psychology in which structural equation modeling has been used to quantify validity (Schimmack , 2021, Schimmack, 2010; Zou et al., 2013). Thus, the improvement of psychological science requires a quantitative research program on construct validity that focuses more firmly on the endeavor of always seeking evidence of the validity of its instruments.

1.7 References

Baptista, M. N. & de Villemor-Amaral, A. E. (2019). *Compêndio de avaliação psicológica*, Editora Vozes.

Bastos, R. V. S., Novaes, F. C., & Natividade, J. C. (2022). Self-Perception of Prejudice and Discrimination Scale: Evidence of Validity and Other Psychometric Properties. *Trends in Psychology*, 1-19. https://doi.org/10.1007/s43076-022-00190-7

Beymer, P. N., Ferland, M., & Flake, J. K. (2022). Validity evidence for a short scale of college students' perceptions of cost. *Current Psychology*, *41*(11), 7937-7956. https://doi.org/10.1007/s12144-020-01218-w

Bosson, J. K., Swann, W. B., Jr., & Pennebaker, J. W. (2000). Stalking the perfect measure of implicit self-esteem: The blind men and the elephant revisited? *Journal of Personality and Social Psychology*, *79*(4), 631-643. https://doi.org/10.1037/0022-3514.79.4.631

Cote, J. A., & Buckley, M. R. (1987). Estimating trait, method, and error variance: Generalizing across 70 construct validation studies. *Journal of Marketing*, *24*, 315-318. https://doi.org/10.1177/002224378702400308

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological bulletin*, *52*(4), 281. https://psycnet.apa.org/doi/10.1037/h0040957

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science*, *8*(4), 370–378. https://doi.org/10.1177/1948550617693063

Kane, M. T. (2016) Explicating validity. *Assessment in Education: Principles, Policy & Practice*, *23*, 198-211. https://doi.org/10.1080/0969594X.2015.1060192

Kane, M. T. (2017) Causal interpretations of psychological attributes. *Measurement: Interdisciplinary Research and Perspectives*, *15*, 79-82. https://doi.org/10.1080/15366367.2017.1369771

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*, (6251), 943-950. https://doi.org/10.1126/science.aac4716

Maul. A. (2017). Moving beyond traditional methods of survey validation. *Measurement: Interdisciplinary Research and Perspectives*, *15*, 103-109. https://doi.org/10.1080/15366367.2017.1369786

Pasquali, L. (2017). *Psicometria: teoria dos testes na psicologia e na educação.* Editora Vozes Limitada.

Rosenberg, M. (1965). *Society and the Adolescent Self-Image.* Princeton University Press.

Schimmack, U. (2010). What multi-method data tell us about construct validity. *European Journal of Personality*, *24*, 241–257. https://doi.org/10.1002/per.771

Schimmack, U. (2021). The Implicit Association Test: A method in search of a construct. *Perspectives on Psychological Science*, *16*(2), 396-414. https://doi.org/10.1177/1745691619863798

Zou, C., Schimmack, U., & Gere, J. (2013). The validity of well-being measures: A multiple-indicator–multiple-rater model. *Psychological Assessment*, *25*(4), 1247-1254. https://doi.org/10.1037/a0033902

2 Classical Test Theory

In Psychology, questionnaires or scales are a crucial part of the assessment. These scales provide importante information about a person. For instance, when doing an educational assessment, a test score on the exam can be a key indicator of the extent to which the person has mastered the knowledge of that domain.

In psychometric latent variable modeling, measurement occurs when a test score reflects a persons' ability in a specific area. However, this score is far from reflecting this ability in a perfect manner. It's easy to see that a person can vary its performance in an educational test if they answer in different times, Thus, the examination is not perfect, and has what we call **measurement error**. In addition, the ability is an unobserved variable (i.e., not observed directly, but inferred by a persons' score). To take all of this into account, a classical test theory (CTT) have been proposed to evaluate scale scores.

2.1 Overview of Classical Test Theory

In CTT, the main idea is that the score of a participant in a given assessment (denoted as X) can be decomposed into their true score (T) and a random error component (E):

$$X = T + E$$

T can be defined as the expected value of the observed score over an infinite number of repeat administrations in the same examination. Or, T can be thought as the score if the scale was perfectly measuring the ability of a given person (i.e., without the measurement error). X is the raw score of the participant, and E is the measurement error. Figure 2.1 shows the relationship between these elements.

The main task for CTT is to elaborate strategies to control or evaluate the magnitude of E. While E can be caused by a number of factors, such as problems with the test, bias from the participants, historical or environmental factors, etc. (Pasquali, 2017). However, there's no way to know the true score (T) of a participant if there were no measurement error (E). In fact, both T and E are unobserved variables. Thus, to use this model, we have to make our first assumption: one can define T as the expected value of the observed scores X (i.e., $E(X) = T$), which leads to the expected value of E being zero ($E(E) = 0$); or one can define the expected value of E as zero, which leads to T being the expected value of X. Both ways of

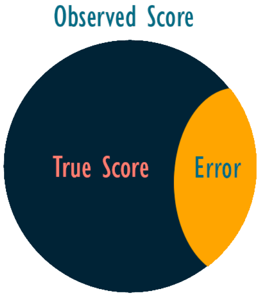


Figure 2.1: The CTT Approach

proceeding with the assumption lead to the same result, but they differ with respect to what is assumed, and what is a consequence of the assumptions (Brennan, 2011).

The structure of the CTT model equation ($X = T + E$) bears a striking resemblance to a straightforward linear regression equation, leading one to interpret E merely as model fitting error in the conventional statistical sense. However, such an interpretation is, at best, misleading. The CTT model operates as a tautology, wherein all variables on the right-hand side remain unobservable, and these unobservable variables lack inherent meaning beyond the assumptions we impose on them. Notably, T does not possess an independent status from the other variables in the model, rendering it inappropriate to characterize E as a residual or model fitting error (Brennan, 2011).

2.2 Reliability in CTT

The standard definition of reliability typically refers to the squared correlation between observed and true scores, denoted as $\rho^2(X, T)$. Additional expressions for reliability are provided below (Brennan, 2011):

$$\rho^2(X,T) = \rho(X,X') = \frac{\sigma^2(T)}{\sigma^2(X)} = \frac{\sigma^2(T)}{\sigma^2(T) + \sigma^2(E)}$$

The last three formulations are typically obtained by assuming that, for an indefinitely large population of participants: (1) test forms (denoted as X and X') are parallel, meaning they share identical observed score means, variances, and covariances, and they exhibit equal covariance with any other measure; (2) the covariance between errors for parallel forms is zero; and (3) the covariance between true and error scores is zero (Brennan, 2011). The reliability estimates tend to align more with the last two expressions in the equation provided earlier, both explicitly addressing true score variance, a value that remains elusive. Typically, these estimates leverage the understanding that the covariance between scores for classically parallel forms equals the true score variance, denoted as $\sigma(X, X') = \sigma^2(T)$. Coefficient α stands out as the most used among these coefficients.

2.3 References

Brennan, R. L. (2010). Generalizability theory and classical test theory. *Applied measurement in education*, *24*(1), 1-21. https://doi.org/10.1080/08957347.2011.532417

Pasquali, L. (2017). *Psicometria: teoria dos testes na psicologia e na educação*. Editora Vozes Limitada.

3 Theory of Latent Variables in Psychometrics

3.1 The Importance of Psychometrics

Psychometrics, a field within psychology, focuses on quantifying and measuring mental attributes, behaviors, performance, and emotions. Despite significant advancements in psychometric modeling over the past centuries, its integration into conventional psychological testing remains limited. This is very concerning, given that measurement problems abound in human research (Cronbach & Meehl, 1955; Messick, 1989; Borsboom et. al., 2004). Scholars argue that applying psychometric models to formalize psychological theory holds promise for addressing these challenges (Borsboom, 2006). However, many psychologists continue to rely on traditional psychometric methods, such as internal consistency coefficients and principal component analyses, without much deviation from past practices. Consequently, the interpretation of psychological test scores often lacks rigor, highlighting the disconnect between psychometrics and psychology (Borsboom, 2006).

3.2 Misunderstandings in Psychometric Practice

Misunderstandings are prevalent in the field of psychometrics. For instance, many studies delve into the structure of individual differences using latent variable theory but employ Principal Component Analysis (PCA) for data analysis. However, PCA does not align with latent variable theory. Thus, extracting a principal component structure alone does not shed light on its correspondence with a supposed latent variable structure. PCA serves as a data reduction technique (Bartholomew, 2004), which, in itself, isn't problematic as long as interpretations remain confined to principal components, which are essentially weighted sum scores.

Another example is the interpretation of group differences through observed scores. The interpretation of differences between groups regarding psychological attributes depends on measurement invariance (or measurement equivalence) between the groups being compared. There are several psychometric models and associated techniques to gain some control over this problem (Mellenbergh, 1989; Meredith, 1993; Millsap & Everson, 1993). However, almost no one cares about this, whether in Brazil or abroad, people simply evaluate the observed scores — without testing the invariance of the measurement models that relate these scores to psychological attributes. If you look at, for example, some of the most influential studies on group differences in intelligence, you rarely see invariance analyses. Consider for example the work of Herrnstein and Murray (1994) and Lynn and Vanhanen (2002). They infer differences in intelligence levels between groups from observed differences in IQ (by race and nationality) without even having performed a single test for invariance.

3.3 Obstacles to the Psychometric Revolution

Borsboom (2006) highlights a significant issue with psychometric models: they often challenge commonly accepted assumptions, such as measurement invariance. This leads researchers into fundamental questions about the structure of the phenomena they study and its relationship to observable data. Developing theories in this context is no simple task, potentially placing researchers in complex situations. Despite the importance of these inquiries in any scientific field, they aren't widely embraced in psychology. Consequently, even if researchers can provide compelling models for their observations, publishing such results proves challenging, as many journal editors and reviewers lack familiarity with psychometric models. Additionally, the perceived complexity of psychometrics exacerbates this issue. Compounding matters are the prevailing research standards in psychology, which demand that scientific articles remain accessible, despite the inherently intricate nature of the subject matter – human behavior and its underlying mental processes.

3.4 What are Constructs and Latent Variables

The notion of what a construct is and how to use it is fundamental for carrying out theories and research in psychology and related areas. Constructs are defined in empirical studies, and findings are interpreted in terms of the construct. What, then, would be a construct?

A construct is a concept that has three characteristics (Cronbach & Meehl, 1955): (i) it is not defined by a single observable referent (for example, 1 item on a scale); (ii) cannot be observed directly; and (iii) its observable referents are not fully inclusive. A latent variable is one (but not the only) statistical tool for studying constructs, and is commonly used in statistical analyzes to evaluate the relationship between constructs and their indicators (Spearman, 1904). A construct is operationally defined in terms of a number of items or indirect indicators, which are taken as an empirical analogue of a construct (Edwards & Bagozzi, 2000). These indicators are also called observed variables, which can be items in a self-report measure, interview, observations, or other means (DeVellis, 1991; Lord & Novick, 1968; Messick, 1995).

The reflective approach to latent variables is used by most methods in psychometrics that deal with measurement issues, often used in areas such as personality (John, 2021), well-being (Diener et al., 2010), criminology (Pechorro et al ., 2021), and others. The reflective approach is based on several assumptions, one of which is critical to its validity. It postulates a causal relationship between the latent variable and its indicators. This means that the variance and covariance in the indicators are dependent on changes in the latent variable (Bollen, 1989).

3.5 Ways to Represent a Construct

Usually, when running an Exploratory of Confirmatory Factor Analysis, we use the common factor model. This model sees the covariance between observable variables as a reflection of the influence of one or more factors and also a variance that is not explained. This would be different from network analysis, which allows covariance between items to have a cause between them. In other words, the psychometric model of factor analysis generally believes that item covariance occurs only because there is a latent factor that explains it. This is a very important assumption to keep in mind, as perhaps your construct does not fit the common factor model, but rather a network analysis. I will explain with an example given by Borsboom and Cramer (2013).

Below, we see the common factor model (Figure 3.1) of an instrument that measures major depression. In it, items measure aspects such as: feeling depressed, insomnia, weight gain, motor problems, fatigue, concentration problems, etc. We see from the image that the variation in scores on the items has a common cause, depression (that is, the higher the person's level of depression, the more they report having these symptoms).

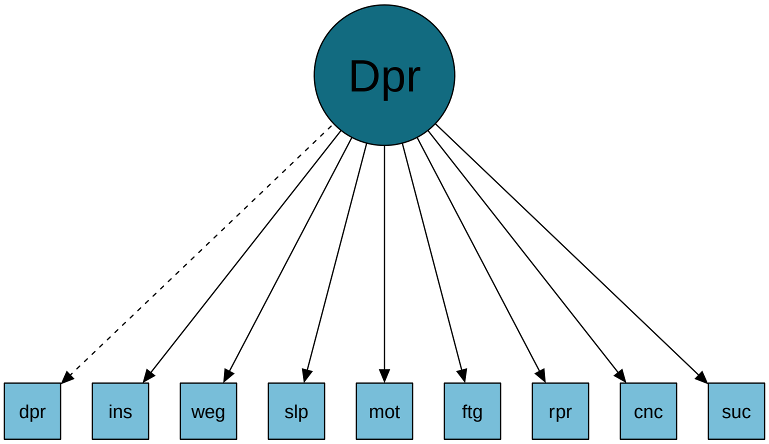


Figure 3.1: Factor Model

However, we can think that some items have relationships with each other that are not just due to depression. An example of this is the cause of concentration problems and its relationship with other symptoms. People who have problems sleeping become fatigued and, therefore, have problems concentrating (problems sleeping → fatigue → concentration problems). In other words, it is possible to infer a causal relationship between one observable variable and another, which breaks with the common factorial model assumption of local independence. A possible representation of this model is in the image below (Figure 3.2), where the items have causal relationships with each other.

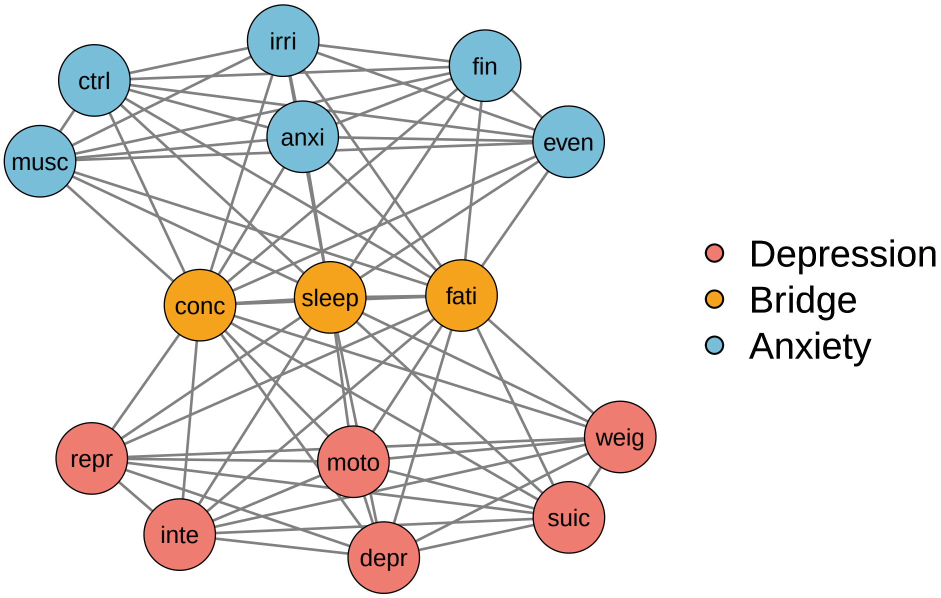


Figure 3.2: Network Model

So how do I know if my construct follows the common factor model or is more like network analysis? Well, often by theory! I know that researchers for a long time only cared about statistics to guide everything, but it is important for us to think about our constructs theoretically again and then test the theory empirically. However, there are also statistics that help verify this! But we'll see that in the next chapter.

3.6 The Platonic Relationship of Cause and Effect in Psychological Measures

The validity of psychometric models depends on the validity of the causal assumptions they make, which are generally implicit to the user. Psychological tests (e.g., self-report questionnaires) are typically constructed to measure constructs, while the responses observed in such tests are believed to reflect the latent variable underlying them (Van Bork et al., 2017). For example, a person's self-esteem is not observed directly, but we assume that it can be measured through items on an instrument. This line of thinking is the basis of the reflective approach, as represented by Figure 3.3. The reflective approach is applied to most psychometric models, such as classical test theory (Lord & Novick, 1968), the common factor model (Bartholomew, 1995; Speaman, 1904), item response theory models (Hambleton et al. al., 1991), latent class and latent profile analysis (B. O. Muthén & L. K. Muthén, 2000; Obersky, 2016), mixture models (Loken & Molenaar, 2008), latent growth models (Meredith & Tisak, 1990), reliability (Nunnally, 1978) and others, all crucial aspects of instrument development and evaluation.

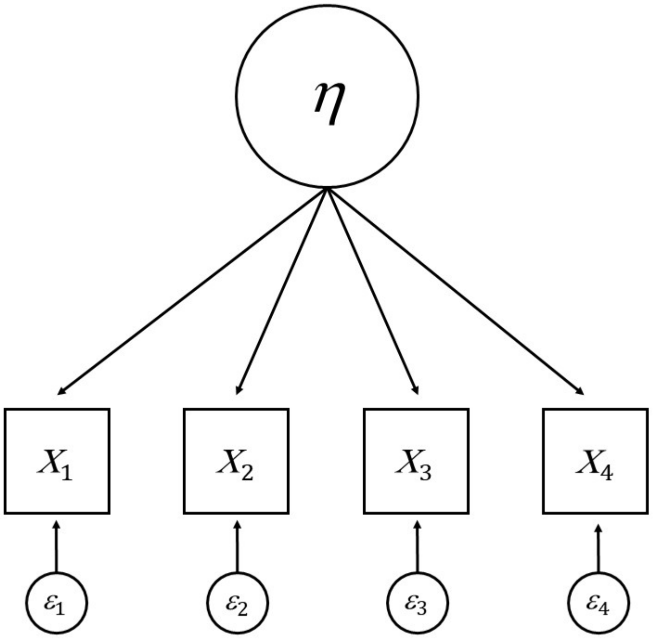


Figure 3.3: The Reflective Measurement Approach.

Causal language is common across a wide range of research areas (Pearl, 2009) and has permeated the definition of reflective measures in psychometric literature. For example, measurement error is often characterized as part of an observed variable that is not "explained" by the construct (or true score; Lord & Novick, 1968; Nunnally, 1978). Furthermore, other authors clearly state the direction of causality from the construct to its indicators (DeVellis, 1991; Long, 1983). However, some authors defend the descriptivist (or formative) approach, which understands latent variables as a parsimonious summary of the data and not the underlying cause of the indicators (Jonas & Markon, 2016; Van Bork et al., 2017). The difference between the causal and descriptive approaches is that the first can be seen as a representation of a real-world phenomenon, while the second does not include a conceptual interpretation and only describes statistical dependencies between indicators (Moneta & Russo, 2014; Van Bork and others, 2017). Causal interpretation is important in many settings (Van Bork et al., 2017): (1) in research, establishing causal relationships is often aligned with the primary goal of explaining correlations between multiple indicators, rather than just summarizing them. them.; (2) a causal interpretation of the construct legitimizes the reflective approach and its shared variance rather than other models that take into account the unique variance of indicators, such as the network model (Borsboom & Cramer, 2013); (3) the causal interpretation resonates with the assumption of local independence (i.e., covariation between indicators disappears when conditioned on their common cause).

However, the simple use of causal language does not necessarily imply that the variables actually have causal relationships, this is an empirical question. To incorporate causality, one must adhere to the principles of causality from the philosophy of science within the psychological, social, and behavioral literature (Asher, 1983; Bagozzi, 1980; Bollen, 1989; Cook & Campbell, 1979; Heise, 1975; James et al. al., 1982). To see ways to test the causal structure of your instrument, see the article by Franco et al., (2023).

3.7 References

Asher, H. B. (1983). *Causal modeling.* Sage.

Bagozzi, R. P. (1980). *Causal models in marketing.* Wiley.

Bartholomew, D. J. (1995). Spearman and the origin and development of factor analysis. *British Journal of Mathematical and Statistical Psychology*, *48*(2), 211-220. https://doi.org/10.1111/j.2044-8317.1995.tb01060.x

Bartholomew, D.J. (2004). *Measuring intelligence: Facts and fallacies.* Cambridge University Press.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, *71*(3), 425-440. https://doi.org/10.1007/s11336-006-1447-6

Borsboom, D., & Cramer, A. O. (2013). Network analysis: an integrative approach to the structure of psychopathology. *Annual review of clinical psychology*, *9*, 91-121. https://doi.org/10.1146/annurev-clinpsy-050212-185608

Borsboom, D., Mellenbergh, G.J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*, 1061–1071. https://doi.org/10.1037/0033-295X.111.4.1061

Cronbach, L.J., & Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281–302. https://doi.org/10.1037/h0040957

DeVellis, R. F., & Thorpe, C. T. (2021). *Scale development: Theory and applications.* Sage publications.

Diener, E., Wirtz, D., Tov, W., Kim-Prieto, C., Choi, D. W., Oishi, S., & Biswas-Diener, R. (2010). New well-being measures: Short scales to assess flourishing and positive and negative feelings. *Social indicators research*, *97*, 143-156. https://doi.org/10.1007/s11205-009-9493-y

Franco, V. R., Bastos, R. V., & Jiménez, M. (2023, June). *Tetrad Fit Index for Factor Analysis Models.* Paper presented at Virtual MathPsych/ICCM 2023. Via mathpsych.org/presentation/1297.

Hambleton, R. K., Swaminathan, H. and Rogers, H. J. (1991) *Fundamentals of Item Response Theory.* Sage

Harman, H. H. (1976). *Modem factor analysis* (3rd ed.). University of Chicago Press.

Heise, D. R. (1975). *Causal analysis.* Wiley.

Herrnstein, R.J., & Murray, C. (1994). *The Bell curve.* The Free Press.

James, L. R., Mulaik, S. A., & Brett, J. M. (1982). *Causal analysis: Assumptions, models and data.* Sage.

John, O. P. (2021). History, measurement, and conceptual elaboration of the Big-Five trait taxonomy: The paradigm matures. In O. P. John & R. W. Robins (Eds.), *Handbook of personality: Theory and research* (pp. 35–82). The Guilford Press.

Jonas, K. G., & Markon, K. E. (2016). A descriptivist approach to trait conceptualization and inference. *Psychological Review*, *123*(1), 90.

Kim, J. O., & Mueller, C. W. (1978). *Factor analysis.* Sage.

Loken & Molenaar (2008). Categories or Continua? The Correspondence Between Mixture Models and Factor Models. In G. R. Hancock & K. M. Samuelsen (Eds), *Advances in Latent Variable Mixture Models.* (pp. 277 - 298).

Long, J. S. (1983). *Confirmatory factor analysis: A preface to LISREL.* Sage.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Addison-Wesley.

Lynn, R., & Vanhanen, T. (2002). *IQ and the wealth of nations.* Praeger.

McDonald, R. P. (2013). *Test theory: A unified treatment.* Psychology Press.

Mellenbergh, G.J. (1989). Item bias and item response theory. *International Journal of Educational Research*, *13*, 127–143. https://doi.org/10.1016/0883-0355(89)90002-5

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika*, *58*, 525–543. https://doi.org/10.1007/BF02294825

Meredith, W., & Tisak, J. (1990). Latent curve analysis. *Psychometrika*, *55*, 107-122. https://doi.org/10.1007/BF02294746

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (pp. 13–103). American Council on Education and National Council on Measurement in Education.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741-749. https://doi.org/10.1037/0003-066X.50.9.741

Millsap, R.E., & Everson, H.T. (1993). Methodology review: Statistical approaches for assessing bias. *Applied Psychological Measurement*, *17*, 297–334. https://doi.org/10.1177/014662169301700401

Moneta, A., & Russo, F. (2014). Causal models and evidential pluralism in econometrics. *Journal of Economic Methodology*, *21*(1), 54-76. https://doi.org/10.1080/1350178X.2014.886473

Muthén, B. O., & Muthén, L. K. (2000). Integrating person-centered and variablecentered analyses: Growth mixture modeling with latent trajectory classes. *Alcoholism: Clinical & Experimental Research*, *24*(6), 882-891. https://doi.org/10.1111/j.1530-0277.2000.tb02070.x

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed). McGraw-Hill.

Oberski, D. (2016). Mixture models: Latent profile and latent class analysis. *Modern statistical methods for HCI*, 275-287. https://doi.org/10.1007/978-3-319-26633-6_12

Pechorro, P., DeLisi, M., Gonçalves, R. A., Quintas, J., & Hugo Palma, V. (2021). The Brief Self-Control Scale and its refined version among incarcerated and community youths: Psychometrics and measurement invariance. *Deviant Behavior*, *42*(3), 425-442. https://doi.org/10.1080/01639625.2019.1684942

Spearman, C. (1904). 'General intelligence' objectively determined and measured. *American Journal of Psychology*, *5*, 201-293. https://doi.org/10.1037/11491-006

Van Bork, R., Wijsen, L. D., & Rhemtulla, M. (2017). Toward a causal interpretation of the common factor model. *Disputatio*, *9*(47), 581-601. https://doi.org/10.1515/disp-2017-0019

4 Measurement Theory: Why it is Possible to Measure Psychological Phenomena

Sometimes, on days of perfect and exact light,

When things are as real as they can possibly be,

I slowly ask myself

Why I even bother to attribute Beauty to things.

Does a flower really have beauty?

Does a fruit really have beauty?

No: they have only color and form

And existence.

Beauty is the name of something that doesn't exist

But that I give to things in exchange for the pleasure they give me.

It means nothing.

So why do I say about things: they're beautiful?

Yes, even I, who live only off living,

Am unwittingly visited by the lies of men

Concerning things that simply exist.

Concerning things,

How hard to be just what we are and see nothing but the visible! (Fernando Pessoa)

Understanding the numerical representation of psychological constructs is essential for advancing the field of psychology. Classical measurement techniques provide a structured framework for assessing complex human behaviors and mental processes. By employing numerical representations, researchers can enhance the validity of their studies, leading to more accurate interpretations and meaningful insights into human cognition and behavior. In this blog post, I will tell you why we should care about measurement theory in psychology by telling you about: - Psychometrics objective; - History of measurement; - How to move forward and define a quantity in psychology; - The additive conjoint measurement framework; - If Rasch Modeling Entail Measurement; - Provide the tools to test measurement axioms.

4.1 The Objective of Psychometrics

Psychometrics is the branch of psychology that is concerned with quantifying and measuring mental attributes, behavior, performance, feelings, and the like. However, psychometrics does not shy away from criticism. As explained by Sijtsma (2012), on the one hand, we have Michell (2000, 2004, 2008) and Kyngdon (2008a, 2008b), who take the position that psychometrics is inadequate for measuring psychological attributes and should be replaced by the additive conjoint measurement (Luce & Tukey, 1964). Still, according to Sijtsma (2012), this perspective requires much of contemporary psychology: its serious implementation would bring psychological research to pause. However, what is the problem with stopping psychological research to improve further such research?

On the other hand, Borsboom and Mellenbergh (2004) and Borsboom and Zand Scholten (2008) argue that modern psychometrics, in particular, item response theory (IRT; Van der Linden & Hambleton, 1997), already successfully facilitated psychological measurement. This is the statistical perspective, which makes the mistake of confusing the prescriptive structure of a statistical measurement model with the theoretical structure of the attribute of interest.

4.2 Dr. Jekyll and Mr. Hyde: Measurement and Validity

When we are measuring an attribute (e.g., personality) from a class of objects (e.g., personality instruments), we associate numbers or other mathematical entities (e.g., Likert scales) within the objects so that the properties of the attribute are faithfully represented as numerical properties (Krantz et al., 1971). This is one of the many objectives of seeking the validity of psychological instruments (e.g., AERA, APA & NCME, 2014; Borsboom, 2005).

As exposed by Bringmann and Eronen (2015), most books, manuals, or monographs on measurement and measurement theory in psychology (for example, AERA, APA & NCME, 2014; Borsboom, 2005; Kline, 2000; McDonald, 1999), the physical measurement rarely appears. In the 2014 edition of Standards for Educational and Psychological Testing, validity is the first topic discussed, and is characterized as “the most fundamental consideration in test development and test evaluation” (AERA et al., 2014, p.11). According to the classical definition, validity refers to the extent to which the test or instrument measures what it is intended to measure (Kline, 2000, p. 17; McDonald, 1999, p. 197), but in the contemporary validity literature, there’s no consensus on how validity should be defined (see Newton & Shaw, 2013). Some of the most prominent approaches to validity are Messick’s (1989) unified treatment of validity, where the focus is on the appropriateness of the inferences that psychologists make based on test results. The approach is based on Kane’s arguments (2001, 2006, 2013), in which validation consists in providing evidence-based arguments for interpretations of test results.

However, all concepts of validity have relationships between the construct and its score, so they depend heavily on the quantification of psychological phenomena, which tries to be covered in psychometric research. I argue that the lack of empirical foundations for psychological measurement is “Mr. Hyde” of psychometrics, a monster for some researchers who, at the same time, is “Dr. Jekyll”, one of the many sources of validity of psychological instruments. This view aligns with that of Michell (2000), who states that “if science is a cognitive enterprise, then I argue that this way of doing it is not normal. It’s pathological”.

4.2.1 History of Psychological Measurement

The pathology of psychometrics began long before this area of knowledge existed. This all started with Pythagoras, who stated that “All things are made of numbers”. This is a very strong assumption, to say that the fundamental structure of all processes in nature is quantitative. In Timaeus, Plato continues saying the same thing, saying that all things are composed of the four basic elements (earth, fire, air, and water), which, in turn, are formed by polyhedra. He continues this logic, stating that polyhedra are made of triangles, which are reducible to lines and angles, and numbers.

But we have someone to disagree with this whole thing about reducing everything to quantities. Aristotle recognized that there are quantities (numbers, sizes, areas, etc.), but there are also qualities. These qualities are not quantitative, and concern things like colors and aromas. This distinction he said was something observable, where quantitative properties had an **additive structure**. Qualities did not have such a structure. Thus, he developed qualitative physics.

Of course, this fight continued between those who said that everything is quantitative and those who said that not everything is. Sometime later, Galileo himself joined the team of Pythagoras and Plato:

> [The universe] cannot be read until we have learnt the language and become familiar with the characters in which it is written. It is written in mathematical language, and the letters are triangles, circles, and other geometrical figures, without which means it is humanly impossible to comprehend a single word. (Galileu - II Saggiatore)

You can see the imprint of this speech to this day. People keep claiming that everything can be expressed in mathematical terms. Of course, if you number everything you associate it with mathematics. But what lies behind this thought is saying that everything in life is quantitatively measurable.

Great contemporaries from the same area as Galileo shared his views, such as Kepler and Descartes. However, this was not what made the quantitative area “win”, and rather the fact that Galileo’s physics dominated European science, taking the focus away from its rival, Aristotle’s physics.

With all this success, Galileo gave birth to the quantitative imperative. With the advent of Newton’s works, which strengthened Galileo’s quantitative views, we increasingly have more philosophers who defend this way of thinking. Kant wrote that

> … In any special doctrine of nature there can be only as much proper science as there is mathematics therein (Kant, 1786, p. 7).

You can see here the birthplace of the thought that all science must be quantitative. It was just in the 19th century that Lord Kelvin (an important physicist), expressed these thoughts more concretely

> When you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind: it may be the beginning of knowledge, but you have scarcely, in your thoughts, advanced to the stage of science, whatever the matter may be. (W. Thomson, 1891, p. 81)

Thus, this speech became the new favorite of the quantitative movement. See, Pearson (one of the darlings of psychology) used this line in 1978. And like many other areas that wanted to claim to be scientific because of quantitative thinking, psychology was no different. Back when psychology emerged in 1860, G. Fechner was also influenced by the thinking of the time about the nature of science. Fechner was a physicist who later became interested in psychological issues, such as the intensity of sensations. Although he was not the first person to attempt to measure psychological variables (perhaps it was Nicole Oresme in the 14th century), he proposed measurement methods.

Well, quantitative thinking continued in the progenitors of psychology (in this branch of psychology, in this case). Eugenicist Francis Galton, who influenced and did many studies on psychology, wrote that

> …until the phenomena of any branch of knowledge have been subjected to measurement and number, it cannot assume the status and dignity of a science (1879, p.147)

His assistant and one of the first psychology professors, James McKeen Cattell, followed this line of thought:

> Psychology cannot attain the certainty and exactness of the physical sciences, unless it rests on a foundation of experiment and measurement (1890, p. 373).

Even the creator of Factor Analysis, Charles Spearman, wrote that:

> … great as may be the potency of this [the experimental method], or of the preceding methods, there is yet another one so vital that, if lacking it, any study is thought by many authorities not to be scientific in the full sense of the word. This further and crucial method is that of measurement. (1937, p. 89).

E. B. Titchener (1905, pp. xxi-xxii) and Kulpe (1895, p. 11) thought the same, but stating that mental processes were measurable. Few who claimed that psychology was a science actually questioned the idea that it could be quantitative. This was probably because the status of science could be lost. A psychologist dared to do this: Franz Brentano. He adhered a little to the Aristotelian thoughts.

> Mathematics appears to me necessary for the exact treatment of all sciences only because we now in fact find magnitudes in every scientific field. If there were a field in which we encountered nothing of the sort, exact description would be possible even without mathematics. (Brentano, 1874, p. 65)

S. S. Stevens, "solved" the tension between people who considered and those who did not consider psychology as a science. Stevens saw that not only additivity could be represented numerically, and placed a new emphasis on these representations. It was Stevens who popularized the concepts of nominal, ordinal, interval, and ratio scales. In 1946 (p. 667), 1951 (p. 1), and 1959 (p. 19), Stevens defined measurement as "..the assignment of numbers to objects or events according to rules". This is one of the most famous measurement definitions in psychology today. You can find it as the main definition in thousands of psychology and psychometrics books.

However, his definition is empty of meaning. Not only are all things measurable, but also all things that can be numbered are forms of measurement. Furthermore, the distinction between quantitative and qualitative variables vanished. In other words, the variable being quantitative is no longer a characteristic of the variable itself, rather it is a pragmatic issue, decided by the researcher. Steven's definition confuses two distinct practices: a) measurement (in the classical sense); and b) numerical coding. Measurement involves the discovery of empirical facts of an intrinsic numeric type. b) Numerical coding is simply a cosmetic use in analysis and presentation of something that is not numeric. It is just a symbolic representation of facts.

4.3 How to Move Forward: Defining Quantity

After all this confusion, we still haven't defined what is a quantity. What distinguishes a quantitative variable from a non-quantitative one? To be quantitative, the variable must be ordered and have an additive structure. Let's do it in steps.

What would be a variable? In general, it is anything relative to which objects can vary. Size is a variable, as different objects have different sizes. Color is a variable, given that we have several colors. Being more detailed about this, the class of variables (size, color, etc.) can only be presented once for each object. Therefore, I do not have two heights at the same time. This is a condition crucial for a variable: not owning the same property more than once. Of course, we can have different properties on the same object, such as being a tall, white, brown-haired person. We have 3 variables (height, race/ethnicity, and hair color). But that's not all that characterizes a variable.

Relationships also form variables. The difference between properties and relationships is important. Things have uniquely shaped properties, like the size of my pen is one. Relationships involve a plurality of things. If the pen is on a table, then the situation involves both the pen and the table. Another example is speed. The speed of X relative to Y is something that involves X and Y. Of course, the speed of X relative to Y is just one. But we can also have another speed, that of X in relation to Z. This does not mean that X has more than one speed at the same time, it only has one, but there is also a relationship between the objects X, Y, and Z.

Another important concept is that of value: the properties and relationships that constitute a variable can be called values of that variable. For example, being 6 meters in size is the value of the size variable. Being a woman is the value of the gender variable, and so on. When we say that a quantitative variable is ordered and additive, we are saying that there are ordinal and additive relationships between the values of that variable.

What constitutes an ordered variable? Well, a simple way is to think that 6 meters is greater than 2 meters. We can also think about education, where higher education is more education than secondary education, which in turn is more than elementary education. More concretely, the values of the variables are ordered according to their magnitudes. We use the symbol $\geq$, which means "greater than or at least equal to", and $>$ meaning "greater than". The symbol $=$ means "equal to" or "identity of the value". Now let's go to the mathematics of the thing.

Consider that X, Y and Z are three values of a variable Q. Then, Q is ordinal if and only if:

1. if $X \geq Y$ and $Y \geq Z$, then $X \geq Z$ (this property is called transitivity. It means that if X is greater than or equal to Y, and Y is greater than or equal to Z, then X must be greater than or equal to Z, given the first relations mentioned).

2. if $X \geq Y$ and $Y \geq X$, then $X = Y$ (also called antisymmetry. It means that If X is greater than Y, and Y is greater than X, how can they not be greater than the other at the same time, so they have to be the same).
3. either $X \geq Y$ or $Y \geq X$ (called strong connectedness; only one variable can be larger, or both are the same).

The relation that has these 3 properties is called the simple order. Q is a ordinal variable if and only if $\geq$ is a simple order of its values. All quantitative variables are ordered by $\geq$, but not every ordinal variable is quantitative. To do this, it is necessary to have additivity.

Additivity is a ternary relationship (made up of 3 parts), symbolized as $X + Y = Z$.

Consider that Q is an ordinal variable, which for any values X, Y, and Z we have:

1. $X + (Y + Z) = (X + Y) + Z$ (associativity; i.e., the order of the sum does not affect the value resulting from the sum)
2. $X + Y = Y + X$ (commutativity; i.e., the order of the operands does not affect the final result).
3. $X \geq Y$ if and only if $X + Z \geq Y + Z$ (monotonicity; that is, if we add the same value on both sides, in X and Y, their order continues in the same direction, where X is greater than or equal to Y).
4. If $X \geq Y$ then there is a value Z that makes $X = Y + Z$ (solvability; means that if a value X is greater than the value Y, there is a third value Z which added to Y makes it a value equal to X).
5. $X + Y \geq X$ (positivity; if X is increased by a value Y, then this result has be greater than the original value of X, given that they are ordinal variables).
6. there is a natural number n such as $nX \geq Y$ (where $1X = X$ and $(n + 1)X = nX + X$ (Archimedean Condition; means that no value Y of the variable is infinitely greater than any other variable X).

These nine conditions (ordinal and additive) are uniformly coexisting. What it means is that they do nothing other than describe the structure of the variable. It does not describe the behavior of objects that have values of this variable.

Thus, it is not just additivity that a measure lives on. But an important criticism of Michell is how psychometricians do not evaluate their models correctly, at the level of measurement theory. As a result, they assume many things that may or may not be true, requiring testing or theorizing about these created measures. If psychometricians really evaluated the level of measurement of their variables, this would already solve the problem of psychometrics being pathological. We cannot keep assuming things that can be testable, or at least theorized in a more concrete way.

4.3.1 Additive Conjoint Measurement

The defense of the Additive Conjoint Measurement will be expressed here according to Michel (2014). Luce and Tukey (1964) proposed the additive conjoint measurement (ACM) specifically for quantification within the social sciences. This measure theory provides a way to identify quantitative structure other than through concatenation (or physical addition) operations. Instead, it allows quantitative structure to be detected through ordinal relationships on a variable. Although psychology lacks concatenation operations, it has many ordinal relationships.

The theory is about the type of situation in which a quantitative variable, , is a non-interactive function of two other variables, A and X. The word "non-interactive" can be understood as "additive" or "multiplicative", although, in fact, it is more general than that. This means that conjoint measurement theory refers to situations like $P = A+X$, or $P = A*X$. Its application is specifically to those instances where no P, A, or X are already quantified. This requires that:

(i) the variable P has an infinite number of values;

(ii) $P = f(A, X)$ (where f is some mathematical function);

(iii) there is a simple order over the values of P; and

(iv) the values of A and X can be identified (i.e. objects can be classified according to the value of A and X).

Let us call a system that satisfies (i)-(iv) a conjoint system. So if $\geq$ in P satisfies three special conditions, it follows that:

(v) P, A, and X are quantitative; and

(vi) f is a non-interactive function.

The three special conditions are:

(1) Double Cancellation;

(2) Solvability; and

(3) the Archimedean condition;

Suppose P is performance on some task (say, the time it takes to run a maze), A is motivation, and X is the amount of prior practice. Of course, it would be a simple matter to order the performances and classify subjects according to motivation (e.g., duration of food or water deprivation) and number of previous practice attempts.

Such conjoint systems are easily visually contemplated if they are thought of as composing a matrix where the rows are values of A, the columns, values of , and the cells, values of P. Let

a, b, c,... etc. be values of A, x, y, z,... etc. be values of X and, since $P = f(A, X)$, the pairs, ax, ay,... , cy, cz,... denote (possibly identical) values of P. Such a matrix is schematically represented by Figure 4.1 to help understand a visual representation of conditions (1) - (3).

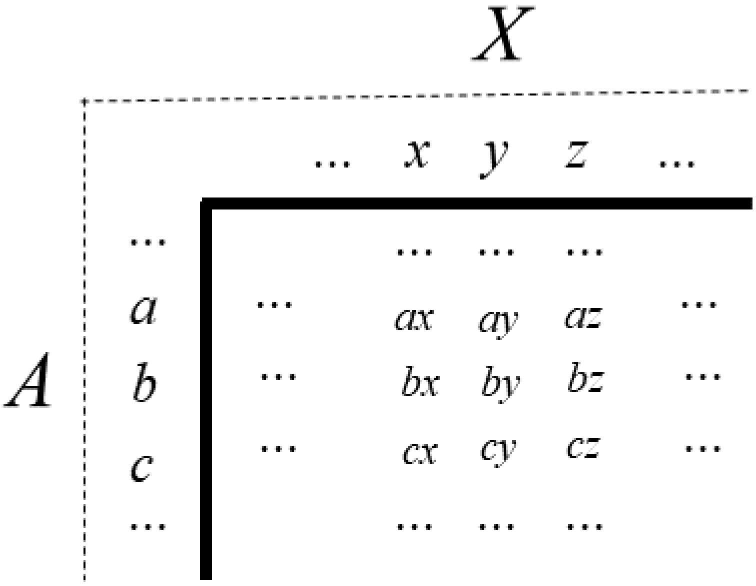


Figure 4.1: A schematic representation of a joint measurement matrix: ... a, b, c ... are values of the variable A, ... x, y, z . .. are values of the variable X and ... ax, ay, ... , cy, cz ... are values of the variable P (ax simply being that value of P produced by the conjunction of a and x, etc.).

4.3.1.1 Double Cancelation

The double cancellation condition states that if certain pairs of values of P are ordered by $\geq$, other pairs of specific values will also be ordered. It's like the transitivity condition that must satisfy (being a simple order). In the context of conjoint measurement, the transitivity of $\geq$ in P is a special case of double cancellation.

Double cancellation takes the following form. Let a, b, and c be any values of A and x, y, and z be any values of X, then $\geq$ in P satisfies double cancellation if and only if

$$\text{we have } ay \geq bx$$

$$\text{and also have } bz \geq cy$$

$$\text{thus, } az \geq cx.$$

Thus the condition appears obscure, but some light is shed if double cancellation is seen as a consequence of that special case of a non-interactive relation between P, A, and X,

$$P = A + X.$$

Given this relation,

$$ay \geq bx \text{ if and only if } a + y \geq b + x$$
$$\text{and } bz \geq ey \text{ if and only if } b + z \geq e + y.$$

Adding the two inequalities on the right-hand side we get

$$a + y + b + z \geq b + x + c + y$$

and since b and y is common on both sides of the inequality, they can be canceled, leaving

$$a + z \geq c + x$$

which, of course, is true if and only if

$$az \geq cx$$

.

Despite its simplicity, double cancellation is a condition that has considerable power. It strongly restricts the order in P. This can be illustrated in a $3X3$ matrix. Let $a1$, $a2$, and $a3$ be three values of A and $x1$, $x2$ and $x3$ be three values of X. The resulting conjoint matrix is illustrated in Figure 4.2.

Now, because a, b, and c in the double cancellation condition are any values of A, then $a1$, $a2$ and $a3$ can be substituted for them in any of the 3! (= 6) different possible ways. Similarly, $x1$, $x2$ and $x3$ can be replaced by x, y and z in 6 different ways. This produces 6 x 6 (= 36) different substitution instances of the double cancellation condition in the 3 x 3 matrix shown above (or in any 3 x 3 conjoint matrix). These 36 different replacement instances are shown in Figure 4.3.

They are not all logically independent of each other. In this, they are in six different sets, each with six. Within each set, the relevant order relations are between the same three values of P (or matrix cells). Arrows have been used to indicate these relationships (i.e., $ax \geq by$ is represented by ax -> by , the single-line arrows represent the antecedent orders and double-line arrows represent the consequent order.

			X	
		x_1	x_2	x_3
	a_1	a_1x_1	a_1x_2	a_1x_3
A	a_2	a_2x_1	a_2x_2	a_2x_3
	a_3	a_3x_1	a_3x_2	a_3x_3

Figure 4.2: Conjoint 3 x 3 Matrix

Within each set of six, if one of the double cancellation instances is true, they all will be. However, between sets, instances of double cancellation are logically independent of each other. Thus, within any 3 x 3 matrix there are six independent tests of the double cancellation condition, this condition is false if in any of the diagrams shown in the figure above, the antecedent order relations are valid, while the consequent is not; otherwise, they are satisfied. Obviously, satisfying double cancellation (in a conjoint matrix, even a 3 x 3 one) is not a trivial issue and very computationally demanding.

4.3.1.2 Solvability

The solvability condition requires that the variables A and X are complex enough to produce any required value of P. It is formally stated as the following.

The order $\geq$ in satisfies solvability if and only if (i) for any a and b in A and x in X, there is a value of X (call it y) such that $ax = by$ (i.e., both $ax \geq by$ and $by \geq ax$); and (ii) for any x and y in X and a in A, there is a value of A (call it b) such that $ax = by$. In other words, given any a, b, x, and y, y exists such that the equation

$$ax = by$$

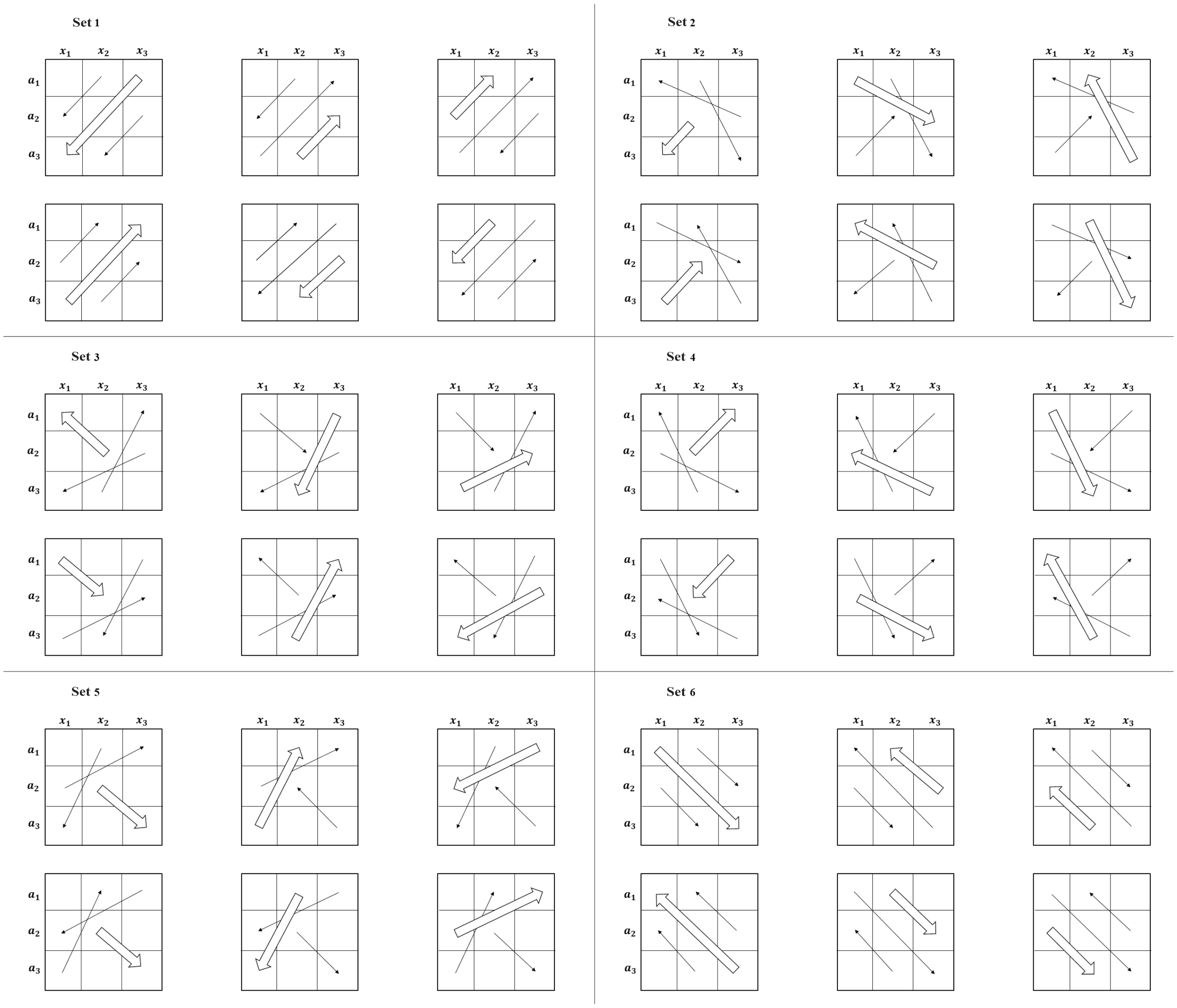


Figure 4.3: Double Cancelation

is solvable.

Thinking in terms of the relationship $P = A + X$, solvability implies that the values of A and X they are equally spaced (as natural numbers are) or they are dense (as rational numbers are).

4.3.1.3 Achimedean Condition

As already explained, the Archimedean condition guarantees that no value of a variable quantity is infinitely greater than any other value. Its meaning here is essentially the same, although in this context its expression is a little more complex. Thinking again in terms of $P = A + X$, a general idea of its content can be stated as follows. Conjoint measurement allows the quantification of differences between the values of A, between the values of X, and between the values of P. Limiting attention to A, the Archimedean condition means that no difference between any two values of A is infinitely greater than the difference between any other two values of A.

4.4 The Tale of Taxometric Analysis

The taxometric method started by Meehl (1995) is designed to assist researchers in determining whether the latent structure of a variable is categorical or continuous (Ruscio et al., 2007). The logic behind such analysis, regardless of the different way of performing it, relies on identifying if the latent distribution is unimodal or multimodal. If the former its' discovered, the researcher will conclude that the measurement level is continuous. In contrast, if the latter is found to be true, there will be evidence in favor of a categorical measurement level (Franco, 2021).

Ruscio and Kaczetow (2009) showed through extensive simulation studies that the curve-comparison fit indexes can identify the measurement level of a latent trait with 93% accuracy. However, in publications using such method, a meta-analysis showed a tendency of studies showing evidence in favor of a numerical and against a categorical latent variable (Haslam et al., 2012). Some possible limitations of taxometrics are due to its' lack of robustness both in statistical and measurement theory. For instance, its' hard to interpret taxometric analysis because the structure of observed covariance allows identical model fit with K taxons in comparison of models with K − 1 factors (Gibson, 1959), which makes taxometric analysis an unfalsifiable method. Moreover, this method has no further developments on current measurement theory, such as testing assumptions of ACM under the psychometric theory.

4.5 Does Rasch Modeling Entail Measurement?

In order to derive a numerical representation of psychological variables, a series of analyses have been developed. Still, conjoint measurement had very little impact on the construction of these psychometric models (Cliff, 1992; Narens & Luce, 1993; Ramsay, 1975, 1991; Schwager, 1991). One of these models is constantly related to conjoint measurement, the Rasch (1960) modeling.

To relate the Rasch model with conjoint measurement, some authors mistakenly argue the relationship via analogy with physical measurement (Kyngdon, 2008a). For instance, Fischer (1995) reached the conclusion that due to the logarithmic transformations yielding additive connections between derived measurements in physics, it logically follows that the constructs of individual ability and item complexity possess adequate complexity to support representation theorems extending to real numbers, which are essentially unique barring linear adjustments. In essence, individual ability and item difficulty are deemed to exhibit additive structures solely based on altering the relationship between these constructs. This relationship, however, is held by analogy to derived measurement through the notion of specific objectivity (Kyngdon, 2008a). To assert that an additive interval measurement of a person's ability and item difficulty is given by Rasch (1960), it's required that the underlying assumption is true: test performance is a multiplicative conjoint structure comprising of a person's ability and the item difficulty. Nonetheless, this has not been proven elsewhere.

Most research regarding the connection between Rasch modeling and conjoint measurement states that the Rasch model is a probabilistic version of conjoint measurement (Kyngdon, 2008b; e.g., Borsboom & Mellenbergh, 2004; Karabatsos, 2001; Kline, 1998). Thus, data that fits the Rasch model should allow for interval scaling. However, this is controversial. As stated by Michel (2008), the theory of conjoint measurement is a theory of ordinal and equivalence relations necessary for quantification, while the Rasch model is not concerned with such relations. Thus, the lack of articulation between Rasch and these relations does not entail that Rasch is mathematically equivalent to the theory of conjoint measurement (Kyngdon, 2008b). The difference between Rasch and the theory of conjoint measurement is mathematically clear. The theory of conjoint measurement proposes axioms for order and additivity (Luce & Tukey, 1964), while Rasch only proposes ordering of persons and items (Borsboom & Zand Scholten, 2008).

In standard research practice, when a person runs a Rasch model, they supposedly check for the consistency between the data and model with measurement axioms using fit statistics (Karabatsos, 2001). However, even if we assume that Rasch is testing those axioms, the test using fit statistics is not straightforward, since the specification of additive conjoint measurement under the Rasch model is data-dependent. This is because the Item Response Function is estimated directly from data, and data contains random or systematic noise. A consequence of such an effect is shown by Nickerson and McClelland (1984) and Karabastos (2001), where they show that Rasch modeling can empirically show perfect data fit, even for data sets containing violations of conjoint measurement axioms.

4.6 The Direct Test of Conjoint Measurement Axioms

A Bayesian technique originated in Karabatsos (2001) and further developed by Domingue (2013) makes it possible to test the ACM axioms. Suppose there's a unidimensional latent variable, which has a function linking the persons' responses to a set of items. Consider that P is an IxJ matrix that contains the true response probabilities for this set of items. Each cell in the conjoint matrix P^{MLE}, with dimensions IxJ, contains the percentage of respondents with a certain ability who answered the appropriate item correctly.

In order to determine whether the axioms are true for P, the order restrictions from the cancellation axioms are imposed stochastically via the Metropolis-Hastings (Metropolis et al., 1953; Hastings, 1970) jumping distribution. Domingue (2013) conducted simulation studies to test the new approach, and the evidence suggests that this approach can discriminate between data generated via the Rasch model and the 3PL model. This is expected, given that the 3PL item response model does not follow the axioms of additive conjoint measurement.

There is an R package called ConjointChecks (Domingue, 2013) that tests the assumptions regarding the cancellation of the additive conjoint measures. To download it, you need the updated devtools library, and install the most current version of the package.

```r
devtools::install_github("https://github.com/cran/ConjointChecks")
```

4.7 References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for Educational and Psychological Testing.*

Bobbs-Merrill. Kline, P. (2000). *Handbook of psychological testing.* Routledge.

Borsboom, D. (2005). *Measuring the Mind.* Cambridge University Press.

Borsboom, D., & Mellenbergh, G. J. (2004). Why psychometrics is not pathological: A comment on Michell.*Theory & Psychology*,*14*(1), 105-120. https://doi.org/10.1177/0959354304040200

Borsboom, D., & Zand Scholten, A. (2008). The Rasch model and conjoint measurement theory from the perspective of psychometrics. *Theory & Psychology*, *18*, 111-117. https://doi.org/10.1177/0959354307086925

Brentano, F. (1874). *Psychology from an empirical standpoint.* (English translation, 1973). Humanities.

Bringmann, L. F., & Eronen, M. I. (2016). Heating up the measurement debate: What psychologists can learn from the history of physics. *Theory & psychology*, *26*(1), 27-43. https://doi.org/10.1177/0959354315617253

Cattell, J. McK. (1890). Mental tests and measurements. *Mind*, *15*, 373-380.

Cliff, N. (1992). Abstract measurement theory and the revolution that never happened. *Psychological Science*, *3*, 186–190. https://doi.org/10.1111/j.1467-9280.1992.tb00024.x

Domingue, B. (2013). Evaluating the equal-interval hypothesis with test score scales. *Psychometrika*, *79*, 1-19. https://doi.org/10.1007/s11336-013-9342-4

Fechner, G. T. (1860). *Elemente der psychophysik*. Breitkopf & Hartel.

Franco, V. R. (2021). É possível identificar o nível de medida de variáveis latentes? [Is it possible to identify the measurement level of latent variables?]. *Avaliação Psicológica* [Psychological Assessment], *20*(2), a-d. http://dx.doi.org/10.15689/ap.2021.2002.ed

Galilei, G. (1864). *Il saggiatore*. G. Barbèra.

Galton, F. (1879). Psychometric experiments. *Brain*, *2*, 147-162

Gibson, W. A. (1959). Three multivariate models: Factor analysis, latent structure analysis, and latent profile analysis. *Psychometrika*, *24*(3), 229-252. https://doi.org/10.1007/BF02289845

Haslam, N., Holland, E., & Kuppens, P. (2012). Categories versus dimensions in personality and psychopathology: A quantitative review of taxometric research. *Psychological medicine*, *42*(5), 903-920. https://doi.org/10.1017/S0033291711001966

Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*(1), 97–109. https://doi.org/10.1093/biomet/57.1.97

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, *38*, 319-342. https://doi.org/10.1111/j.1745-3984.2001.tb01130.x

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.; pp. 17–64). American Council on Education/Praeger.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*, 1-73.

Kant, I. (1786). *Metaphysicalfoundations o f natural science.* (J. Ellington Trans. 1970). https://doi.org/10.1111/jedm.12000

Karabatsos, G. (2001). The Rasch model, additive conjoint measurement, and new models of probabilistic measurement theory. Journal of applied measurement, 2(4), 389-423.

Krantz, D. H., Luce, R. D., Suppers, P., & Tversky, A. (1971). *Foundations of measurement* (Vol 1: Additive and Polynomial Representations). Academic Press.

Kulpe, O. (1895). *Outline of psychology.* Sonnenschein.

Kyngdon, A. (2008a). The Rasch model from the perspective of the representational theory of measurement. *Theory & Psychology*, *18*, 89–109. https://doi.org/10.1177/0959354307086924

Kyngdon, A. (2008b). Conjoint measurement, error and the Rasch model: A reply to Michell, and Borsboom and Zand Scholten. *Theory & Psychology*, *18*, 125–131 https://doi.org/10.1177/0959354307086927

Luce, R. D., & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology*, *1*, 1-27. https://doi.org/10.1016/0022-2496(64)90015-X

McDonald, R. P. (1999). *Test theory: A unified treatment.* Erlbaum.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, *21*(6), 1087–1092. https://doi.org/10.1063/1.1699114

Meehl, P. E. (1995). Bootstraps taxometrics: Solving the classification problem in psychopathology. *American Psychologist*, *50*(4), 266–275. https://doi.org/10.1037/0003-066X.50.4.266

Michell, J. (2000). Normal science, pathological science and psychometrics.*Theory & Psychology*, *10*(5), 639-667. https://doi.org/10.1177/0959354300105004

Michell, J. (2004). Item response models, pathological science and the shape of error: Reply to Borsboom and Mellenbergh.*Theory & Psychology*, *14*(1), 121-129. https://doi.org/10.1177/0959354304040201

Michell, J. (2008). Is psychometrics pathological science?. *Measurement*, *6* (1-2), 7-24. https://doi.org/10.1080/15366360802035489

Michell, J. (2014). *An introduction to the logic of psychological measurement.* Psychology Press.

Narens, L., & Luce, R.D. (1993). Further comments on the 'nonrevolution' arising from axiomatic measurement theory. *Psychological Science*, *4*, 127–130. https://doi.org/10.1111/j.1467-9280.1993.tb00475.x

Newton, P. E., & Shaw, S. D. (2013). *Validity in Educational & Psychological Assessment.* Sage.

Pearson, K. (1978). *The history of statistics in the seventeenth and eighteenth centuries.* Griffin

Ramsay, J.O. (1975). Review of Foundations of measurement, Volume I. *Psychometrika*, *40*, 257–262.

Ramsay, J.O. (1991). Reviews of Foundations of measurement, Volumes II and III. *Psychometrika*, *56*, 355–358.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Danish Institute for Educational Research.

Ruscio, J., & Kaczetow, W. (2009). Differentiating categories and dimensions: Evaluating the robustness of taxometric analyses. *Multivariate Behavioral Research*, *44*(2), 259-280. https://doi.org/10.1080/00273170902794248

Ruscio, J., Ruscio, A. M., & Meron, M. (2007). Applying the bootstrap to taxometric analysis: Generating empirical sampling distributions to help interpret results. *Multivariate Behavioral Research*, *42*(2), 349-386. https://doi.org/10.1080/00273170701360795

Schwager, K.W. (1991). The representational theory of measurement: An assessment. *Psychological Bulletin*, *110*, 618–626. https://doi.org/10.1037/0033-2909.110.3.618

Spearman, C. (1937). *Psychology down the ages* (Vol. 1). Macmillan.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*, 667-680. https://doi.org/10.1126/science.103.2684.677

Stevens, S. S. (1951). Mathematics, measurement and psychophysics. In S. S. Stevens (Ed.), *Handbook o f experimental psychology* (pp. 1-49). Wiley.

Stevens, S. S. (1959). Measurement, psychophysics and utility. In C. W. Churchman, & P. Ratoosh (Eds.), *Measurement: definition and theories* (pp. 18-63). Wiley.

Thomson, W. (1891). Popular lectures and addresses (Vol. 1). Macmillan. Titchener, E. B. (1905). *Experimental psychology* (Vols. 1-3).

Macmillan. Van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. Springer-Verlag.

5 Exploratory Factor Analysis

Exploratory Factor Analysis (EFA) is a statistical tool that serves several purposes. In social sciences (e.g., Psychology, Education) it has served the general purpose of reducing the number of dimensions/factors of a scale or instrument. That is, reducing the number of parameters to the number of latent traits/psychological constructs. It serves the purpose of seeking evidence of validity of internal structure of an instrument.

Thus, we can define the objective of EFA as follows: Evaluate the dimensionality of a series of indicators in order to identify the smallest number of latent traits that explain the pattern of correlations (Osborne, 2014).

More formally, the Common Factor Model sees the covariance between observable variables as a reflection of the influence of one or more factors and unexplained variance. The items are considered indicators that vary according to the level of the latent trait, that is, the higher your level of Depression, the greater your agreement with the item "I have been feeling depressed".

What would be the point of carrying out an EFA? To reduce the number of parameters we have and group it into one or more latent traits. In other words, instead of having 21 different indicators that assess Depression/Anxiety/Stress, we reduce it to 3 indicators (latent traits) that explain the variance of the items. EFA divides between common variance and unique variance. Common variance concerns the shared influence of latent traits on an indicator. Unique variance can represent two things: item variation that reflects unknown latent causes; and random error given unreliability or measurement error.

The common factor model is based on the mechanics of linear regression, and specifies that the observable data reflect a linear combination of latent trait influence. If we have 1 indicator/item, representing m factors, we have the following notation:

Item1 $= \boldsymbol{\lambda}_{i1}\eta_1 + \boldsymbol{\lambda}_{i2}\eta_2 + ... + \boldsymbol{\lambda}_{im}\eta_m + \boldsymbol{\varepsilon}_i$

Where: $\boldsymbol{\lambda}_{im}$ = the strength of the association between the factor m and the indicator i;$\boldsymbol{\varepsilon}_i$ = the error in the indicator i; η = the factor of number m.

If we have 5 indicators/items represented by 3 factors we have the following notation:

Item1 $= \boldsymbol{\lambda}_{11}\eta_1 + \boldsymbol{\lambda}_{12}\eta_2 + \boldsymbol{\lambda}_{13}\eta_3 + \boldsymbol{\varepsilon}_1$

Item2 $= \boldsymbol{\lambda}_{21}\eta_1 + \boldsymbol{\lambda}_{22}\eta_2 + \boldsymbol{\lambda}_{23}\eta_3 + \boldsymbol{\varepsilon}_2$

Item3 $= \boldsymbol{\lambda}_{31}\eta_1 + \boldsymbol{\lambda}_{32}\eta_2 + \boldsymbol{\lambda}_{33}\eta_3 + \boldsymbol{\varepsilon}_3$

Item4 $= \boldsymbol{\lambda}_{41}\eta_1 + \boldsymbol{\lambda}_{42}\eta_2 + \boldsymbol{\lambda}_{43}\eta_3 + \varepsilon_4$

Item5 $= \boldsymbol{\lambda}_{51}\eta_1 + \boldsymbol{\lambda}_{52}\eta_2 + \boldsymbol{\lambda}_{53}\eta_3 + \varepsilon_5$

5.1 Why is it Called Exploratory?

It is generally more used as data-driven, that is, it does not presuppose the behavior of the relationship between the variables and their factors. In EFA, the number of factors that appear in the data is generally tested, and the items have factor loadings on both their hypothesized factor and the other factors. On the other hand, in a Confirmatory Factor Analysis, the parameters are fixed and the items load (generally) only on their respective factors. However, they are not atheoretical, since you need to have a theory to build a scale. For instance, I might have a theory about behavioral intent that encompasses a 3-factor model. Thus, I'd build items that measures those 3 factors.

5.2 EFA Step-by-step

EFA has, mainly, 4 steps.

a) Verification of data adequacy

b) Factor Retention

c) Factor Extraction

d) Factor Rotation

I'll explain each one of them below.

5.2.1 Data adequacy

To test whether the data we have is suitable for doing an EFA, we generally use two criteria: Bartlett test of sphericity and Kaiser-Meyer-Olkin (KMO).

5.2.1.1 Bartlett's test of sphericity

This test verifies the hypothesis that the variables are not correlated in the population. Thus, its hypothesis says that the population correlation matrix is an identity matrix. If the correlation matrix is an identity matrix, the factor model is inappropriate, given that there is no correlation between the variables. See an example of an identity matrix below, imagine that this matrix below is the correlation matrix between items of an instrument.

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

The statistical part behind the test is given by the following equation:

$\chi^2 = -[(n-1) - \frac{(2v+5)}{6}] * log(det(R))$

Where:

n = sample size;

v = number of variables;

det(R) = determinant of the correlation matrix;

Values from Bartlett's test of sphericity with significance levels of $p < 0.05$ indicate that we can proceed with an EFA (Tabachnick & Fidell, 2007).

5.2.1.2 Kaiser-Meyer-Olkin (KMO)

Evaluates the adequacy of the factor analysis, indicating the proportion of variance in the items that may be caused by factors. KMO checks whether the inverse correlation matrix is close to the diagonal matrix by comparing the values of the observed linear correlations with the values of the partial correlations. The formula of KMO is

$KMO = \frac{\sum\sum_{j \neq k} r_{jk}^2}{\sum\sum_{j \neq k} r_{jk}^2 + \sum\sum_{j \neq k} q_{jk}^2}$

Where:

r_{jk}^2 = is the square of the elements of the original off-diagonal correlation matrix;

q_{jk}^2 = is the square of the partial correlation between the variables;

The KMO index values indicating the appropriateness of factor analysis can vary among different authors. For instance, Hair et al. (2006) suggest that KMO values between 0.5 and 1.0 are acceptable, with values below 0.5 indicating that factor analysis may not be suitable for the dataset. On the other hand, Kaiser & Rice (1974) propose a more stringent criterion, indicating that for the factor analysis model to have adequate fit, the KMO value should exceed 0.7.

5.2.2 Factor Retention

Given that computer software will extract as many factors as there are items in the analyses, and for the purpose of EFA, we have to decide how many extracted factors we should retain for subsequent analyses. We have some methods to decide:

5.2.2.1 Kaiser Criterion (1960, 1970)

This criterion proposes that eigenvalues greater than 1 are a good parameter for the factor to be significant. This rule reflects the intuition that the factor must take into account the variance of at least one indicator. Thus, the eigenvalue is the sum of the squared factor loadings of the items, which represents the variance in each item that can be explained by the factor. The Kaiser criterion should not be used in isolation, because it both underestimates the number of factors and also overestimates them in some cases (Zwick & Velicer, 1986).

5.2.2.2 Scree Plot

It involves analyzing the eigenvalue graph and evaluating the “elbow break” in the data where the slope of the curve changes (flattens) sharply. An example is given in Figure 5.1, where this is clear is in the following data (this data was created randomly to illustrate). In Panel A, we see the elbow breaking when we have 5 factors, that is, we can say that this scale has 4 factors that explain the variance of the data. However, sometimes it won’t be so clear where the “elbow break” is (as we can see in Panel B).

Thus, identifying this “elbow break” can become an interpretative exercise, and is not recommended for determining the number of factors to extract if used alone.

5.2.2.3 Parallel Analysis

It was a method proposed by Horn (1965), which uses Monte-Carlo simulation and which involves generating random and uncorrelated data to compare the eigenvalues of the EFA with the eigenvalues of the random data. In this simulation, a hypothetical set of variable correlation matrices is created with the same dimensionality as your data. This simulated data is then factored as many times as the researcher wants and the average of the eigenvalues of this simulation is calculated. Therefore, the number of factors to be retained must be those that explain more than random data.

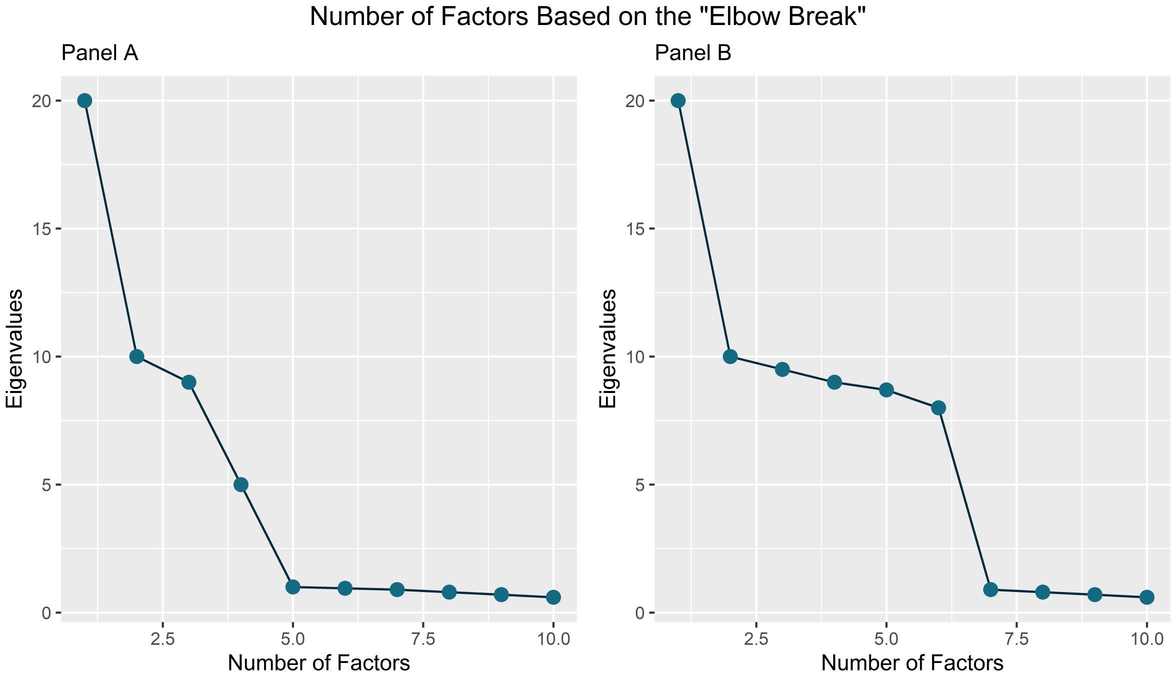


Figure 5.1: Number of Factors Based on the “Elbow Break”.

5.2.2.4 Theory

Always remember, factor retention criteria, even with more “objective” measures, such as those mentioned above, have a subjective criteria. It is always important to worry about a crucial point for the reproducibility of psychology, the THEORY behind it. If we don’t have a good theory behind it, our conclusions may fall apart both in our research and in future replications. Therefore, some researchers argue that theory can be a criterion for selecting the number of factors in a scale. Of course, if we have a solid theory, we are likely to see this reflected in the other indicators.

5.2.3 Factor Extraction

An extraction technique is a group of methods that examine the correlation/covariance between all variables and seek to extract a latent variable from the measured variables. For a long time, in the literature, some authors used Principal Component Analysis to perform dimension reduction of latent traits. Thus, we need to differentiate Exploratory Factor Analysis from Principal Component Analysis (PCA).

5.2.3.1 Exploratory Factor Analysis vs Principal Component Analysis

Both techniques have the same objective: to reduce a given number of items to a smaller number of variables. Both methods assume that the variance of an item is composed of specific variance, common variance and error variance, as explained previously.

PCA is based on the linear correlation of observed variables, without differentiating common variance from specific variance between items. In other words, when items are retained in a given component, both common variance and specific variance are taken into account. While in EFA only the common variance is taken into account.

In Figure 5.2, we see the difference between PCA and EFA. In general, PCA is based on the formative model, that is, the latent variables are formed by the manifest variables (or items). An example of a variable in the formative model is socioeconomic level, which can be explained by items such as income, place of residence, education, etc. Thus, this latent variable is a representation of the items. EFA, on the other hand, is based on the reflective model, that is, we have a latent trait that explains the variation of the variables. It is no longer changing the items that change the level of the latent variable, but the opposite. Thus, our items are a representation of the latent trait. An example of a variable in the reflective model is subjective well-being, where the greater the person's subjective well-being, the more they will tend to agree with the item "I am satisfied with my life".

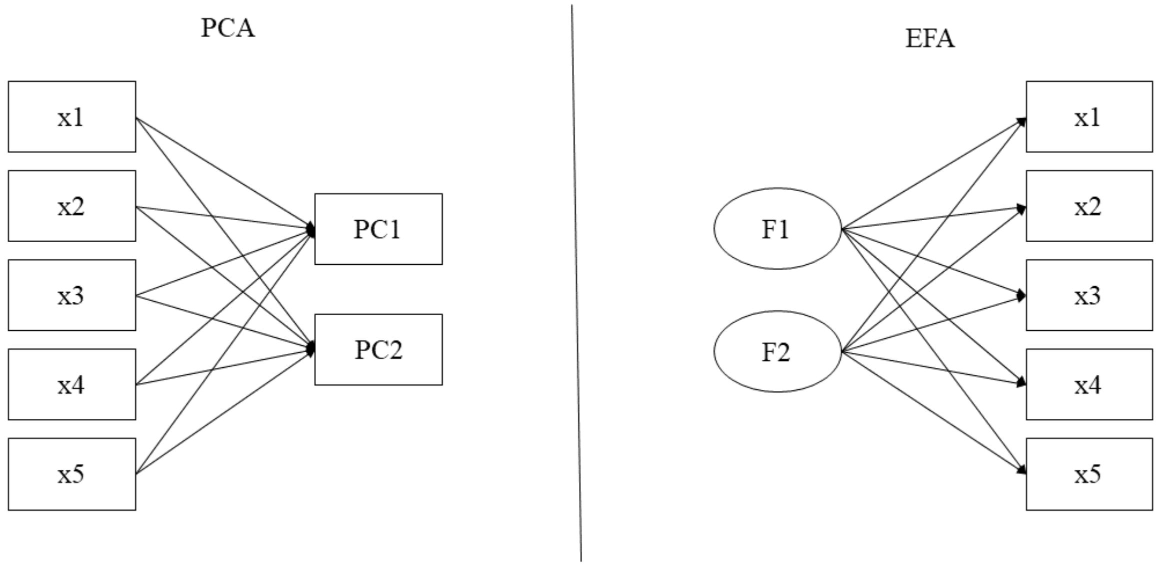


Figure 5.2: EFA vs PCA

In EFA, there are a number of extraction methods to choose from: unweighted least squares, generalized least squares, maximum likelihood, principal axis factoring, alpha factoring, image factoring, etc. Fabrigar et al (1999) argue that, if the data are relatively normally distributed, maximum likelihood is the best choice to make, as it allows the calculation of a variety of model fit indices and allows the statistical significance test of factor loadings, correlations between factors and the calculation of confidence intervals. However, if the assumption of multivariate normality is violated, one of the principal factor's methods (e.g., principal axis factoring) is recommended.

5.2.4 Factor Rotation

After selecting the number of factors, rotation is done to facilitate data interpretation. The term rotation is used because the axes are being rotated so that the clusters of items fall as close to themselves as possible. In other words, the group of items that are close together become even closer. Although this method changes the eigenvalue, the overall percentage of variance will remain the same.

We have two ways to rotate: orthogonally and obliquely. When we rotate orthogonally, the axes move while remaining orthogonal to each other (that is, they continue to have an angle of 90° between them). We generally perform an orthogonal rotation when it is assumed that the factors are not related to each other. In oblique rotation, as the name suggests, the axes move without necessarily maintaining a 90° angle between them. We generally perform an oblique rotation when we do not have the assumption of orthogonality between factors, that is, the factors can be related.

5.3 How to Run an Exploratory Factor Analysis in R

To run an Exploratory Factor Analysis, we must first install the *psych* (Revelle, 2023) package and *EFA.MRFA* (Navarro-Gonzalez & Lorenzo-Seva, 2021).

```r
install.packages("psych")

install.packages("EFA.MRFA")
```

So, we tell the program that we are going to use the functions of these packages.

To run the analyses, we will use the BFI database (Big Five Personality Factors Questionnaire) that already exists in the *psych* package.

5.3.1 Data Adequacy in R

To see how suitable the data is for factorization, we will perform **Bartlett's test of sphericity**.

First, we will calculate the correlation matrix of the 25 BFI items, omitting missing values (i.e., *NA*), with:

```r
correlation <- cor(na.omit(psych::bfi[,1:25]))
```

Then we calculate the sphericity test where the first argument we put the correlation matrix and the second we put the sample size. We will have the following code and output.

```r
psych::cortest.bartlett(correlation, n = nrow(na.omit(psych::bfi[,1:25])))
```

```
$chisq
[1] 18146.07

$p.value
[1] 0

$df
[1] 300
```

The level of significance was small enough for R to say that it is 0. Assuming that values lower than 0.05 indicate that a factor analysis can be useful for our data, our data proved to be suitable for this indicator. Remember, in this test we are looking at the difference between an identity matrix and our correlation matrix, so if it is significant, we have a statistically significant difference between the two matrices.

Now we will do another data adequacy test, this time using the **Kaiser-Meyer-Olkin Measure of Sampling Adequacy**, or KMO for those more familiar with it. Use the following code, where the argument is your items.

```r
psych::KMO(psych::bfi[,1:25])
```

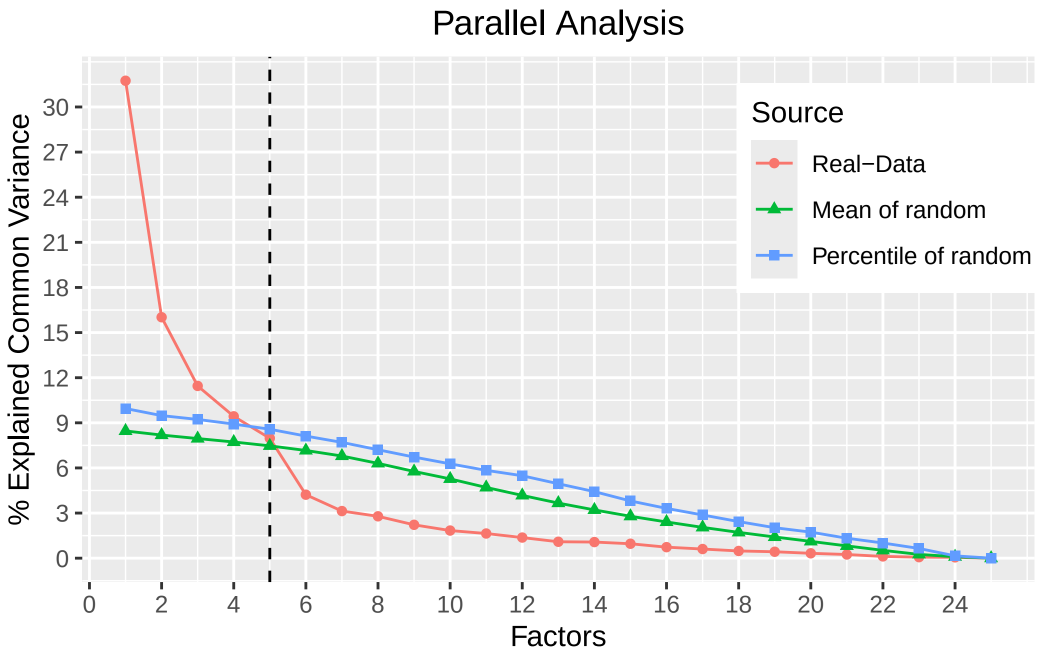
```
Kaiser-Meyer-Olkin factor adequacy
Call: psych::KMO(r = psych::bfi[, 1:25])
Overall MSA =  0.85
MSA for each item =
  A1   A2   A3   A4   A5   C1   C2   C3   C4   C5   E1   E2   E3   E4   E5   N1
0.74 0.84 0.87 0.87 0.90 0.83 0.79 0.85 0.82 0.86 0.83 0.88 0.89 0.87 0.89 0.78
  N2   N3   N4   N5   O1   O2   O3   O4   O5
0.78 0.86 0.88 0.86 0.85 0.78 0.84 0.76 0.76
```

We see that the overall KMO value was 0.85, and we also have the value for each item. Values close to 1.0 generally indicate that factor analysis can be useful for our data. Therefore, we will proceed with the Exploratory Factor Analysis.

5.3.2 Parallel Analysis in R

As one of the objectives of EFA is to reduce the number of parameters to the number of psychological constructs, it is important to carry out analyzes to select their number of factors. To select the number of factors we will use the parallel analysis of the *EFA.MRFA* package. This is the same parallel analysis as the FACTOR software. It does a *Parallel Analysis using Minimum Rank Factor Analysis*, using the following function.

```
resultsPA <- EFA.MRFA::parallelMRFA(na.omit(psych::bfi[,1:25]),
                                   Ndatsets = 500, percent = 95,
                                   corr = "Polychoric", graph = TRUE,
                                   display = FALSE
                                   )
```



where

1. The first argument is our data, that is, the items;
2. `Ndatsets` = number of datasets simulated for parallel analysis;

3. `percent` = confidence interval;
4. `corr`= type of correlation (if it is polychoric, spearman, kendall, etc.). The polychoric correlation matrix deals best with ordinal data that comes from a latent variable;
5. `graph`= to output the image of the eigenvalues.

Initially, we also see in the output (not available here, but available if you run the same code in your computer) the value of Bartlett's test of sphericity and the KMO, so we can just use the above function to calculate both and the parallel analysis. We see in the figure that the number of factors to be extracted by the average percentage of variance of the parallel analysis was 5, equal to the BFI theory. It also shows in the output the number of recommended factors:

```
cat("The number of factors based on the average of the simulations was ",
 resultsPA$N_factors_mean,
 ".\nThe number of factors based on the percentile was",
 resultsPA$N_factors_percentiles,".")
```

```
The number of factors based on the average of the simulations was  5 .
The number of factors based on the percentile was 4 .
```

5.3.3 Factor Extraction in R

Now let's do exploratory factor analysis with the right number of factors (i.e., five).

```
fit <- psych::fa(na.omit(psych::bfi[,1:25]),
                 nfactors = 5,
                 n.obs = nrow(na.omit(bfi[,1:25])),
                 rotate = "oblimin",
                 cor = "poly",
                 fm = "minrank")
```

```
Loading required namespace: Rcsdp

Loading required namespace: GPArotation
```

where

1. The first argument is our data, that is, the items.
2. `nfactors` = number of factors that emerged in the parallel analysis
3. `n.obs` = number of participants.
4. `rotate`= type of rotation. Here I chose an oblique rotation, but the package has several ways to do both oblique and orthogonal rotations.

5. `cor`= type of correlation (if it is polychoric, spearman, kendall, etc.)
6. `fm` = the method of doing factor analysis. "minrak" does the *Minimum Rank Factor Analysis*.

Let's ask for the result:

```r
print(fit, sort=TRUE)
```

```
Factor Analysis using method =  minrank
Call: psych::fa(r = na.omit(psych::bfi[, 1:25]), nfactors = 5, n.obs = nrow(na.omit(bfi[,
    1:25])), rotate = "oblimin", fm = "minrank", cor = "poly")
Standardized loadings (pattern matrix) based upon correlation matrix
   item MRFA2 MRFA1 MRFA3 MRFA5 MRFA4   h2   u2 com
N2   17  0.86  0.06  0.02 -0.11  0.02 0.73 0.27 1.0
N1   16  0.85  0.11  0.00 -0.12 -0.06 0.72 0.28 1.1
N3   18  0.76 -0.08 -0.04  0.07  0.01 0.62 0.38 1.0
N5   20  0.55 -0.22 -0.01  0.24 -0.19 0.45 0.55 2.0
N4   19  0.52 -0.40 -0.14  0.11  0.10 0.56 0.44 2.2
E2   12  0.12 -0.71 -0.04 -0.05 -0.06 0.60 0.40 1.1
E4   14 -0.01  0.70  0.02  0.26 -0.09 0.66 0.34 1.3
E1   11 -0.07 -0.67  0.13 -0.07 -0.08 0.46 0.54 1.1
E3   13  0.09  0.48  0.01  0.23  0.32 0.52 0.48 2.3
E5   15  0.17  0.48  0.30  0.02  0.23 0.49 0.51 2.5
C2    7  0.18 -0.09  0.76  0.07  0.06 0.58 0.42 1.2
C4    9  0.18  0.01 -0.70  0.02 -0.04 0.57 0.43 1.1
C5   10  0.20 -0.14 -0.63  0.03  0.12 0.53 0.47 1.4
C3    8  0.03 -0.07  0.62  0.10 -0.08 0.38 0.62 1.1
C1    6  0.06 -0.04  0.60  0.00  0.18 0.41 0.59 1.2
A2    2 -0.02  0.02  0.07  0.77  0.02 0.64 0.36 1.0
A3    3 -0.02  0.18  0.03  0.68  0.05 0.60 0.40 1.2
A5    5 -0.12  0.29  0.01  0.55  0.05 0.54 0.46 1.7
A1    1  0.21  0.18  0.08 -0.53 -0.07 0.29 0.71 1.7
A4    4 -0.04  0.12  0.23  0.48 -0.20 0.39 0.61 2.0
O3   23  0.03  0.21  0.02  0.07  0.66 0.55 0.45 1.2
O5   25  0.12  0.10 -0.05  0.04 -0.61 0.39 0.61 1.2
O1   21  0.00  0.13  0.08  0.02  0.57 0.40 0.60 1.1
O2   22  0.20  0.05 -0.09  0.16 -0.53 0.33 0.67 1.6
O4   24  0.15 -0.35 -0.05  0.22  0.47 0.38 0.62 2.6

                      MRFA2 MRFA1 MRFA3 MRFA5 MRFA4
SS loadings            3.00  2.78  2.56  2.40  2.05
Proportion Var         0.12  0.11  0.10  0.10  0.08
```

```
Cumulative Var         0.12  0.23  0.33  0.43  0.51
Proportion Explained   0.24  0.22  0.20  0.19  0.16
Cumulative Proportion  0.24  0.45  0.65  0.84  1.00

 With factor correlations of
      MRFA2 MRFA1 MRFA3 MRFA5 MRFA4
MRFA2  1.00 -0.20 -0.18 -0.05  0.01
MRFA1 -0.20  1.00  0.24  0.29  0.13
MRFA3 -0.18  0.24  1.00  0.20  0.18
MRFA5 -0.05  0.29  0.20  1.00  0.17
MRFA4  0.01  0.13  0.18  0.17  1.00

Mean item complexity =  1.5
Test of the hypothesis that 5 factors are sufficient.

df null model =  300  with the objective function =  9.59 with Chi Square =  23262.17
df of  the model are 185  and the objective function was  1.01

The root mean square of the residuals (RMSR) is  0.03
The df corrected root mean square of the residuals is  0.04

The harmonic n.obs is  2436 with the empirical chi square  1596.69  with prob <  5.9e-223
The total n.obs was  2436  with Likelihood Chi Square =  2439.42  with prob <  0

Tucker Lewis Index of factoring reliability =  0.841
RMSEA index =  0.071  and the 90 % confidence intervals are  0.068 0.073
BIC =  996.77
Fit based upon off diagonal values = 0.98
Measures of factor score adequacy
                                                  MRFA2 MRFA1 MRFA3 MRFA5 MRFA4
Correlation of (regression) scores with factors    0.96  0.94  0.94  0.94  0.90
Multiple R square of scores with factors           0.93  0.89  0.88  0.88  0.81
Minimum correlation of possible factor scores      0.85  0.77  0.75  0.75  0.63
```

In the first table, we see that the items presented higher factor loadings in their respective factors, also in accordance with the theory. Below this table we see the amount of variance explained. The MRFA2 factor explained 24% of the data variance, while the MRFA4 factor explained 16%. Other information is the correlation between factors, χ^2, degrees of freedom, TLI, RMSEA, BIC, RMSR, among other adjustment indices.

We will hear more about fit indices. In general, we use adequacy indices to know whether our tested model (i.e., the 5-factor model) is adequate enough to explain our data.

5.4 How to report a Factor Analysis

Exploratory factor analysis showed that the data were suitable for analysis KMO = 0.85; Bartlett's test of sphericity, ^{2}(300; $N = 2436$)= 23262.2, $p < 0.001$. Parallel analysis suggested the extraction of five factors. The fifth empirical factor explained 7.98% of the data variance, while the fifth simulated average factor explained 7.51% of the variance. Additionally, for the adequacy indices, the scale presented the following statistics ^{2}(185, $N = 2436$) = 23262.17, p $<$ 0.001; TLI = 0.941; RMSEA = 0.071 (90% CI 0.068–0.073).

5.5 References

Brown, T. A. (2015). *Confirmatory factor analysis for applied research Second Edition.* The Guilford Press.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*(3), 272-299. https://doi.org/10.1037/1082-989X.4.3.272

Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (2006). *Multivariate data analysis with readings* (Vol. 6). Pearson Prentice Hall.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*(2), 179-185. https://doi.org/10.1007/BF02289447

Hutcheson, G. D. & Sofroniou, N. (1999). *The multivariate social scientist: Introductory statistics using generalized linear models.* Sage Publications

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and psychological measurement.*

Kaiser, H. F. (1970). A second generation little jiffy. *Psychometrika*, *35*(4), 401-415. https://doi.org/10.1007/BF02291817

Kaiser, H. F., & Rice, J. (1974). Little jiffy, mark IV. *Educational and psychological measurement*, *34*(1), 111-117. https://doi.org/10.1177/001316447403400115

Navarro-Gonzalez D, Lorenzo-Seva U (2021). *EFA.MRFA: Dimensionality Assessment Using Minimum Rank Factor Analysis.* R package. https://CRAN.R-project.org/package=EFA.MRFA.

Osborne, J. W. (2014). *Best Practices in Exploratory Factor Analysis.* CreateSpace Independent Publishing. ISBN-13: 978-1500594343, ISBN-10:1500594342.

R Core Team (2023). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Revelle, W. (2023). *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, Illinois. R package. https://CRAN.R-project.org/package=psych.

Tabachnick, B. G. & Fidell, L. S. (2007). *Using Multivariate Statistics* (5th. ed.). Allyn and Bacon.

Zwick, W.R. & Velicer, W.F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin, 99*, 432-442. https://doi.org/10.1037/0033-2909.99.3.432

6 Confirmatory Factor Analysis

In this text, I will present the fundamentals of the Confirmatory Factor Analysis (CFA), and how it works, and we will compare CFA with Exploratory Factor Analysis.

6.1 What is it and When Do We Apply Confirmatory Factor Analysis?

The CFA is a multivariate statistic that serves to estimate the structure of an instrument, verifying how well the measured variables represent the number of constructs. That is, it verifies whether an instrument's structure can be, but is not necessarily, true. For this, we need to state which structure we want to test. Generally, the CFA is used when there is a previous study that tells us the dimensionality of that instrument. For instance, we would have a North American study that uses an EFA to verify the instrument's dimensionality and you use a CFA to verify how well this structure happens with Brazilian data. However, this is not the only way you can use the CFA! You can, for example, have the EFA in the same study (to explore the dimensionality), but still test different theoretical models using the CFA.

Thus, both EFA and CFA are applied when you want to estimate the dimensionality of an instrument (note that I said estimate, not explore/discover dimensionality). For example, we can apply the CFA in self-report instruments, where items represent behaviors, thoughts, or feelings. Another example, we can apply it to a set of other measures, such as psychophysical measures of anxiety. Thus, CFA applies to instruments that measure some attributes such as well-being, anxiety, prejudice, etc.

6.2 Model Specification

The model from a CFA is similar, but not equal to, the model from an EFA. The model can be described as:

$$x = \Lambda_x \xi + \delta$$

$$y = \Lambda_y \eta + \epsilon$$

Where x and y are observed variables, ξ and η are latent factors, and δ and ϵ are measurement errors. Both formulas yield the same basic model, where an observed variable depends on one or more latent variable and a measurement error. Remember that the measurement error is considered to be uncorrelated with the latent variables.

Imagine we have eight items ($x1$ to $x8$), where the first four items' measures extroversion, and the last four measures neuroticism. Let's assume extroversion has no effects on the indicators of neuroticism. In addition, each indicator contains a measurement error that is assumed uncorrelated with the latent variables. The matrix equation the represents these relations are:

$$\begin{bmatrix} x1 \\ x2 \\ x3 \\ x4 \\ x5 \\ x6 \\ x7 \\ x8 \end{bmatrix} = \begin{bmatrix} \lambda_{1,1} & 0 \\ \lambda_{2,1} & 0 \\ \lambda_{3,1} & 0 \\ \lambda_{4,1} & 0 \\ 0 & \lambda_{5,2} \\ 0 & \lambda_{6,2} \\ 0 & \lambda_{7,2} \\ 0 & \lambda_{8,2} \end{bmatrix} \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} + \begin{bmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \\ \delta_4 \\ \delta_5 \\ \delta_6 \\ \delta_7 \\ \delta_8 \end{bmatrix}$$

$$COV(\eta_i, \delta_j) = 0$$

for all i and j

$$E(\delta_j) = 0$$

for all j.

We can think η_1 as an extroversion latent variable and η_2 as a neuroticism latent variable, where the first column of λ's are factor loadings (the direct structural relation between a latent and observed variable; may be viewed as regression coefficients) for extroversion, and the second column are factor loadings for neuroticism. The double subscript of $\lambda_{i,j}$ represent the row (item) and column (latent variable) positions. A zero in λ represent that the corresponding observed variable is not influenced by the latent variable in that column.

6.3 Differences Between Saturated/Unrestricted Model and Restricted Model

The EFA model can be called the saturated/unrestricted model. This is because all latent dimensions explain the variation in all items, as exemplified in the image below (Figure 6.1).

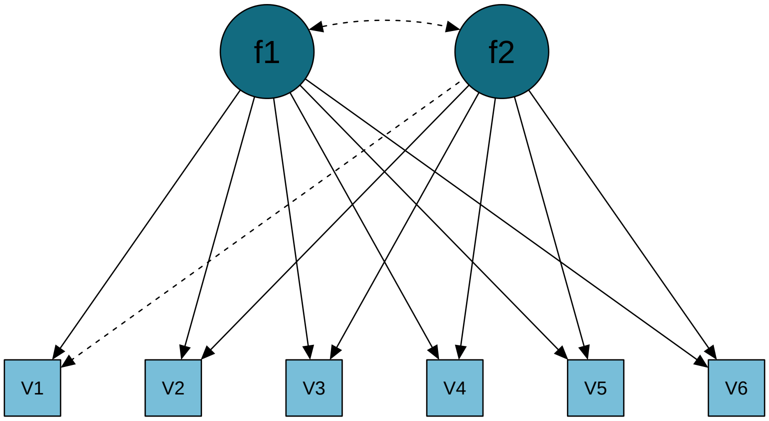


Figure 6.1: Exploratory Factor Analysis Model

As for the CFA, we can call it the Restricted Model, that is, we impose some restrictions on the model, for example, not having cross-loadings of a factor with items from another factor. The restricted model is exemplified in the image below (Figure 6.2).

Of course, there are some practical differences between one model and another. The first is that, generally, the output of the factor loadings from a CFA is different from the EFA. While in EFA we have cross loads on all factors, in the CFA some loadings are set at 0 (Table 6.1).

Table 6.1: Factor Loadings of a CFA.

Items	Factor 1	Factor 2
V1	0.6	0.0
V2	0.7	0.0
V3	0.8	0.0
V4	0.0	0.8
V5	0.0	0.5
V6	0.0	0.9

I made Table 6.2 that shows the differences between the unrestricted model (EFA) and the restricted model (CFA).

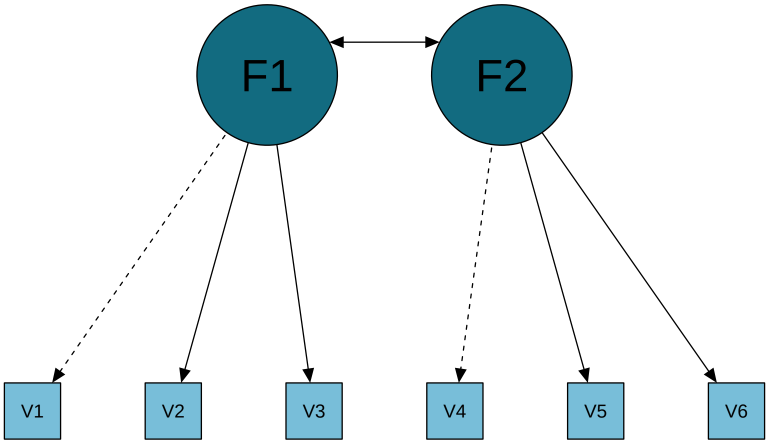


Figure 6.2: Confirmatory Factor Analysis Model.

Table 6.2: Differences Between EFA and CFA.

	EFA	CFA
Explore data dimensionality	Yes	No
Require defined hypothesis	No	Yes
Choose items	Yes	No
Test Models	Yes	Yes
Prove models	No	No
Fit Indices Available	Yes	Yes
Restricted Model	No	Yes
Unrestricted/Saturated Model	Yes	No
Modifications/Residual Correlations	No	Yes

We see in the table above that, for the confirmatory factor analysis, we need to have a defined hypothesis, that is, there must be a theory behind that will directly guide our analyses, we cannot just keep exploring without a proper justification. This is a little different from EFA, which has a theory behind the structure, but you test whether this structure will be corroborated in the data (through parallel analysis and the like). Of course, in EFA we can extract the factors based on theory, which, in a way, would resemble CFA in terms of the hypothesis guiding the analysis directly.

It is also important to emphasize again that in the CFA we can test different models, being able to make modifications and allow residual correlations. We can even test more complex models, such as a hierarchical model or a bifactor model. In short, because CFA makes restrictions on the model, we have the possibility to test a multitude of things! One use of CFA is through multi-group CFA.

6.4 Model Identification

We have to deal with a "problem" called model identification when we talk about a restricted model (Bollen, 1989). In other words, we need our data to have enough "information" to be able to do the necessary statistics.

Imagine if we were to estimate a one-factor model with 4 items (for example, estimating depression with a 4-item questionnaire). We therefore estimate 4 factor loadings (one per item), 4 residues (one per item), that is, we have 8 "information" to be discovered/estimated. The information we actually have is the item scores (for example, people's scores on items Item1, Item2, Item3, Item4) and the correlation between them. Count the cells of the correlation matrix between 4 items in Table 6.3.

Table 6.3: Correlation Table Between Items1 to Item4.

	Item1	Item2	Item3	Item4
Item1				
Item2	-0.05			
Item3	-0.13	0.03		
Item4	-0.04	0.04	-0.04	

Thus, we have 4 scores + 6 correlations = 10 pieces of information. In other words, with 4 items we can estimate the 8 pieces of factor loadings and residues since we have 10 pieces of information in our hands. Following this logic, it is easy to see that, in order to be able to identify the model, the minimum number of items is 3 items per latent factor. See, in a unifactorial model with 3 items, we will estimate 3 factor loadings + 3 residues = 6 necessary information. We have information for 3 items + 3 correlations = 6 information in our sleeve. So we will have 0 degrees of freedom (DF).

- If DF $<$ 0, the unidentified model (nothing will be estimated);
- If DF = 0, the model is under-identified (only factor loadings will be arbitrarily estimated; no fit indexes will be generated);
- If DF$>$ 1, the overidentified model (everything can be estimated).

A model should only be interpreted if DF> 1, as this is the only way to solve the covariance equation of items and latent variables, allowing the output of fit indices.

6.5 Fit Indices

The validity of psychometric models depends on the validity of the causal assumptions they make, which are generally implicit to the user. Psychological tests (e.g., self-report questionnaires) are typically constructed to measure constructs, while the responses observed in such tests are believed to reflect the underlying latent variable (Van Bork et al., 2017). For example, a person's self-esteem is not observed directly, but we assume that it can be measured through the use of items from an instrument. Various fit indices have been created trying to figure out whether the data fits a specific model. However, the causal assumption regarding the relationship between constructs and their indicators is often ignored in commonly used fit indices (Bartholomew, 1995; Franco et al., 2023).

6.5.1 Fit Indices Commonly Used in Factor Analysis

To test whether the theoretical model reflects the data causally, several fit indices have been developed to try to achieve this. Two main classes of fit indices have been proposed to try to operationalize the "goodness" (or "badness") of models (Xia & Yang, 2019): absolute fit indices and incremental fit indices.

Absolute fit indices assess how far the fitted model is from a "perfect" model, while a "perfect" model is defined as the model that can perfectly predict the values of the observed correlation matrix. One of the most used absolute fit indices is the *Root Mean Squared Error of Approximation* (RMSEA; Steiger & Lind, 1980; Franco et al., 2023). Incremental fit indices, on the other hand, evaluate the performance of the fitted model compared to a "baseline" model. The base model, in this context, is normally defined as the model where all variables are considered independent and, therefore, should be the model with the worst possible fit. The *Comparative Fit Index* (CFI; Bentler, 1990) and the *Tucker-Lewis Index* (TLI; Tucker & Lewis, 1973) are two of the most commonly used incremental fit indices. Regardless of whether a fit index is incremental or absolute, the "quality" of the fit is defined according to the objective function of the factor model, which is usually defined in terms of some type of difference between the observed correlation matrix and the implied correlation matrix from the adjusted model (Franco et al., 2023).

The main objective of factor analysis is to find a structure of latent causes that can be used to explain the correlational structure of observed data. Fit indices are, then, a way of checking whether the identified model is, in fact, good enough to explain the data. For a researcher to tell whether a model is "good enough" to explain the correlation structure of a data set, decisions based on fit indices depend on a set of cutoff criteria (Bentler & Bonett, 1980; Jöreskog & Sörbom, 1993). For example, Hu and Bentler (1999) demonstrated, through simulation studies, that an RMSEA less than 0.06 and a CFI and TLI greater than 0.95 indicate a relatively good fit of the model data to observed continuous variables. With nominal and ordinal data, however, these fit indices tend to be biased in the direction of good fit. Therefore, with nominal and ordinal data, more rigorous criteria or another decision criterion must be used for model selection (Xia & Yang, 2019).

For example, the article by Bonifay and Cai (2017) verified how the fit indexes of some models behaved. For this, a quasi-unrestricted model was tested (similar to the Exploratory Factor Analysis model, but with 2 loads being restricted to identify the model); a bifactor model; two hierarchical models; and a one-dimensional model. For this, fit indices were analyzed in 1000 simulated datasets. They found that, within all possible fits in these databases, the quasi-unconstrained model and the bifactor model almost always present good fit indices. This implies that we cannot interpret fit indices as good model indicators in these cases. For example, if you compare a bifactor model with a single-factor model, you will most likely find better fit indices in the bifactor, but this is not necessarily the best model to explain the data. Of course, one would have to compare fit indices of nested models, but the example serves as an illustration.

6.5.2 Criticisms of Factor Analysis

Some authors are critical of factorial models due to the lack of testing of causal assumptions, as shown by the network literature as an alternative way of explaining/describing the correlation patterns found between observed variables (Epskamp et al., 2018; Schmittmann et al., 2013). For example, McFarland's work (McFarland, 2020) states that psychometric networks of Gaussian graphical models and latent variable modeling (Kline, 2023) are alternatives to each other, where both can be applied to describe or explain the variance-covariance structure of the observed variables of interest. . In fact, some simulation (e.g., van Bork et al., 2021) and theoretical (e.g., Kruis & Maris, 2016) studies have shown that network and factor analytic models can sometimes explain the same patterns of correlation. This highlights a limitation of fit indices such as RMSEA, CFI and TLI for assessing the "quality" of factor models: they do not necessarily consider the causal assumptions embedded in factor models.

Regardless, both absolute and incremental fit indices have been used to assess whether we have support for a factor or network model, models that have different causal assumptions (Kan et al., 2020; McFarland, 2020; see also Aristodemou et al., 2023). Therefore, a fit index that takes into account the causal structure assumed by factor models could, in principle, provide additional information necessary to perform a more appropriate model selection.

6.6 How to Run a Confirmatory Factor Analysis in R.

To run a Confirmatory Factor Analysis, we must first install the *lavaan* (Rosseel, 2012) package.

```
install.packages("lavaan")
```

And tell the program that we are going to use the functions from this package.

```
library(lavaan)
```

Then, we must have information on which model we should test. In other words, we have to know the theory behind some instrument: how many factors we have, which items represent which factors, whether or not the factors are correlated, etc.

Let's use the Holzinger and Swineford (1939) model as an example. We will save the model in the `HS.model` variable.

```
HS.model <- ' visual  =~ x1 + x2 + x3
              textual =~ x4 + x5 + x6
              speed   =~ x7 + x8 + x9 '
```

You can see in this code that `=~` is used when we have a latent variable (on the left), and we inform after `=~` which items belong to that factor (summing the items). By default, *lavaan* will correlate the factors. Let's leave it like that for now. Now we will run the analysis and save in the object `cfa.fit`.

```
cfa.fit <- cfa(model = HS.model,
               data = HolzingerSwineford1939,
               estimator = 'ml',
               ordered = FALSE
               )

summary(cfa.fit,
        fit.measures=TRUE,
        standardized=TRUE
        )
```

```
lavaan 0.6.17 ended normally after 35 iterations

  Estimator                                         ML
  Optimization method                           NLMINB
  Number of model parameters                        21

  Number of observations                           301

Model Test User Model:
  Test statistic                                85.306
  Degrees of freedom                                24
  P-value (Chi-square)                           0.000

Model Test Baseline Model:

  Test statistic                               918.852
  Degrees of freedom                                36
  P-value                                        0.000

User Model versus Baseline Model:

  Comparative Fit Index (CFI)                    0.931
  Tucker-Lewis Index (TLI)                       0.896

Loglikelihood and Information Criteria:

  Loglikelihood user model (H0)              -3737.745
  Loglikelihood unrestricted model (H1)      -3695.092

  Akaike (AIC)                                7517.490
  Bayesian (BIC)                              7595.339
  Sample-size adjusted Bayesian (SABIC)       7528.739

Root Mean Square Error of Approximation:

  RMSEA                                          0.092
  90 Percent confidence interval - lower         0.071
  90 Percent confidence interval - upper         0.114
  P-value H_0: RMSEA <= 0.050                    0.001
  P-value H_0: RMSEA >= 0.080                    0.840

Standardized Root Mean Square Residual:

  SRMR                                           0.065

Parameter Estimates:

  Standard errors                             Standard
  Information                                 Expected
  Information saturated (h1) model          Structured
```

```
Latent Variables:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  visual =~
    x1                1.000                               0.900    0.772
    x2                0.554    0.100    5.554    0.000    0.498    0.424
    x3                0.729    0.109    6.685    0.000    0.656    0.581
  textual =~
    x4                1.000                               0.990    0.852
    x5                1.113    0.065   17.014    0.000    1.102    0.855
    x6                0.926    0.055   16.703    0.000    0.917    0.838
  speed =~
    x7                1.000                               0.619    0.570
    x8                1.180    0.165    7.152    0.000    0.731    0.723
    x9                1.082    0.151    7.155    0.000    0.670    0.665

Covariances:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  visual ~~
    textual           0.408    0.074    5.552    0.000    0.459    0.459
    speed             0.262    0.056    4.660    0.000    0.471    0.471
  textual ~~
    speed             0.173    0.049    3.518    0.000    0.283    0.283

Variances:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
   .x1                0.549    0.114    4.833    0.000    0.549    0.404
   .x2                1.134    0.102   11.146    0.000    1.134    0.821
   .x3                0.844    0.091    9.317    0.000    0.844    0.662
   .x4                0.371    0.048    7.779    0.000    0.371    0.275
   .x5                0.446    0.058    7.642    0.000    0.446    0.269
   .x6                0.356    0.043    8.277    0.000    0.356    0.298
   .x7                0.799    0.081    9.823    0.000    0.799    0.676
   .x8                0.488    0.074    6.573    0.000    0.488    0.477
   .x9                0.566    0.071    8.003    0.000    0.566    0.558
    visual            0.809    0.145    5.564    0.000    1.000    1.000
    textual           0.979    0.112    8.737    0.000    1.000    1.000
    speed             0.384    0.086    4.451    0.000    1.000    1.000
```

The first argument you have to put the variable where you configured the model. The `data` argument must come with your database. As the data follows a normal distribution and is continuous, we will consider the Maximum Likelihood estimator and the items will not be considered as ordinal.

Now, let's analyze the result with the following function, where we ask for the fit indices, standardized loads and correlations.

The "Model Test User Model" represents the chi-square of the configured model. We also have several other adjustment indices, such as CFI, TLI, RMSEA and SRMR. We report fit indices as follows.

The Holzier and Swineford (1939) model had the following fit indices: ^{2}*(gl = 24) = 85,306, p < 0,001, CFI = 0,931, TLI = 0,896, RMSEA [IC 95%]= 0,092 [0,071 - 0,0114], SRMR = 0,065.*

The standardized factor loadings are in the "Latent Variables" part in the "Std.all" column. The p-values of each item are in the "P(>|z|)" column. We see that we do not have the loading of the first item of each factor. This is because we have to fix one of the loads to have the magnitude of the others as a parameter, and *lavaan* always fixes the first one by default (look at the *Estimate* column, which represents the non-standardized load). We can set other items and leave the first one to be estimated, just put `NA*` in front of the first item and set another item at `1*`. That way:

```
HS.model <- ' visual  =~ NA*x1 + 1*x2 + x3
              textual =~ NA*x4 + x5 + 1*x6
              speed   =~ NA*x7 + x8 + 1*x9 '

cfa.fit <- cfa(model = HS.model,
               data = HolzingerSwineford1939,
               estimator = 'ml',
               ordered = FALSE)

summary(cfa.fit,
        fit.measures=TRUE,
        standardized=TRUE)
```

```
lavaan 0.6.17 ended normally after 40 iterations

  Estimator                                         ML
  Optimization method                           NLMINB
  Number of model parameters                        21

  Number of observations                           301

Model Test User Model:

  Test statistic                                85.306
  Degrees of freedom                                24
  P-value (Chi-square)                           0.000

Model Test Baseline Model:

  Test statistic                               918.852
  Degrees of freedom                                36
  P-value                                        0.000

User Model versus Baseline Model:

  Comparative Fit Index (CFI)                    0.931
  Tucker-Lewis Index (TLI)                       0.896

Loglikelihood and Information Criteria:

  Loglikelihood user model (H0)              -3737.745
  Loglikelihood unrestricted model (H1)      -3695.092

  Akaike (AIC)                                7517.490
  Bayesian (BIC)                              7595.339
  Sample-size adjusted Bayesian (SABIC)       7528.739

Root Mean Square Error of Approximation:

  RMSEA                                          0.092
  90 Percent confidence interval - lower         0.071
  90 Percent confidence interval - upper         0.114
  P-value H_0: RMSEA <= 0.050                    0.001
  P-value H_0: RMSEA >= 0.080                    0.840

Standardized Root Mean Square Residual:

  SRMR                                           0.065

Parameter Estimates:

  Standard errors                             Standard
  Information                                 Expected
  Information saturated (h1) model          Structured

Latent Variables:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  visual =~
    x1                1.807    0.325    5.554    0.000    0.900    0.772
    x2                1.000                               0.498    0.424
    x3                1.318    0.239    5.509    0.000    0.656    0.581
  textual =~
    x4                1.080    0.065   16.703    0.000    0.990    0.852
    x5                1.202    0.072   16.760    0.000    1.102    0.855
    x6                1.000                               0.917    0.838
  speed =~
    x7                0.925    0.129    7.155    0.000    0.619    0.570
    x8                1.091    0.145    7.517    0.000    0.731    0.723
    x9                1.000                               0.670    0.665

Covariances:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  visual ~~
    textual           0.209    0.048    4.322    0.000    0.459    0.459
    speed             0.157    0.040    3.967    0.000    0.471    0.471
  textual ~~
    speed             0.174    0.048    3.592    0.000    0.283    0.283

Variances:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
   .x1                0.549    0.114    4.833    0.000    0.549    0.404
   .x2                1.134    0.102   11.146    0.000    1.134    0.821
   .x3                0.844    0.091    9.317    0.000    0.844    0.662
   .x4                0.371    0.048    7.779    0.000    0.371    0.275
   .x5                0.446    0.058    7.642    0.000    0.446    0.269
   .x6                0.356    0.043    8.277    0.000    0.356    0.298
   .x7                0.799    0.081    9.823    0.000    0.799    0.676
   .x8                0.488    0.074    6.573    0.000    0.488    0.477
   .x9                0.566    0.071    8.003    0.000    0.566    0.558
    visual            0.248    0.077    3.214    0.001    1.000    1.000
    textual           0.840    0.098    8.541    0.000    1.000    1.000
    speed             0.449    0.087    5.152    0.000    1.000    1.000
```

See that now items x2, x6 and x9 are fixed with a charge equal to 1.

Well, we see the covariances below, in the “Covariances” part. The standardized column for the covariance (the correlation) between the factors is also “Std.all”. We see that visual was correlated with textual (r = 0.459), visual with speed (r = 0.471), and textual with speed (r = 0.283), with all correlations being significant (column “P(>|z|))”.

What if in our model we theorize that there is no correlation between factors? We have a few more things to add to the code. See in the previous output that the correlation is expressed by `~~`. Also, remember that to set a parameter to some number, we multiply with `*` in the model.

Then, the code with all orthogonal (i.e., uncorrelated) factors.

```
HS.model <- ' visual  =~ x1 + x2 + x3
              textual =~ x4 + x5 + x6
              speed   =~ x7 + x8 + x9

              visual ~~ 0*textual
              visual ~~ 0*speed
              textual ~~ 0*speed
              '

cfa.fit <- cfa(model = HS.model,
               data = HolzingerSwineford1939,
               estimator = 'ml',
               ordered = FALSE
               )

summary(cfa.fit,
        fit.measures=TRUE,
        standardized=TRUE
        )
```

```
lavaan 0.6.17 ended normally after 32 iterations

  Estimator                                         ML
  Optimization method                           NLMINB
  Number of model parameters                        18

  Number of observations                           301

Model Test User Model:

  Test statistic                               153.527
  Degrees of freedom                                27
  P-value (Chi-square)                           0.000

Model Test Baseline Model:
  Test statistic                               918.852
  Degrees of freedom                                36
  P-value                                        0.000

User Model versus Baseline Model:

  Comparative Fit Index (CFI)                    0.857
  Tucker-Lewis Index (TLI)                       0.809

Loglikelihood and Information Criteria:

  Loglikelihood user model (H0)              -3771.856
  Loglikelihood unrestricted model (H1)      -3695.092

  Akaike (AIC)                                7579.711
  Bayesian (BIC)                              7646.439
  Sample-size adjusted Bayesian (SABIC)       7589.354

Root Mean Square Error of Approximation:

  RMSEA                                          0.125
  90 Percent confidence interval - lower         0.106
  90 Percent confidence interval - upper         0.144
  P-value H_0: RMSEA <= 0.050                    0.000
  P-value H_0: RMSEA >= 0.080                    1.000

Standardized Root Mean Square Residual:

  SRMR                                           0.161

Parameter Estimates:

  Standard errors                             Standard
  Information                                 Expected
  Information saturated (h1) model          Structured

Latent Variables:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  visual =~
    x1                1.000                               0.724    0.621
    x2                0.778    0.141    5.532    0.000    0.563    0.479
    x3                1.107    0.214    5.173    0.000    0.801    0.710
  textual =~
```

```
    x4                1.000                               0.984    0.847
    x5                1.133    0.067   16.906    0.000    1.115    0.866
    x6                0.924    0.056   16.391    0.000    0.910    0.832
  speed =~
    x7                1.000                               0.661    0.608
    x8                1.225    0.190    6.460    0.000    0.810    0.801
    x9                0.854    0.121    7.046    0.000    0.565    0.561

Covariances:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  visual ~~
    textual           0.000                               0.000    0.000
    speed             0.000                               0.000    0.000
  textual ~~
    speed             0.000                               0.000    0.000

Variances:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
   .x1                0.835    0.118    7.064    0.000    0.835    0.614
   .x2                1.065    0.105   10.177    0.000    1.065    0.771
   .x3                0.633    0.129    4.899    0.000    0.633    0.496
   .x4                0.382    0.049    7.805    0.000    0.382    0.283
   .x5                0.416    0.059    7.038    0.000    0.416    0.251
   .x6                0.369    0.044    8.367    0.000    0.369    0.308
   .x7                0.746    0.086    8.650    0.000    0.746    0.631
   .x8                0.366    0.097    3.794    0.000    0.366    0.358
   .x9                0.696    0.072    9.640    0.000    0.696    0.686
    visual            0.524    0.130    4.021    0.000    1.000    1.000
    textual           0.969    0.112    8.640    0.000    1.000    1.000
    speed             0.437    0.097    4.520    0.000    1.000    1.000
```

See that all correlations are set to 0. You can set any value for any parameter, but remember to have a theory behind it to support it.

Of course, we can also do the analysis for ordinal data. Generally, for ordinal data we use another estimator, "WLSMV", and put the argument ordered = TRUE. It would look like this (but will not work on this data, since the data is not ordinal):

```
cfa.fit <- cfa(model = HS.model,
               data = HolzingerSwineford1939,
               estimator = 'WLSMV',
               ordered = TRUE
               )
```

We can calculate from people's factor scores. Factor scores work like when you calculate the average of an instrument to correlate with others, but calculating averages has certain assumptions, while factor scores have others. So, to calculate the factor scores just use the following code.

```
data_with_scores <- lavPredict(
  cfa.fit,
  type = "lv",
  method = "EBM",
  label = TRUE,
  append.data = TRUE,
  optim.method = "bfgs"
  )
```

We see that in the variable `data_with_scores` the factor scores of each subject were calculated and these scores were added to their database.

6.7 References

Aristodemou, M. E., Kievit, R. A., Murray, A. L., Eisner, M., Ribeaud, D., & Fried, E. I. (2023). Common Cause Versus Dynamic Mutualism: An Empirical Comparison of Two Theories of Psychopathology in Two Large Longitudinal Cohorts. *Clinical Psychological Science*, 21677026231162814. https://doi.org/10.1177/21677026231162814

Bartholomew, D. J. (1995). Spearman and the origin and development of factor analysis. *British Journal of Mathematical and Statistical Psychology*, *48*(2), 211-220. https://doi.org/10.1111/j.2044-8317.1995.tb01060.x

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*, 238–246. https://doi.org/10.1037/0033-2909.107.2.238

Bentler, P.M.,& Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, *88*, 588–606. https://doi.org/10.1037/0033-2909.88.3.588

Bollen, K. A. (1989). *Structural equations with latent variables.* John Wiley & Sons.

Bonifay, W., & Cai, L. (2017). On the complexity of item response theory models. *Multivariate behavioral research*, *52*(4), 465-484. https://doi.org/10.1080/00273171.2017.1309262

Borsboom, D., & Cramer, A. O. (2013). Network analysis: an integrative approach to the structure of psychopathology. *Annual review of clinical psychology*, *9*, 91-121. https://doi.org/10.1146/annurev-clinpsy-050212-185608

Epskamp, S., Maris, G., Waldorp, L. J., & Borsboom, D. (2018). Network psychometrics. T*he Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development*, 953-986. https://doi.org/10.1002/9781118489772.ch30

Franco, V. R., Bastos, R. V., & Jiménez, M. (2023, June). *Tetrad Fit Index for Factor Analysis Models*. Paper presented at Virtual MathPsych/ICCM 2023. Via mathpsych.org/presentation/1297.

Holzinger, K. J., & Swineford, F. (1939). A study in factor analysis: The stability of a bi-factor solution. *Supplementary educational monographs*.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1–55. https://doi.org/10.1080/10705519909540118

Kan, K. J., de Jonge, H., van der Maas, H. L., Levine, S. Z., & Epskamp, S. (2020). How to compare psychometric factor and network models. *Journal of Intelligence, 8*(4), 35. https://doi.org/10.3390/jintelligence8040035

Kline, R. B. (2023). *Principles and practice of structural equation modeling*. Guilford publications.

Kruis, J., & Maris, G. (2016). Three representations of the Ising model. *Scientific reports*, *6*(1), 34175. https://doi.org/10.1038/srep34175

McFarland, D. (2020). The effects of using partial or uncorrected correlation matrices when comparing network and latent variable models. *Journal of Intelligence*, *8*(1), 7. https://doi.org/10.3390/jintelligence8010007

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2), 1-36. https://doi.org/10.18637/jss.v048.i02

Schmittmann, V. D., Cramer, A. O., Waldorp, L. J., Epskamp, S., Kievit, R. A., & Borsboom, D. (2013). Deconstructing the construct: A network perspective on psychological phenomena. *New ideas in psychology*, *31*(1), 43-53. https://doi.org/10.1016/j.newideapsych.2011.02.007

Steiger, J. H.,&Lind, J. C. (1980). Statistically based tests for the number of common factors. *Paper presented at the Annual Meeting of the Psychometric Society*, Iowa City, IA.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, *38*, 1–10. https://doi.org/10.1007/BF02291170

Van Bork, R., Wijsen, L. D., & Rhemtulla, M. (2017). Toward a causal interpretation of the common factor model. *Disputatio*, *9*(47), 581-601. https://doi.org/10.1515/disp-2017-0019

Xia, Y., & Yang, Y. (2019). RMSEA, CFI, and TLI in structural equation modeling with ordered categorical data: The story they tell depends on the estimation methods. *Behavior Research Methods*, *51*, 409-428. https://doi.org/10.3758/s13428-018-1055-2

7 Exploratory Graph Analysis

In psychology, education, and behavioral sciences, we use scales/instruments to measure a particular construct (e.g., anxiety, happiness). To do this, we usually have a questionnaire with X number of items and we want to know the number of latent factors that arise from these items. This is usually done with Factor Analysis, where the number of dimensions is usually estimated by examining the patterns of eigenvalues. Two of the most common methods that use eigenvalues are the Kaiser-Guttman criterion that selects factors that have eigenvalue > 1 and parallel analysis. However, many criticisms have been made regarding the performance of these methods in estimating dimensionality. For example, the Kaiser-Guttman criterion can either underestimate or overestimate the number of factors (Zwick & Velicer, 1986). Traditional parallel analysis tends to have inflated Type I errors in binary data (Green et al., 2016).

Due to these and other limitations, Golino & Epskamp (2017) proposed a new method for estimating the dimensionality of a scale, called Exploratory Graph Analysis (EGA). This chapter will be a brief introduction to recent developments in EGA, aiming to disseminate this method.

7.1 About the Method

Network psychometrics methods have recently gained attention in psychological sciences literature. This may be due to the change in theoretical interpretation of correlations observed in the data. Traditionally, as done by EFA, psychometric models assume that latent causes explain observed behavior (i.e., items). Emerging areas, such as network psychometrics, present promising models for research in psychology, as they support theoretical perspectives on complexity, that is, they consider psychological attributes as systems of observed behaviors that reinforce each other in a dynamic way.

There is a relationship between a typical latent variable in traditional psychometrics and in network clusters. As said by Golino & Epskamp (2017), if the data-generating mechanism is the reflective model, the data cluster items into its' factors. For instance, if the data-generating mechanism of the Big-5 personality model is reflective, we will have 5 clusters in a network, one of each factor.

Thus, EGA is a dimensionality estimation method, just like factor analysis. EGA is also an exploratory method that does not depend on a priori assumptions, therefore, it does not require guidance from the researcher. In EGA, vertices represent variables (i.e., items) and edges represent the relationship (i.e., correlations) between two vertices. Golino et al. (2020) showed that the EGA method performs as well as the best techniques for selecting the number of factor analysis dimensions. Furthermore, EGA was one of the methods with the highest accuracy in general.

To run an EGA, we have, basically, 3 steps:

a. Determine redundancies in items through Unique Variable Analysis.

b. Perform the EGA itself.

c. Check the stability of the structure found by EGA, through bootstrap.

7.2 Unique Variable Analysis (UVA)

In reflective models (i.e., the Factor Analysis model), observed variables correlate only because they have a common cause (i.e., the latent trait). The name for this is local independence, that is, the idea that the correlation, or dependence, observed between items is explained exclusively by the latent trait.

Items are considered redundant when, even after considering the latent trait as the cause of the correlation, the items still have a strong correlation. This correlation of items that are independent of the latent trait may cause unintended effects when estimating dimensionality in psychometric modeling. Correlation between items for latent variable models has the potential to cause a violation of the principle of local independence, resulting in poor fit (Christensen et al., 2023).

For these reasons, we first use Unique Variable Analysis (UVA) to detect local independence before estimating any factors. UVA (Christensen et. al., 2023) uses the weighted topological overlap measure (Nowick et al., 2009) in an estimated network. The weighted topological overlap measure is calculated as:

$$\boldsymbol{\omega} = \frac{\boldsymbol{\Sigma}_u a_{iu} a_{uj} + a_{ij}}{min\{k_i, k_j\} + 1 - a_{ij}}$$

where a_{ij} is the weight of the edge between vertices i and j. u represents the shared connections with other edges for edges i and j. k represents the sum of all connections for a given edge.

The UVA algorithm first calculates the association structure of observed data and then uses a threshold or significance test to determine redundancy between pairs of variables (Christensen et al., 2023). Values greater than 0.25 are determined to have considerable local dependence (i.e., redundancy) that must be addressed. By default, UVA will remove all redundant variables ($\boldsymbol{\omega}$ 0.25) except one based on the following rules:

- duplets (two variables): The variable with the smallest maximum weighted topological overlap for all other variables (except the one with which it is redundant) is kept and the other is removed.
- triplets (three or more variables): The variable with the highest weighted average topological overlap for all other variables that are redundant with each other is kept and all others are removed.

7.3 Exploratory Graph Analysis

Let's explain a little about EGA step by step: what a partial correlation is, glasso, EBIC and Walktrap.

7.3.1 Partial Correlation

Just like a standard linear correlation (which we often use), partial correlation represents the degree of association between two variables. However, unlike the standard linear correlation, the partial correlation calculates this association between two variables by controlling for all other variable correlations that you put into the model. In the EGA example, we see the relationship between an Item 1 and Item 2 controlling the effect of all other items. To calculate, simply calculate the inverse of the covariance matrix. As the aim is not to teach how to calculate the inverse of a matrix, we leave it as homework.

7.3.2 What is EBICglasso

When we are calculating relationships between variables, we can have several spurious correlations. In the case of network analysis, these spurious correlations are removed to better identify the model. The glasso algorithm directly penalizes the elements of the variance-covariance matrix, turning them into zero when we have correlations that are low. It works as follows:

GLASSO is a regularization technique that reduces parameter estimates with some estimates, making them exactly zero. LASSO uses a parameter called lambda ($\boldsymbol{\lambda}$), which controls the sparsity of the network. Smaller values of $\boldsymbol{\lambda}$ remove fewer edges (i.e., relationships between variables), increasing the possibility of including spurious correlations, while higher values of $\boldsymbol{\lambda}$ remove more edges, increasing the possibility of removing relevant edges. When $\boldsymbol{\lambda} = 0$, the estimates are equal to the ordinary least squares solution to the partial correlation matrix.

The popular approach in the network psychometrics literature is to compute models at various values of (usually 100) and select the model that minimizes the Extended Bayesian Information Criterion (EBIC). EBIC model selection uses a gamma hyperparameter ($\boldsymbol{\gamma}$) to control how much it prefers simpler models (i.e. models with fewer edges). Larger values of $\boldsymbol{\gamma}$ lead to simpler models, while smaller values of $\boldsymbol{\gamma}$ lead to denser models. When $\boldsymbol{\gamma} = 0$, the EBIC is equal to the Bayesian Information Criterion.

7.3.3 The Clustering Algorithm: An Example with Walktrap

This is a hierarchical clustering algorithm, and uses the following step by step:

1. The algorithm begins by computing a transition matrix, where each element represents the probability of one vertex arriving at another (based on the strength of the vertex).
2. Start random walks for a certain step number (e.g., 4) using the matrix for possible destinations.
3. Uses Ward's (1963) clustering procedure, where each vertex begins with its own cluster; then it joins adjacent clusters (reducing the squared distances between other clusters).
4. Modularity (Newman, 2006) is used to determine an optimal partition of clusters.
5. Each detected cluster represents a latent trait.

7.4 EGA's Stability

To verify the stability of the EGA results, the method used is the bootstrap. The EGA Bootstrap (Christensen & Golino, 2021) performs a parametric or resampling (non-parametric) procedure to determine the robustness of the EGA empirical analysis. Generally, 500 iterations/simulated databases are simulated with the same correlation pattern as your database. The output of the EGA bootstrap graph (bootEGA) is the median network structure that represents the median value of each pairwise partial correlation across the bootstraps. After obtaining the median value for each pairwise partial correlation, a community detection algorithm is applied.

The Dimension Stability output produces a graph of how many times each variable is replicating in its empirical structure through bootstraps. Structural consistency is defined as the extent to which each empirically derived dimension is exactly (i.e., identical variable composition) recovered from the replicated bootstrap samples (Christensen, Golino, & Silvia, 2020). In general, structural consistency and item stability values greater than 0.70-0.75 reflect sufficient stability (Christensen & Golino, 2021).

7.5 How to Run the EGA Step By Step in R

To run EGA, we first have to install the *EGAnet* (Golino & Christensen, 2023) package to perform the analyses, and the *psychTools* (Revelle, 2023) package to get a database, and *lavaan* (Rosseel, 2012) to request the fit indices.

```
install.packages("EGAnet")
install.packages("psychTools")
install.packages("lavaan")
```

And tell the program that we are going to use the functions of these packages.

```
library(EGAnet)
library(psychTools)
library(lavaan)
```

```
This is lavaan 0.6-17
lavaan is FREE software! Please report any bugs.
```

To demonstrate the step-by-step process I mentioned, we will use the bfi dataset from the *psychTools* package. We will use 25 items from a self-report personality questionnaire based on the Big-5 model. The bank has data on 2,800 subjects.

We will first select the bank's personality items, without considering the sociodemographic items.

```
bfi_items <- psych::bfi[,1:25]
```

7.5.1 UVA in R

To run the UVA in R, we use the following code in this database.

```
bfi_uva <- UVA(
              data = bfi_items,
              key = as.character(psychTools::bfi.dictionary$Item[1:25])
              )

# Show Results
bfi_uva
```

```
Variable pairs with wTO > 0.30 (large-to-very large redundancy)

            node_i                node_j   wto
 Get angry easily. Get irritated easily. 0.431

----

Variable pairs with wTO > 0.25 (moderate-to-large redundancy)

----

Variable pairs with wTO > 0.20 (small-to-moderate redundancy)

                                    node_i
                         Don't talk a lot.
                   Am exacting in my work.
 Am indifferent to the feelings of others.
           Do things in a half-way manner.
               Know how to comfort others.
                         Get angry easily.
                Have frequent mood swings.
         Inquire about others' well-being.
                                node_j   wto
 Find it difficult to approach others. 0.226
 Continue until everything is perfect. 0.225
     Inquire about others' well-being. 0.219
                        Waste my time. 0.209
             Make people feel at ease. 0.207
            Have frequent mood swings. 0.205
                      Often feel blue. 0.204
           Know how to comfort others. 0.203
```

Based on the above result, there are a couple of variables that are above the acceptable threshold: Getting angry easily. and gets irritated easily. ($\boldsymbol{\omega}$ = 0.431). Variables that were removed in this automated process can be viewed using:

```
bfi_uva$keep_remove
```

```
$keep
[1] "Get irritated easily."

$remove
```

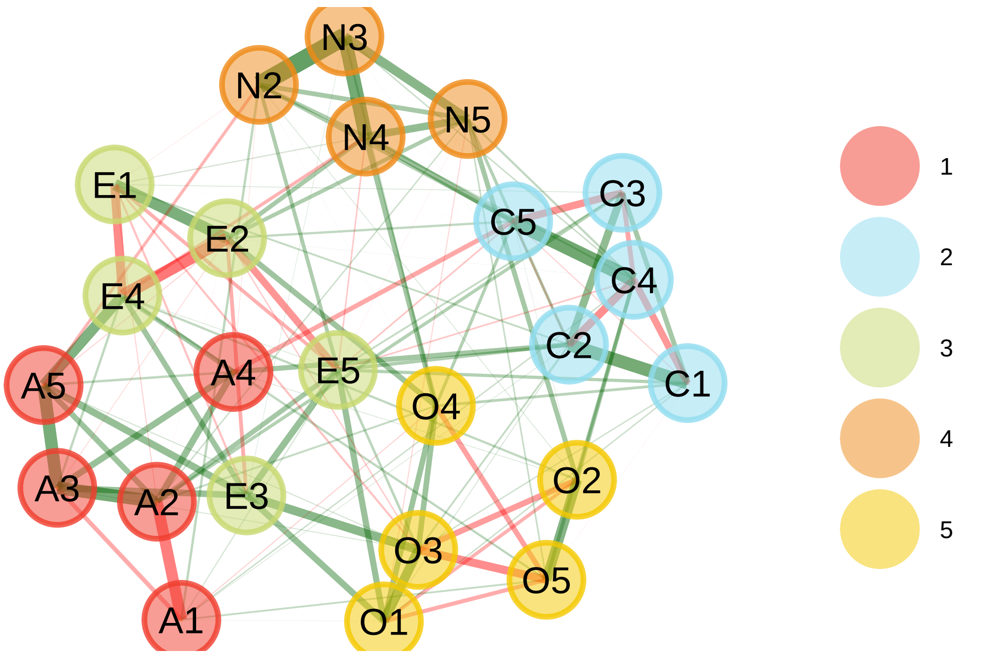
```
[1] "Get angry easily."
```

Next, we will work with the dataset without the redundant item obtained from the UVA function.

7.5.2 EGA in R

With item redundancies addressed, EGA is ready to be applied to your questionnaire. Just use the following function.

```
bfi_ega <- EGA(data = bfi_uva$reduced_data)
```



We see that five dimensions are estimated, as each color represents a factor (see the dots on the right). Therefore, the result is consistent with the five-factor model of personality. We can get a summary of this output in text format as follows.

```
summary(bfi_ega)
```

```
Model: GLASSO (EBIC with gamma = 0.5)
Correlations: auto
```

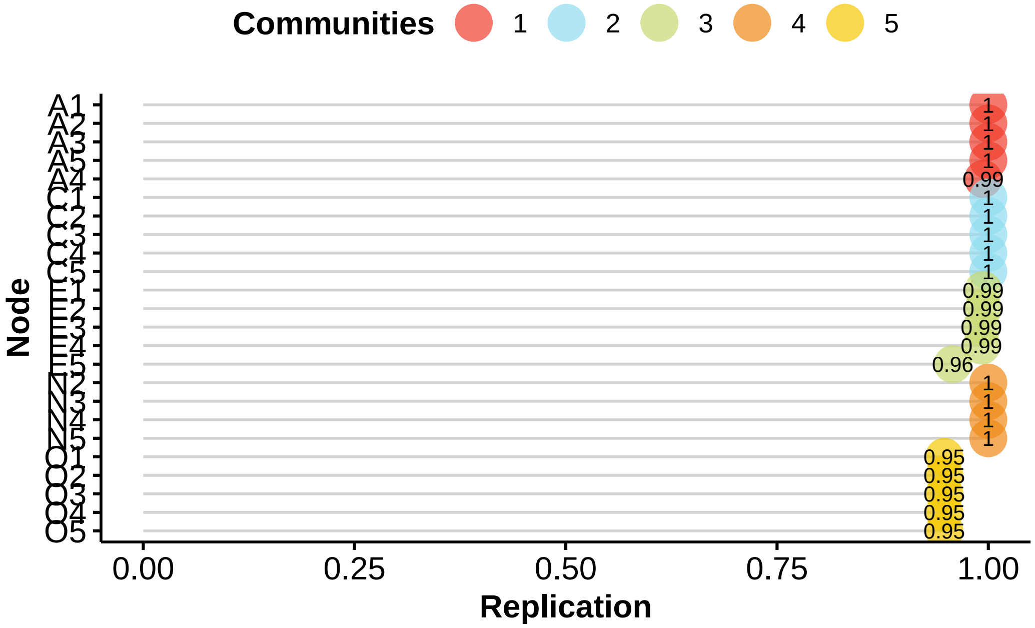
```
Lambda: 0.0597096451199323 (n = 100, ratio = 0.1)

Number of nodes: 24
Number of edges: 125
Edge density: 0.453

Non-zero edge weights:
     M    SD    Min   Max
 0.041 0.112 -0.270 0.396

----

Algorithm:  Walktrap

Number of communities:  5

A1 A2 A3 A4 A5 C1 C2 C3 C4 C5 E1 E2 E3 E4 E5 N2 N3 N4 N5 O1 O2 O3 O4 O5
 1  1  1  1  1  2  2  2  2  2  3  3  3  3  3  4  4  4  4  5  5  5  5  5

----

Unidimensional Method: Louvain
Unidimensional: No

----

TEFI: -24.989
```

The summary tells us which model was used to estimate the network (i.e., "glasso" explained earlier) and which parameters were used for that model, such as gamma ($\boldsymbol{\gamma}$=0.5) and lambda ($\boldsymbol{\lambda}$=0, 0597). Then there are descriptions about the network, such as the number of vertices, edges, edge density, and descriptive statistics about the edges. Third, it informs which community detection algorithm was used, the number of communities (dimensions) and the association of each variable. Fourth, the method to check whether the model is one-dimensional. Finally, the Total Entropy Fit Index (or tefi) is provided, which can be used for model comparison (see Golino et al., 2021).

7.5.3 EGA's Stability in R

To check how stable EGA is in your database, simply use the following code.

```
bfi_boot <- bootEGA(
  data = bfi_uva$reduced_data,
  seed = 2024
)
```



In this example, the median structure found in the bootstrap matches our empirical structure.

Although this result without much error is common, it is not always the case. This is because as a community detection algorithm is applied ad-hoc to the median network structure in the bootstrap, it is possible that the number and content of communities do not match the empirical structure. This possibility happens from time to time and does not mean there is anything wrong with your analysis, but it may suggest some instability in the structure.

Let's look at some basic descriptive statistics about bootstrap analysis.

```
summary(bfi_boot)
```

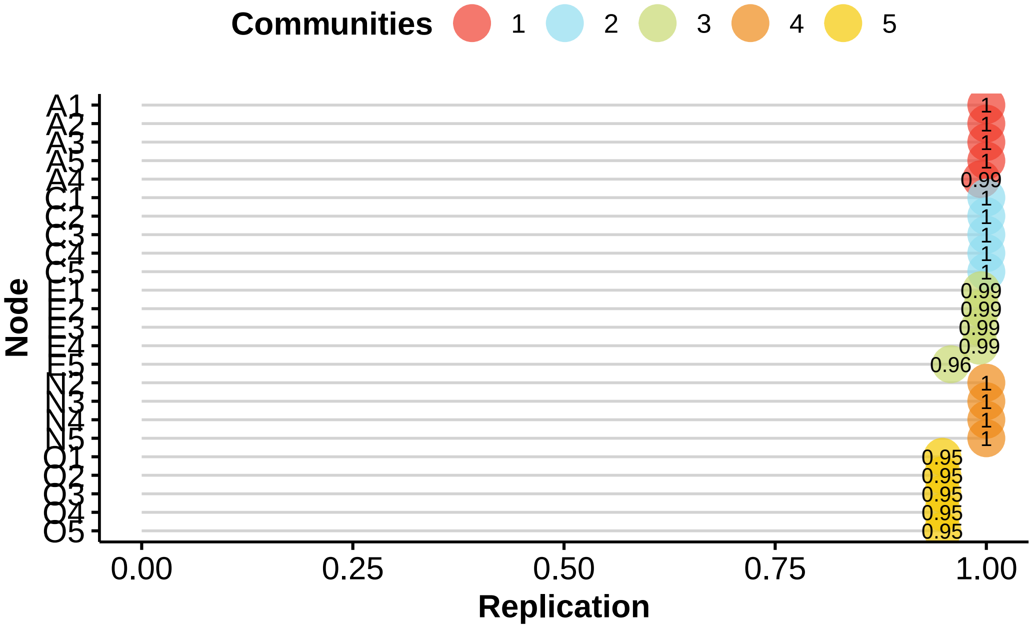
```
Model: GLASSO (EBIC)
Correlations: auto
Algorithm:  Walktrap
Unidimensional Method:  Louvain
```

```
----

EGA Type: EGA
Bootstrap Samples: 500 (Parametric)

                4     5
Frequency:  0.058 0.942

Median dimensions: 5 [4.54, 5.46] 95% CI
```

As with the empirical procedure, the first information is about the estimation methods and algorithms used. Then there is information about the bootstrap procedure including how often each number of communities was observed and the median number of communities (with 95% confidence intervals). In this example, the structure is quite stable and can be taken as preliminary evidence of a robust structure.

This summary alone cannot tell us if everything is ok with stability. To check this we have to use the following code.

```
dimensionStability(bfi_boot)
```



```
EGA Type: EGA
```

```
Bootstrap Samples: 500 (Parametric)

Proportion Replicated in Dimensions:

   A1    A2    A3    A4    A5    C1    C2    C3    C4    C5    E1    E2    E3 
1.000 1.000 1.000 0.994 1.000 1.000 1.000 1.000 1.000 1.000 0.994 0.994 0.992 
   E4    E5    N2    N3    N4    N5    O1    O2    O3    O4    O5 
0.992 0.958 1.000 1.000 1.000 1.000 0.948 0.948 0.948 0.948 0.948 

----

Structural Consistency:

    1     2     3     4     5 
0.994 1.000 0.966 1.000 0.948 
```

Our results demonstrate that the five-dimensional structure we identified is quite robust, given that all items replicated in their given dimension more than 70% or 75% of the simulated databases.

We can see network loadings (similar to factor loadings), with the code:

```
Network_loadings <- net.loads(bfi_boot$EGA)

# Print Results
print(Network_loadings, minimum = 0)
```

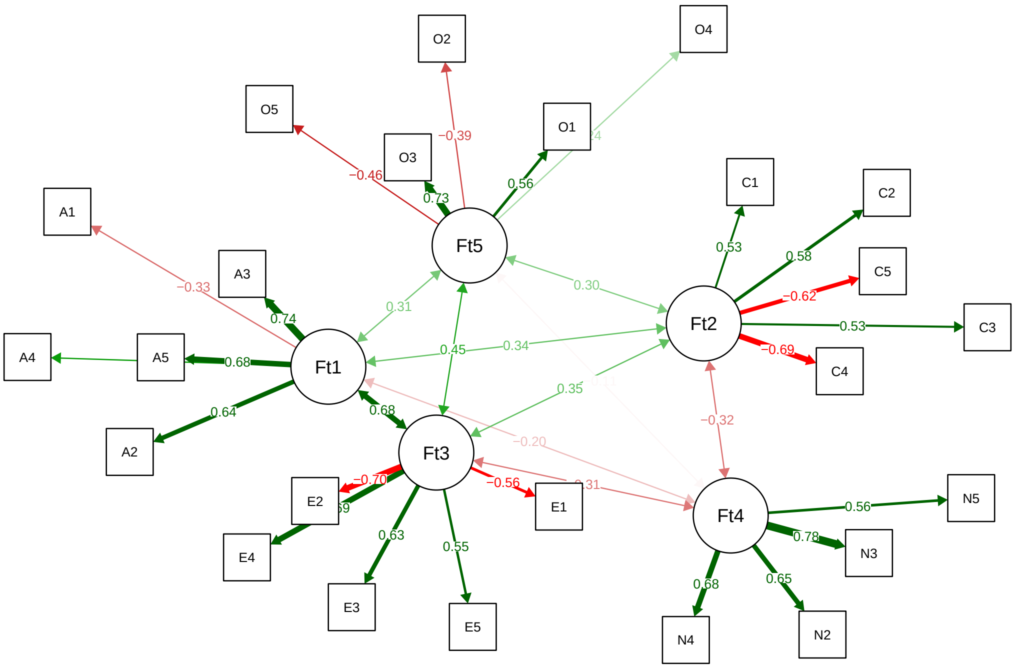
```
Loading Method: BRM

   1      2      3      4      5     
A2  0.384  0.034  0.058  0.016  0.021
A3   0.38      0  0.092      0   0.01
A5  0.217      0  0.178 -0.034  0.012
A4  0.163    0.1  0.046 -0.006 -0.012
A1 -0.166  0.022  0.011  0.027 -0.034
C2  0.062  0.311  0.052  0.036  0.029
C1      0  0.264  0.031      0   0.06
C3  0.035  0.253  0.026      0      0
C5  -0.05 -0.253 -0.038  0.107  0.039
C4 -0.015 -0.347 -0.018  0.062  -0.08
E4  0.188      0  0.263 -0.037 -0.048
E3  0.142      0  0.172  0.007  0.159
```

```
E5  0.064  0.122  0.172 -0.063  0.093
E1 -0.011  0.028  -0.25  -0.02 -0.023
E2 -0.021 -0.025 -0.333  0.095  0.074
N3      0  0.016  0.006  0.458  0.022
N2 -0.064 -0.036  0.041  0.298   0.01
N4      0 -0.081 -0.101  0.281  0.056
N5  0.015  0.054 -0.041  0.238 -0.079
O3  0.022  0.032  0.142      0  0.308
O1 -0.003  0.023   0.12 -0.011  0.239
O4  0.032  0.061 -0.053  0.083  0.174
O2  0.008 -0.045  0.022  0.064 -0.204
O5  0.023 -0.038  0.022  0.018 -0.289
Standardized loadings >= |0.00| are displayed. To change this 'minimum', use `print(net.loads
```

We were also able to obtain the adjustment through a Confirmatory Factor Analysis using *EGAnet*.

```
fit <- EGAnet::CFA(bfi_ega,
                   data = bfi_uva$reduced_data,
                   estimator = "WLSMV",
                   plot.CFA = TRUE,
                   layout = "spring"
                  )
```

```
[1] "A1" "A2" "A3" "A4" "A5"
[1] "C1" "C2" "C3" "C4" "C5"
[1] "E1" "E2" "E3" "E4" "E5"
[1] "N2" "N3" "N4" "N5"
[1] "O1" "O2" "O3" "O4" "O5"
```



To request fit indices we can use *lavaan*.

```
lavaan::fitMeasures(fit$fit,
                    fit.measures = "all"
                    )
```

```
                   npar                          fmin
                 82.000                         0.738
                  chisq                            df
               4132.221                       242.000
                 pvalue                  chisq.scaled
                  0.000                      3570.714
              df.scaled                 pvalue.scaled
                242.000                         0.000
   chisq.scaling.factor                baseline.chisq
                  1.157                     17655.215
            baseline.df               baseline.pvalue
                276.000                         0.000
  baseline.chisq.scaled            baseline.df.scaled
              14781.836                       276.000
baseline.pvalue.scaled baseline.chisq.scaling.factor
                  0.000                         1.194
```

```
                         cfi                           tli
                       0.776                         0.745
                  cfi.scaled                    tli.scaled
                       0.771                         0.738
                  cfi.robust                    tli.robust
                       0.777                         0.746
                        nnfi                           rfi
                       0.745                         0.733
                         nfi                          pnfi
                       0.766                         0.672
                         ifi                           rni
                       0.777                         0.776
                 nnfi.scaled                    rfi.scaled
                       0.738                         0.725
                  nfi.scaled                   pnfi.scaled
                       0.758                         0.665
                  ifi.scaled                    rni.scaled
                       0.771                         0.771
                 nnfi.robust                    rni.robust
                       0.746                         0.777
                        logl            unrestricted.logl
                 -109988.969                   -107922.858
                         aic                           bic
                  220141.938                    220628.802
                      ntotal                          bic2
                    2800.000                    220368.260
            scaling.factor.h1             scaling.factor.h0
                       1.155                         1.147
                       rmsea                rmsea.ci.lower
                       0.076                         0.074
               rmsea.ci.upper                rmsea.ci.level
                       0.078                         0.900
                rmsea.pvalue                rmsea.close.h0
                       0.000                         0.050
       rmsea.notclose.pvalue             rmsea.notclose.h0
                       0.000                         0.080
                rmsea.scaled         rmsea.ci.lower.scaled
                       0.070                         0.068
       rmsea.ci.upper.scaled           rmsea.pvalue.scaled
                       0.072                         0.000
rmsea.notclose.pvalue.scaled                  rmsea.robust
                       0.000                         0.076
       rmsea.ci.lower.robust         rmsea.ci.upper.robust
```

```
              0.074                          0.078
rmsea.pvalue.robust  rmsea.notclose.pvalue.robust
              0.000                          0.001
                rmr                     rmr_nomean
              0.144                          0.150
               srmr                   srmr_bentler
              0.070                          0.070
srmr_bentler_nomean                           crmr
              0.073                          0.073
        crmr_nomean                     srmr_mplus
              0.076                          0.070
  srmr_mplus_nomean                          cn_05
              0.073                        190.246
              cn_01                            gfi
            201.638                          0.992
               agfi                           pgfi
              0.990                          0.741
                mfi                           ecvi
              0.499                          1.534
```

7.6 References

Christensen, A. P., Garrido, L. E., & Golino, H. (2023). Unique variable analysis: A network psychometrics method to detect local dependence. *Multivariate Behavioral Research*, 1-18. https://doi.org/10.1080/00273171.2023.2194606

Christensen, A. P., & Golino, H. (2021). Estimating the stability of psychological dimensions via bootstrap exploratory graph analysis: A Monte Carlo simulation and tutorial. *Psych*, *3*(3), 479-500. https://doi.org/10.3390/psych3030032

Christensen, A. P., Golino, H., & Silvia, P. J. (2020). A psychometric network perspective on the validity and validation of personality trait questionnaires. *European Journal of Personality*, *34*(6), 1095-1108. https://doi.org/10.1002/per.2265

Green, S. B., Redell, N., Thompson, M. S., & Levy, R. (2016). Accuracy of revised and traditional parallel analyses for assessing dimensionality with binary data. *Educational and Psychological Measurement*, *76*(1), 5–21. https://doi.org/10.1177/0013164415581898

Golino, H., & Christensen, A. P. (2023). *EGAnet: Exploratory Graph Analysis – A framework for estimating the number of dimensions in multivariate data using network psychometrics*. R package.

Golino, H., & Epskamp, S. (2017). Exploratory graph analysis: A new approach for estimating the number of dimensions in psychological research. *PloS one*, *12*(6), e0174035. https://doi.org/10.1371/journal.pone.0174035

Golino, H., Moulder, R., Shi, D., Christensen, A. P., Garrido, L. E., Nieto, M. D., ... & Boker, S. M. (2021). Entropy fit indices: New fit measures for assessing the structure and dimensionality of multiple latent variables. *Multivariate Behavioral Research*, *56*(6), 874-902. https://doi.org/10.1080/00273171.2020.1779642

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Revelle, W. (2023). *psychTools: Tools to Accompany the 'psych' Package for Psychological Research*. Northwestern University, Evanston, Illinois. R package. https://CRAN.R-project.org/package=psychTools.

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2), 1-36. https://doi.org/10.18637/jss.v048.i02

Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, *103*, 8577–8582. https://doi.org/10.1073/pnas.0601602103

Nowick, K., Gernat, T., Almaas, E., & Stubbs, L. (2009). Differences in human and chimpanzee gene expression patterns define an evolving network of transcription factors in brain. *Proceedings of the National Academy of Sciences*, *106*(52), 22358-22363. https://doi.org/10.1073/pnas.0911376106

Ward, J. H. (1963). Hierarchical clustering to optimise an objective function. *Journal of the American Statistical Association*, *58*, 238–244.

Zwick, W.R. & Velicer, W.F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin*, *99*, 432–442. https://doi.org/10.1037/0033-2909.99.3.432

8 Social Desirability Bias

Much of the research carried out with human beings that measures behaviors, affections, personality, etc. use self-report scales (Lange & Dewitte, 2019; Peterson & Kerin, 1981). When responding to a questionnaire, some factors influence the response given to items that may or may not be associated with the latent trait being measured. Ideally, when we measure a construct, we want to measure it without many errors or spurious variations; however, it is possible that there is bias/response style that introduces spurious variations into our analyses. Some examples of these responses are: social desirability, acquiescence, and extreme responses.

8.1 Faking: The Good, The Bad, and the Ugly

Faking depends on the context of the application and the questionnaire applied. The person who uses *faking* aims to provide a representation of themselves that helps achieve a personal objective (Ziegler et al., 2011). Therefore, *faking* occurs when this set of responses is activated by situational demands and personal characteristics to produce systematic differences in test scores that are not due to the construct of interest. *Faking* is a behavior that is influenced by different factors and is, in essence, a matter of measurement (Ziegler et al., 2011).

Faking can be conceptualized as *faking good* and *faking bad. Faking good* is a conscious effort to manipulate responses to an instrument to make a positive impression (Zickar & Robie, 1999). *Faking bad* includes both the fabrication of clinical and/or diagnostic symptoms and the exaggeration of symptoms to obtain a specific secondary gain (Ziegler et al., 2011). One question that remains is: what makes people pretend?

Variables in *faking* models can be classified based on the type of belief a given variable is likely to impact. The expectancy theory of Ziegler et al. (2011) states that the choice to do *faking* or not is caused by: a) Belief that one is capable of doing *faking*; b) Belief that doing *faking* is important; c) Belief that the opportunity is valued. The belief that someone is capable of *faking* comes from different variables, such as personality traits, cognitive ability, knowledge and experience, as well as situational factors such as the degree of transparency of the item and the use of verification warnings that make an individual more or less capable of faking (Griffith et al., 2006; McFarland & Ryan, 2000; Raymark & Tafero, 2009; Riggio et al., 1988; Snell et al., 1999).

8.2 Faking Good: Social Desirability

A comprehensive survey of existing literature indicates that Social Desirability (SD) scales stand out as the most frequently utilized and explored measures of faking behavior (Ziegler et al., 2011). When individuals are prompted to evaluate how well certain traits reflect them, they often exhibit a tendency to endorse those traits if they are socially desirable (Edwards, 1957). As early as 1953, Edwards expressed skepticism regarding the accuracy of item scores on personality assessments, questioning whether respondents' answers genuinely reflected their personal attributes (Edwards, 1953). Furthermore, Edwards' research from that period demonstrated a direct correlation between the likelihood of endorsing an item and its level of social desirability. This inclination toward response bias may stem from various factors such as the experimental or testing environment, the motives of the subjects (e.g., aspirations for achievement, the desire for approval, etc.), or the individuals' anticipation of the evaluative outcomes of their actions (King & Bruner, 2000).

Concerning the impact of Social Desirability (SD) bias, it stands as one of the most prevalent sources of bias affecting the credibility of research findings within psychology and the social sciences (King & Bruner, 2000; Malhotra, 1988; Nederhof, 1985; Paulhus, 1991; Peltier and Walsh, 1990). When relying on self-reported data featuring socially desirable responses, it can lead to false associations between variables, potentially obscuring or weakening the relationships among the variables of interest (Connelly & Chang, 2016; Ganster et al., 1983; Kaiser et al., 2008; Paunonen & LeBel, 2012). Moreover, such bias can skew the average scores on trait questionnaires (Ziegler et al., 2007) and alter the internal structure of measurement instruments (Pettersson et al., 2012). Thus, it is strongly advised to implement methods for controlling or mitigating the influence of SD in research endeavors.

8.3 How to Represent Social Desirability

In the SD literature, there is an ongoing debate on the dimensionality of SD. The single-factor model has been challenged by two-factor models, suggesting that SD consists of two different (but related) factors (e.g., Paulhus, 1984). For example, Paulhus (1984, Paulhus & John, 1998) has presented evidence for the two-factor model of SD, where one factor is labeled impression management and the other is self-deception. More specifically, Paulhus and John (1998) state that SD consists of two self-favoring tendencies: 1) Alpha: an egoistic tendency to see oneself as an exceptionally talented and socially prominent member of society; 2) Gamma: a moralistic tendency, the view of oneself as an exceptionally good member of society. Ziegler et al. (2011) present an argument for the one-factor model. The authors state that it is necessary to show a correlation between scales to introduce a method factor such as SD; if there is no correlation then there is nothing to explain the importance of SD since it won't affect the correlations between other constructs. They also state that there can be method factors on a more specific level, however, scholars are often concerned only with factors influencing instruments in general, and the single-factor model is often enough for this (Ziegler et al., 2011).

8.4 Modeling Faking with Classical Test Theory

Since faking is a measurement issue, it's a necessary task to conceptualize faking within the psychometric theory. In a Classical Test Theory perspective, an individual's observed score (X) on a test can be expressed as a function of the person's true score (T) and error (E), such that

$$X = T + E$$

Then, in a set of observed scores for a sample of test takers, the variance in the observed scores can be expressed as a function of the variance in the true scores and the variance of the errors. Note that, in the equation below, there is an assumption that the error is random and unrelated to true scores. Then, when incorporating faking into the equation, the observed scores associated with faking cannot be due to random error. In other words, faking must be conceptualized as a component of a psychological true score (Ziegler et al., 2011).

$$\sigma^2_X = \sigma^2_T + \sigma^2_E$$

In a psychometric approach, it's common to conceptualize faking as a single, unitary source of systematic variance (e.g., Komar et al., 2008; Schmitt & Oswald, 2006). However, as stated in Ziegler et al. (2011), conceptualizing faking as a single source of systematic variance is an oversimplification, because it is a complex behavior, and the degree to which one fakes is a function of dispositional, attitudinal, and situational factors. In a motivating setting (where people will fake), we can express the observed scores as follows in the following equation:

$$X_{Motivated} = (T_T + (T_{F1} + ... + T_{Fn})) + E$$

where T_{F1} to T_{Fn} are systematic individual attitudinal, and situational factors that influence observed scores in motivating contexts. In a sample of scores, we can express the variance in observed scores obtained in motivated settings as follows in the equation:

$$\sigma^2_{XMotivated} = \sigma^2_{T_t} + (\sigma^2_{F1} + ... + \sigma^2_{Fn}) + (2\sigma^2_{T_T,F1} + ... + 2\sigma^2_{Fn-1,Fn}) + \sigma^2_E$$

8.5 Some Ways to Control Social Desirability

8.5.1 Correlations and Social Desirability Scales

Controlling SD bias has been a challenge in the literature (Leite & Cooper, 2009; Paulhus,1981; Ziegler et al., 2011). One of the basic forms of control is to use scales that measure 'pure' desirability, i.e., that measures an SD construct independent of specific content. With this scale, it is possible to measure whether there is a correlation between other instruments and the latent variable of desirability (e.g., Greenblatt et al. 1984). These instruments describe socially desirable but statistically infrequent behaviors. For example, "I don't gossip about other people's business" or "I always obey laws, even if I am unlikely to get caught" (Paulhus, 1991). Then, it is interpreted that, if individuals agree with such statements, they lie because it's virtually impossible to do so. SD scales have been used by many studies across disciplines. Thus, individuals' scores on these scales are interpreted as indicating how positively a person wants to present themselves.

Much research was conducted using SD scales, for example, in public health, medicine, criminology, and politics (e.g., Hebert et al., 1997; Ng et al., 2020; Vecina et al., 2016; Williams et al., 2009). However, this method assumes that the scale only assesses desirability in isolation (orthogonal) to the target construct (i.e., the construct the researcher wants to measure), that is, these scales fail in terms of the discriminant validity of other constructs related to desirability, since desirability is related to other constructs. In addition to the limitations of using an SD instrument, there is an ongoing debate about what these scales measure (e.g., Connelly & Chang, 2016; Tourangeau & Yan, 2007). For example, some argue that they assess socially desirable personality traits (also called "substance"). Connelly and Chang (2016) provided evidence that these instruments contain both method variance (i.e., response style) and trait variance (i.e., substance). Another meta-analysis showed that SD scales do not measure what they intend to measure (i.e., a positive self-representation; Lanz et al., 2021). This claim comes from the fact that SD scale scores have close to zero correlation with prosocial behavior, even in high-stakes settings (i.e., using monetary incentives, and more anonymous vs. less anonymous research; Lanz et al., 2021).

Another limitation is the validity of SD scales as "faking detectors". de Vries and colleagues (2014) and Uziel (2010) argue that scores on the SD scale reflect substantive (socially desirable) traits rather than a general response bias. Tourangeau and Yan (2007) state that the key limitation of these scales is the interpretation of the scores. More specifically, it is impossible to differentiate between a truly honest respondent who is virtuous (e.g., people that don't gossip and always obey the law) and a dishonest respondent who actively lies to present themselves positively.

Meta-analyses have shown that controlling for SD does not increase the predictive validity of scales (Li & Bagger, 2006; Ones et al., 1996). Nonetheless, the studies included in these meta-analyses are based on correlation coefficients and control desirability using partial correlation.

These methods have strong assumptions about the psychometric properties of scale items (Leite & Cooper, 2009). To avoid this, newer and more robust methods can be used.

8.5.2 Ferrando et al. (2009)

Ferrando and colleagues' (2009) model aims to control biases, including social desirability. The model uses the followingequation (summarized here):

$$X_{ij} = {}_{jc}\boldsymbol{\theta}_{ic} + {}_{jd}\boldsymbol{\theta}_{id} + \boldsymbol{\varepsilon}_{ij}$$

where represents the factor loading, θ_{ic} represents the content factor (the construct to be measured), θ_{id} represents the social desirability factor, and $\boldsymbol{\varepsilon}$ represents the residue/ error. Considering that we are going to use a social desirability scale to measure θ_{id}, we have that, for each k number of items on a SD scale, the model reduces to.

$$X_{ij} = {}_{kd}\boldsymbol{\theta}_{id} + \boldsymbol{\varepsilon}_{ij}$$

Using the framework proposed by Ferrando et al. (2009), model adjustment is made using *minimum rank* factor analysis (MRFA; Ten Berge & Kiers, 1991). Following the requirements of this analysis, MRFA minimizes common variance when multiple r-factors are maintained. To estimate the social desirability factor, it is expected that, in a good test, this factor will be weaker than the other construct to be measured. Therefore, to obtain a stable solution with the present method, the authors suggest having at least three markers of social desirability. The first item is used as a proxy for SD, while the remaining items are taken as instrumental variables.

Ferrando and colleagues' (2009) Semi-Restricted Three-Dimensional Factor-Analytic Model aims to control for acquiescence and social desirability. The model proposed by the authors has two strong assumptions, but only one is of interest for desirability. The assumption states that we have at least one item that measures "pure" social desirability (a proxy variable). On the one hand, some researchers believe that there is no a priori reason why desirability should be related to other latent traits (e.g., Edwards, 1967). On the other hand, social desirability is expected to be related to personality traits such as conscientiousness, emotional stability, agreeableness, and socialization (e.g., Connelly & Chang, 2016; Graziano & Tobin, 2002). Therefore, Ferrando's model does not completely solve the problem of a "pure" social desirability scale, as it requires at least one desirability item completely orthogonal to the psychological dimension.

Another limitation of Ferrando et al. (2009) is that the method only applies to unidimensional scales or with multidimensional measurements that approximate an independent cluster structure. Furthermore, there is a tendency for this method to overcontrol (Uziel, 2010), that is, it tends to overestimate the importance of the bias. This is because in self-report instruments it is common to have a general content dimension (for example, the general kindness factor, which brings together several facets). When estimating a general dimension such as bias, part of that general dimension can be attributed to the true variance of the descriptive content (psychological dimension). Thus, the more orthogonal the desirability factor is in relation to the other latent trait, the lower the chance of excess control occurring. Therefore, control methods within its scale can alleviate this limitation.

8.5.3 Leite and Cooper (2010)

This model uses factor mixture models as an extension of Ferrando's (2005) method for detecting SD bias. Their method has two contrasting hypotheses: the null hypothesis states that the SD bias factor is not related to subjects' responses on the content scale; the other states that the SD bias factor predicts responses to the focal items for all respondents. Regarding the extent of their work, Leite & Cooper's (2010) method is an intermediate hypothesis: for some individuals in the sample (but not for all individuals), the SD bias factor predicts responses to the items of the content instrument. Although this method can differentiate between SD responders and nonresponders, it still does not solve the problem of a 'pure' SD instrument, because there is still a need to use SD instruments that have 'pure' social desirability items.

8.5.4 Ziegler and Buehner (2009)

The authors conclude that *faking* can be understood as a systematic measurement error, resulting from the interaction between context and person. If *faking* were seen as this interaction, it would then be systematic variance (Ziegler & Buehner, 2009). Thus, spurious measurement errors are systematic because it is assumed that this error does not always occur, but always occurs in identical circumstances.

Ziegler & Buehner (2009) propose a new way to separate trait variance from *faking*, stating that it is a method to control SD. The logic behind this modeling is that spurious measurement error (i.e., *faking*) contributes to correlations between instruments, however, not between scales that measure a latent trait, but between scales that contain *faking*. Thus, a systematic measurement error can be viewed as common method variance (CMV; Podsakoff et al., 2003), which is modeled as a latent variable using structural equation modeling.

The proposed method works as follows. The questionnaire is administered twice to two groups, and spurious measurement errors must occur at both measurement points in both groups if SD always occurs to some extent. The two groups are separated as 1) a control group, which is asked to answer honestly both times (low stakes); 2) an experimental group, which receives a specific forgery instruction for the second time (high stakes). At the first measurement point, both common factors of the method must have the same character. However, at Time 2, the character of the common method factor in the experimental group should have changed due to the specific *faking* instruction.

One of the limitations of using CMV can be interpreted by Podsakoff et al. (2003), who alerted CMV users, as this latent factor of the method could comprise several things, with *faking* being just one of the explanations. Furthermore, this method assumes that all respondents in a specific condition are *faking* their answers. However, we have no evidence to assume that virtuous people will *fake* their answers in specific scenarios, or that "fakers" will *fake* every time. Another limitation is related to resource constraints, researchers must spend more time, money and other resources collecting more data and ensuring that participants answer the questionnaire twice. The third limitation is stated by the authors:

> To use the presented structural equation model to extract *faking* from variance, at least two different characteristics must be faked by participants. Otherwise, the spurious measurement error variance and the trait variance could not be separated, because the trait and the method factor would attempt to explain the same variance.

8.5.5 Peabody quadruplets (1967)

To gauge Social Desirability (SD) while preserving the integrity of item content, one approach to managing social desirability is by meticulously crafting items within the scale itself (e.g., Peabody, 1967; Pettersson et al., 2012). This involves dividing items into two distinct components: one addressing the construct under evaluation (descriptive content), and the other focusing on the social desirability of the behavior described by the item (evaluative content). Through this method, balanced quadruplets can be constructed to encompass both the lower and upper extremes of a given characteristic. Consequently, quadruplets offer the additional advantage of mitigating another form of bias known as acquiescence bias – the tendency to endorse items irrespective of their content polarity, whether positive or negative (Mirowsky & Ross, 1991). An illustrative example of a quadruple is presented in the Table 8.1.

Table 8.1: Hypothetical Descriptors to Assess Extraversion in Quadruplets

	Low Desirability	High Desirability
Low Descriptive	Withdrawn	Introspective
High Descriptive	Chatty	Communicative

Peabody's (1967) approach exclusively relies on quadruplets to assess social desirability. However, this method can pose operational challenges. Firstly, the content of the items within the quadruplets may not always lend itself to manipulation or may not naturally fit into a question format. Secondly, employing quadruplets necessitates a substantial number of items to evaluate the same content, which may not contribute additional insights into the construct while potentially inducing respondent fatigue. As a result, one alternative is to estimate social desirability within the quadruplets and utilize this estimation to regulate desirability outside the quadruplets, such as through the employment of multiple indicators multiple causes (MIMIC) modeling techniques.

From the creation of quadruplets it is possible to extract a general factor of social desirability, and then separate the bias from the other factors that we wish to measure (Pettersson et al., 2014; Pettersson et al., 2012; Saucier et al., 2001). The quadruple model can be represented by the following matrix equation:

$$\begin{bmatrix} x1 \\ x2 \\ x3 \\ x4 \end{bmatrix} = \begin{bmatrix} -\boldsymbol{\lambda}_{1c} & +\boldsymbol{\lambda}_{1d} \\ -\boldsymbol{\lambda}_{2c} & -\boldsymbol{\lambda}_{2d} \\ +\boldsymbol{\lambda}_{3c} & +\boldsymbol{\lambda}_{3d} \\ +\boldsymbol{\lambda}_{4c} & -\boldsymbol{\lambda}_{4d} \end{bmatrix} \begin{bmatrix} \eta_c \\ \eta_d \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \boldsymbol{\varepsilon}_3 \\ \boldsymbol{\varepsilon}_4 \end{bmatrix}$$

Where x_n is the observed response for item n within the quadruple; $\boldsymbol{\lambda}_{nc}$ is the factor loading of item n in content dimension c ; $\boldsymbol{\lambda}_{nd}$ is the factor loading of item n on the desirability dimension d; η represents the constructs; and $\boldsymbol{\varepsilon}$ the measurement errors.

Bastos and Valentini (2023) have run two simulation studies in order to see if controlling the social desirability using the MIMIC model recovers the MIMIC regressions from the social desirability factors to items outside of the quadruplets manipulations. The first simulation showed that, under certain conditions, the MIMIC-Quadruplets model for Likert-type recovered the SD regressions to extra items. In addition, in the MIMIC-Quadruplets model for forced-choice, all conditions simulated in this study recovered (based on bias and coverage indicators) the regressions from social desirability to extra items.

For this, I have developed two intuitive shiny apps where researchers can input their model and see if there's enough power, low bias, and high coverage to estimate the social desirability of items outside of the quadruplets. One of the apps (called quadSimple; https://peabody-mimic.shinyapps.io/quadSimple/) is more user-friendly and requires little information regarding the instrument. The quadSimple is recommended to be used before the construction of an instrument, to give light to the required number of quadruplets they need to build. The other app (called quadSim; https://peabody-mimic.shinyapps.io/quadSim/) is more precise and requires more information about the instrument. The quadSim is recommended for scales where researchers already have information on the model parameters.

8.6 How to Control Social Desirability in R

8.6.1 Controlling Desirability with Ferrando et al. (2009)

To run with the analysis by Ferrando et al. (2009), we first have to install the *vampyr* (Navarro-Gonzalez et al., 2021) package to run the analyses.

```r
install.packages("vampyr")
```

And tell the program that we are going to use the functions of these packages.

```r
library(vampyr)
```

To run the analyses, we will use a database from the package itself. Let's see what the dataset looks like.

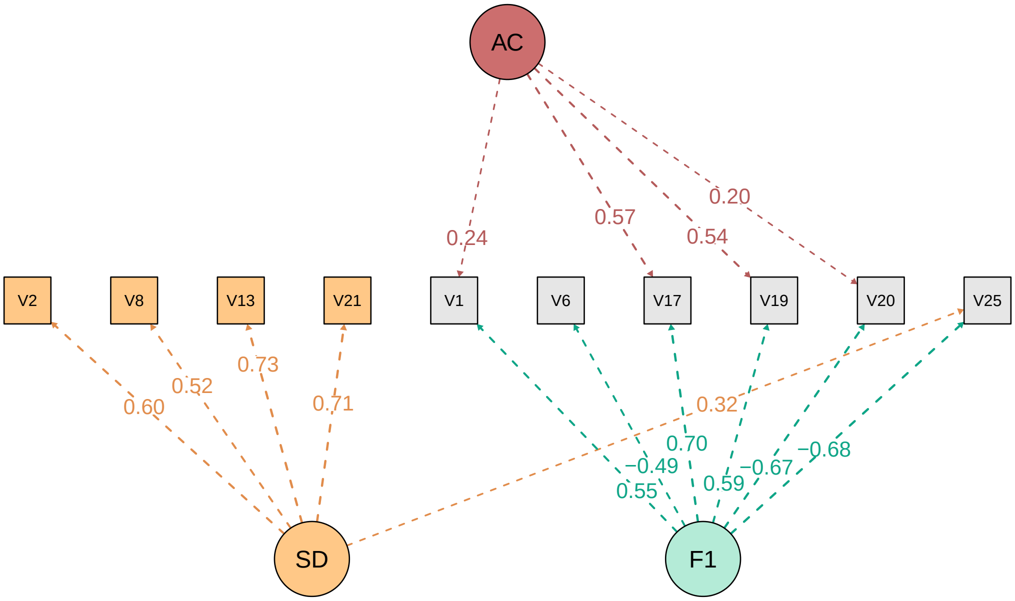
```r
summary(vampyr::vampyr_example)
```

```
      V2              V8             V13             V21
Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
1st Qu.:3.000   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:2.000
Median :4.000   Median :4.000   Median :2.000   Median :3.000
Mean   :3.667   Mean   :3.263   Mean   :2.317   Mean   :2.947
3rd Qu.:5.000   3rd Qu.:4.000   3rd Qu.:3.000   3rd Qu.:4.000
Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000
      V1              V6             V17            V19             V20
Min.   :1.000   Min.   :1.000   Min.   :1.00   Min.   :1.000   Min.   :1.000
1st Qu.:3.000   1st Qu.:1.000   1st Qu.:3.00   1st Qu.:3.000   1st Qu.:1.000
Median :4.000   Median :2.000   Median :4.00   Median :4.000   Median :2.000
Mean   :3.643   Mean   :2.467   Mean   :3.71   Mean   :3.493   Mean   :1.997
3rd Qu.:5.000   3rd Qu.:3.000   3rd Qu.:5.00   3rd Qu.:5.000   3rd Qu.:3.000
Max.   :5.000   Max.   :5.000   Max.   :5.00   Max.   :5.000   Max.   :5.000
     V25
Min.   :1.000
1st Qu.:1.000
Median :1.000
Mean   :1.687
3rd Qu.:2.000
Max.   :5.000
```

According to the package, we have a dataset with 300 observations and 10 variables, where 6 items measure physical aggression and we have 4 markers of social desirability. Items 1, 2, 3, and 4 are markers of SD ("pure" measures of SD), and the remaining 6 items measure physical aggression. Items 5, 7 and 8 are in the positive pole of the target construct and items 6, 9 and 10 are written in the negative pole of the target construct.

To perform the analysis controlling both desirability and acquiescence, simply use the following code.

```r
res <-  ControlResponseBias(vampyr_example,
                            content_factors = 1,
                            SD_items = c(1,2,3,4),
                            corr = "Polychoric",
                            contAC = TRUE,
                            rotat = "promin",
                            PA = FALSE,
                            factor_scores = FALSE,
                            path = TRUE)
```



```
DETAILS OF ANALYSIS

Number of participants                      :   300
Number of items                             :    10
Items selected as SD items                  :  1, 2, 3, 4
Items selected as unbalanced                :  0
Dispersion Matrix                           : Polychoric Correlations
Method for factor extraction                : Unweighted Least Squares (ULS)
Rotation Method                             : none
```

```
-----------------------------------------------------------------------

Univariate item descriptives

Item       Mean        Variance    Skewness     Kurtosis (Zero centered)

Item   1   3.667       1.260      -0.555       -0.566
Item   2   3.263       1.760      -0.379       -1.005
Item   3   2.317       1.695       0.601       -0.880
Item   4   2.947       1.924      -0.033       -1.284
Item   5   3.643       1.374      -0.565       -0.535
Item   6   2.467       1.802       0.487       -0.967
Item   7   3.710       1.678      -0.652       -0.716
Item   8   3.493       1.629      -0.411       -0.862
Item   9   1.997       1.515       1.041       -0.011
Item  10   1.687       0.925       1.293        0.838

Polychoric correlation is advised when the univariate distributions of ordinal items are
asymmetric or with excess of kurtosis. If both indices are lower than one in absolute value,
then Pearson correlation is advised. You can read more about this subject in:

Muthen, B., & Kaplan D. (1985). A Comparison of Some Methodologies for the Factor Analysis o
Non-Normal Likert Variables. British Journal of Mathematical and Statistical Psychology, 38,

Muthen, B., & Kaplan D. (1992). A Comparison of Some Methodologies for the Factor Analysis o
Non-Normal Likert Variables: A Note on the Size of the Model. British Journal of Mathematical
and Statistical Psychology, 45, 19-30.

-----------------------------------------------------------------------

Adequacy of the dispersion matrix

Determinant of the matrix     = 0.047816437916934
Bartlett's statistic          =   896.4 (df =    45; P = 0.000000)
Kaiser-Meyer-Olkin (KMO) test = 0.76664 (fair)

-----------------------------------------------------------------------
EXPLORATORY FACTOR ANALYSIS CONTROLLING SOCIAL DESIRABILITY AND ACQUIESCENCE
-----------------------------------------------------------------------

Robust Goodness of Fit statistics

          Root Mean Square Error of Approximation (RMSEA) = 0.032
```

```
 Robust Mean-Scaled Chi Square with 23 degrees of freedom = 30.146

              Non-Normed Fit Index (NNFI; Tucker & Lewis) = 0.989
                              Comparative Fit Index (CFI) = 0.994
                              Goodness of Fit Index (GFI) = 0.977

-----------------------------------------------------------------------

                  Root Mean Square Residuals (RMSR) = 0.0452
Expected mean value of RMSR for an acceptable model = 0.0578 (Kelley's criterion)

-----------------------------------------------------------------------

Unrotated loading matrix

         Factor SD Factor AC Factor 1
Item   1   0.60257   0.00000  0.00000
Item   2   0.51525   0.00000  0.00000
Item   3   0.72710   0.00000  0.00000
Item   4   0.71130   0.00000  0.00000
Item   5  -0.07851   0.23763  0.54830
Item   6   0.27520   0.00221 -0.49052
Item   7  -0.16413   0.57414  0.70138
Item   8  -0.14320   0.54065  0.59105
Item   9   0.26560   0.19600 -0.66800
Item  10   0.31732   0.06249 -0.68222
```

This analysis allows controlling the effects of two response biases: Social Desirability and Acquiescence, extracting the variance due to these factors before extracting the content variance. If you don't have or want to control acquiescence, simply change the argument `contAC = TRUE` to `contAC = FALSE` .

We see that Bartlett's test of sphericity and KMO were calculated before proceeding with Exploratory Factor Analysis. Furthermore, the model fit indices were calculated. We also see that items 6, 9 and 10 have even high loadings on the desirability factor ("Factor SD"), and items 5, 7 and 8 on the acquiescence factor ("Factor AC").

The cool thing is that it allows you to calculate people's factor scores. Factor scores work like when you calculate the mean score of an instrument to correlate with others, but calculating mean scores has certain assumptions, while factor scores have others. So, to calculate the factor scores while controlling the SD and acquiescence biases, simply leave the factor scores argument as `TRUE` (`factor_scores = TRUE)` and save the result in some variable. In our case, we save the results in the `res` variable.

To save only the factor scores, simply extract the scores from the list.

```
factor_scores <- res$Factor_scores
```

This way, just put this column of factor scores together with your data (using `cbind()`) and then calculate whatever analysis you want.

8.6.2 Controlling with MIMIC and Quadruples (Bastos & Valentini, 2023)

To run a MIMIC with Quadruples, we first have to install the *lavaan* (Rosseel, 2012) package to run the analyzes and for database simulation and *semplot* (Epskamp, 2022) package for visualization.

```
install.packages("lavaan")
install.packages("semPlot")
```

And tell the program that we are going to use the functions of these packages.

```
library(lavaan)
library(semPlot)
```

Let's simulate the data with quadruplets for us to use.

```
#Quadruple Factor Loadings on Social Desirability
FactorLoadingsSDQ<- rep(0.3, 16)*c(1,-1,1,-1)

#Quadruple Factor Loadings on the Target Construct
RandomFactorLoadingsQ<-rep(0.7, 16)*c(-1,-1,1,1)

# Factor Loads of the extra item in the Target Construct
set.seed(2021)
RandomFactorLoadings <- round(runif((10), min = .3, max = .8), 3)

# Desirability Regressions for Target Construct items
set.seed(2021)
RandomSDregression <- round(runif((10), min = .1, max = .5), 3)

# Item Thresholds
set.seed(2020)
thld1Vet<-round(runif(26, min=-2, max=.5),3)
thld2Vet<-round(thld1Vet +.5,3)
```

```
thld3Vet<-round(thld1Vet + 1,3)
thld4Vet<-round(thld1Vet + 1.5,3)

# Simulated Model
simModel <- paste0("fator1 =~",RandomFactorLoadings[1],"*it1 +",
                         RandomFactorLoadings[2],"*it2 +",
                         RandomFactorLoadings[3],"*it3 +",
                         RandomFactorLoadings[4],"*it4 +",
                         RandomFactorLoadings[5],"*it5 +",
                         RandomFactorLoadingsQ[1],"*sd1 +",
                         RandomFactorLoadingsQ[2],"*sd2 +",
                         RandomFactorLoadingsQ[3],"*sd3 +",
                         RandomFactorLoadingsQ[4],"*sd4 +",
                         RandomFactorLoadingsQ[5],"*sd5 +",
                         RandomFactorLoadingsQ[6],"*sd6 +",
                         RandomFactorLoadingsQ[7],"*sd7 +",
                         RandomFactorLoadingsQ[8],"*sd8\n",

                         "fator2 =~", RandomFactorLoadingsQ[6],"*it6 +",
                         RandomFactorLoadingsQ[7],"*it7 +",
                         RandomFactorLoadingsQ[8],"*it8 +",
                         RandomFactorLoadingsQ[9],"*it9 +",
                         RandomFactorLoadingsQ[10],"*it10 +",
                         RandomFactorLoadingsQ[9],"*sd9 +",
                         RandomFactorLoadingsQ[10],"*sd10 +",
                         RandomFactorLoadingsQ[11],"*sd11 +",
                         RandomFactorLoadingsQ[12],"*sd12 +",
                         RandomFactorLoadingsQ[13],"*sd13 +",
                         RandomFactorLoadingsQ[14],"*sd14 +",
                         RandomFactorLoadingsQ[15],"*sd15 +",
                         RandomFactorLoadingsQ[16],"*sd16\n",

                         "SD =~", FactorLoadingsSDQ[1], "*sd1 +",
                         FactorLoadingsSDQ[2],"*sd2 +",
                         FactorLoadingsSDQ[3],"*sd3 +",
                         FactorLoadingsSDQ[4],"*sd4 +",
                         FactorLoadingsSDQ[5], "*sd5 +",
                         FactorLoadingsSDQ[6],"*sd6 +",
                         FactorLoadingsSDQ[7],"*sd7 +",
                         FactorLoadingsSDQ[8],"*sd8 +",
                         FactorLoadingsSDQ[9], "*sd9 +",
                         FactorLoadingsSDQ[10],"*sd10 +",
                         FactorLoadingsSDQ[11],"*sd11 +",
                         FactorLoadingsSDQ[12],"*sd12 +",
                         FactorLoadingsSDQ[13], "*sd13 +",
                         FactorLoadingsSDQ[14],"*sd14 +",
                         FactorLoadingsSDQ[15],"*sd15 +",
                         FactorLoadingsSDQ[16],"*sd16\n",

                         "SD ~~ 1*SD\n",
                         "fator1 ~~ 1*fator1\n",
                         "fator2 ~~ 1*fator2\n",
                         "fator1 ~~ 0*SD\n",
                         "fator2 ~~ 0*SD\n",
                         "fator1 ~~ .3*fator2\n",

                         "it1 ~",RandomSDregression[1],"*SD\n",
                         "it2 ~",RandomSDregression[2],"*SD\n",
                         "it3 ~",RandomSDregression[3],"*SD\n",
                         "it4 ~",RandomSDregression[4],"*SD\n",
                         "it5 ~",RandomSDregression[5],"*SD\n",
                         "it6 ~",RandomSDregression[6],"*SD\n",
                         "it7 ~",RandomSDregression[7],"*SD\n",
                         "it8 ~",RandomSDregression[8],"*SD\n",
                         "it9 ~",RandomSDregression[9],"*SD\n",
                         "it10 ~",RandomSDregression[10],"*SD\n",

                         "sd1 |",thld1Vet[1],"*t1 +", thld2Vet[1], "*t2 +",
                         thld3Vet[1],"*t3 +",thld4Vet[1],"*t4\n",
                         "sd2 |",thld1Vet[2],"*t1 +", thld2Vet[2], "*t2 +",
                         thld3Vet[2],"*t3 +",thld4Vet[2],"*t4\n",
                        "sd3 |",thld1Vet[3],"*t1 +", thld2Vet[3], "*t2 +",
                        thld3Vet[3],"*t3 +",thld4Vet[3],"*t4\n",
                        "sd4 |",thld1Vet[4],"*t1 +", thld2Vet[4], "*t2 +",
                        thld3Vet[4],"*t3 +",thld4Vet[4],"*t4\n",
                        "it1 |",thld1Vet[5],"*t1 +", thld2Vet[5], "*t2 +",
                        thld3Vet[5],"*t3 +",thld4Vet[5],"*t4\n",
                        "it2 |",thld1Vet[6],"*t1 +", thld2Vet[6], "*t2 +",
                        thld3Vet[6],"*t3 +",thld4Vet[6],"*t4\n",
                        "it3 |",thld1Vet[7],"*t1 +", thld2Vet[7], "*t2 +",
                        thld3Vet[7],"*t3 +",thld4Vet[7],"*t4\n",
                        "it4 |",thld1Vet[8],"*t1 +", thld2Vet[8], "*t2 +",
                        thld3Vet[8],"*t3 +",thld4Vet[8],"*t4\n",
                        "it5 |",thld1Vet[9],"*t1 +", thld2Vet[9], "*t2 +",
                        thld3Vet[9],"*t3 +",thld4Vet[9],"*t4\n",
                        "it6 |",thld1Vet[10],"*t1 +", thld2Vet[10], "*t2 +",
                        thld3Vet[10],"*t3 +",thld4Vet[10],"*t4\n",
                        "it7 |",thld1Vet[11],"*t1 +", thld2Vet[11], "*t2 +",
                        thld3Vet[11],"*t3 +",thld4Vet[11],"*t4\n",
                        "it8 |",thld1Vet[12],"*t1 +", thld2Vet[12], "*t2 +",
                        thld3Vet[12],"*t3 +",thld4Vet[12],"*t4\n",
                        "it9 |",thld1Vet[13],"*t1 +", thld2Vet[13], "*t2 +",
                        thld3Vet[13],"*t3 +",thld4Vet[13],"*t4\n",
                        "it10 |",thld1Vet[14],"*t1 +", thld2Vet[14], "*t2 +",
                        thld3Vet[14],"*t3 +",thld4Vet[14],"*t4\n",
                        "sd5 |",thld1Vet[15],"*t1 +", thld2Vet[15], "*t2 +",
                        thld3Vet[15],"*t3 +",thld4Vet[15],"*t4\n",
                        "sd6 |",thld1Vet[16],"*t1 +", thld2Vet[16], "*t2 +",
                        thld3Vet[16],"*t3 +",thld4Vet[16],"*t4\n",
                        "sd7 |",thld1Vet[17],"*t1 +", thld2Vet[17], "*t2 +",
                        thld3Vet[17],"*t3 +",thld4Vet[17],"*t4\n",
                        "sd8 |",thld1Vet[18],"*t1 +", thld2Vet[18], "*t2 +",
                        thld3Vet[18],"*t3 +",thld4Vet[18],"*t4\n",
                        "sd9 |",thld1Vet[19],"*t1 +", thld2Vet[19], "*t2 +",
                        thld3Vet[19],"*t3 +",thld4Vet[19],"*t4\n",
                        "sd10 |",thld1Vet[20],"*t1 +", thld2Vet[20], "*t2 +",
                        thld3Vet[20],"*t3 +",thld4Vet[20],"*t4\n",
                        "sd11 |",thld1Vet[21],"*t1 +", thld2Vet[21], "*t2 +",
                        thld3Vet[21],"*t3 +",thld4Vet[21],"*t4\n",
                        "sd12 |",thld1Vet[22],"*t1 +", thld2Vet[22], "*t2 +",
                        thld3Vet[22],"*t3 +",thld4Vet[22],"*t4\n",
                        "sd13 |",thld1Vet[23],"*t1 +", thld2Vet[23], "*t2 +",
                        thld3Vet[23],"*t3 +",thld4Vet[23],"*t4\n",
                        "sd14 |",thld1Vet[24],"*t1 +", thld2Vet[24], "*t2 +",
                        thld3Vet[24],"*t3 +",thld4Vet[24],"*t4\n",
                        "sd15 |",thld1Vet[25],"*t1 +", thld2Vet[25], "*t2 +",
                        thld3Vet[25],"*t3 +",thld4Vet[25],"*t4\n",
                        "sd16 |",thld1Vet[26],"*t1 +", thld2Vet[26], "*t2 +",
                        thld3Vet[26],"*t3 +",thld4Vet[26],"*t4")

#Simulating the Data
simulatedData <- lavaan::simulateData(model = simModel,
                                       model.type = "sem",
                                       sample.nobs = 4000,
                                       seed = 2024,
                                       return.type = "data.frame",
                                       standardized = TRUE
                                       )
```

In the simulated data, we have items from it1 to it10 (which are items that are not made in quadruple format), items sd1 to sd16 (which are items in quadruple format), and are in the 5-point Likert format. See a summary of the items below.

```
summary(simulatedData)
```

```
     it1             it2             it3             it4
Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
1st Qu.:3.000   1st Qu.:4.000   1st Qu.:4.000   1st Qu.:2.000
Median :5.000   Median :5.000   Median :5.000   Median :4.000
Mean   :4.146   Mean   :4.311   Mean   :4.163   Mean   :3.381
3rd Qu.:5.000   3rd Qu.:5.000   3rd Qu.:5.000   3rd Qu.:5.000
Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000
     it5             sd1            sd2             sd3             sd4
Min.   :1.000   Min.   :1.00   Min.   :1.000   Min.   :1.000   Min.   :1.000
1st Qu.:4.000   1st Qu.:1.00   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:2.000
Median :5.000   Median :2.00   Median :4.000   Median :2.000   Median :3.000
Mean   :4.445   Mean   :2.52   Mean   :3.353   Mean   :2.585   Mean   :3.092
3rd Qu.:5.000   3rd Qu.:4.00   3rd Qu.:5.000   3rd Qu.:4.000   3rd Qu.:4.000
Max.   :5.000   Max.   :5.00   Max.   :5.000   Max.   :5.000   Max.   :5.000
     sd5             sd6             sd7             sd8
Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.000
Median :3.000   Median :3.000   Median :1.000   Median :2.000
Mean   :3.314   Mean   :2.846   Mean   :1.655   Mean   :2.485
3rd Qu.:5.000   3rd Qu.:4.000   3rd Qu.:2.000   3rd Qu.:4.000
Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000
     it6             it7             it8             it9
Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.000
Median :2.000   Median :2.000   Median :2.000   Median :1.000
Mean   :2.623   Mean   :2.139   Mean   :2.186   Mean   :1.983
3rd Qu.:4.000   3rd Qu.:3.000   3rd Qu.:3.000   3rd Qu.:3.000
Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000
     it10            sd9             sd10            sd11
Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
```

```
 1st Qu.:2.000   1st Qu.:1.000   1st Qu.:3.000   1st Qu.:3.000  
 Median :3.000   Median :3.000   Median :4.000   Median :4.000  
 Mean   :3.297   Mean   :2.853   Mean   :3.792   Mean   :3.942  
 3rd Qu.:5.000   3rd Qu.:4.000   3rd Qu.:5.000   3rd Qu.:5.000  
 Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000  
      sd12            sd13            sd14            sd15      
 Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000  
 1st Qu.:4.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.000  
 Median :5.000   Median :1.000   Median :1.000   Median :1.000  
 Mean   :4.264   Mean   :1.994   Mean   :1.686   Mean   :1.814  
 3rd Qu.:5.000   3rd Qu.:3.000   3rd Qu.:2.000   3rd Qu.:2.000  
 Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000  
      sd16      
 Min.   :1.000  
 1st Qu.:3.000  
 Median :5.000  
 Mean   :4.019  
 3rd Qu.:5.000  
 Max.   :5.000  
```

You will get to understand the model better now, when we configure it. We have 26 items, 4 of which are quadruples (16 items), and 10 items outside the quadruples to try to control social desirability. Let's configure the model the way you would with your database, that is, we will place all items (quadruples or not) of a given factor to estimate that factor. For example, items it1 to it4 and items sd1 to sd8 are Factor 1 items, so we will estimate Factor 1 with these items. The same logic applies to Factor 2. To estimate desirability, we will only use the quadruples, given that only in the quadruples we manipulated the items to have desirability. We also have to maintain the content factors (Factor 1 and Factor 2) with a correlation equal to 0 with the desirability factor. This is a necessary step to be able to carry out the calculation, otherwise we will have to estimate more parameters than we have information about. Finally, we will perform a desirability regression for the items that were not manipulated in quadruples, to control for the desirability of these extra items.

```
empiricalModel <- "
                factor1 =~ NA*it1 + it2 + it3 + it4 + it5 + sd1 + sd2 + sd3 +
                sd4 + sd5 + sd6 + sd7 + sd8

                factor2 =~ NA*it6 + it7 + it8 + it9 + it10 + sd9 + sd10 + sd11 +
                sd12 + sd13 + sd14 + sd15 + sd16

                SD =~ NA*sd1 + sd2 + sd3 + sd4 + sd5 + sd6 + sd7 + sd8 + sd9 +
                sd10 + sd11 + sd12 + sd13 + sd14 + sd15 + sd16
```

```
              SD ~~ 1*SD
              factor1 ~~ 1*factor1
              factor2 ~~ 1*factor2

              factor1 ~~ 0*SD
              factor2 ~~ 0*SD
              factor1 ~~ factor2

              it1 ~ SD
              it2 ~ SD
              it3 ~ SD
              it4 ~ SD
              it5 ~ SD
              it6 ~ SD
              it7 ~ SD
              it8 ~ SD
              it9 ~ SD
              it10 ~SD"
```

E agora, rodaremos a análise da seguinte forma. Como temos itens ordinais, falamos para o programa que os itens são ordinais e usamos o estimador "WLSMV".

```
sem.fit <- sem(model = empiricalModel,
               data = simulatedData,
               estimator = "WLSMV",
               ordered = TRUE
               )

summary(sem.fit,
        standardized=TRUE,
        fit.measures = TRUE
        )
```

```
lavaan 0.6.17 ended normally after 39 iterations

  Estimator                                       DWLS
  Optimization method                           NLMINB
  Number of model parameters                       157

  Number of observations                          4000
```

```
Model Test User Model:
                                              Standard      Scaled
  Test Statistic                               134.500     262.032
  Degrees of freedom                               272         272
  P-value (Chi-square)                           1.000       0.657
  Scaling correction factor                                  0.828
  Shift parameter                                           99.666
    simple second-order correction

Model Test Baseline Model:

  Test statistic                            182093.691   68350.757
  Degrees of freedom                               325         325
  P-value                                        0.000       0.000
  Scaling correction factor                                  2.672

User Model versus Baseline Model:

  Comparative Fit Index (CFI)                    1.000       1.000
  Tucker-Lewis Index (TLI)                       1.001       1.000

  Robust Comparative Fit Index (CFI)                         1.000
  Robust Tucker-Lewis Index (TLI)                            1.000

Root Mean Square Error of Approximation:

  RMSEA                                          0.000       0.000
  90 Percent confidence interval - lower         0.000       0.000
  90 Percent confidence interval - upper         0.000       0.005
  P-value H_0: RMSEA <= 0.050                    1.000       1.000
  P-value H_0: RMSEA >= 0.080                    0.000       0.000

  Robust RMSEA                                               0.001
  90 Percent confidence interval - lower                     0.000
  90 Percent confidence interval - upper                     0.010
  P-value H_0: Robust RMSEA <= 0.050                         1.000
  P-value H_0: Robust RMSEA >= 0.080                         0.000

Standardized Root Mean Square Residual:

  SRMR                                           0.011       0.011

Parameter Estimates:
  Parameterization                               Delta
  Standard errors                           Robust.sem
  Information                                 Expected
  Information saturated (h1) model        Unstructured

Latent Variables:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  factor1 =~
    it1               0.533    0.015   34.800    0.000    0.533    0.533
    it2               0.713    0.013   54.803    0.000    0.713    0.713
    it3               0.657    0.013   49.768    0.000    0.657    0.657
    it4               0.468    0.015   31.011    0.000    0.468    0.468
    it5               0.631    0.015   42.576    0.000    0.631    0.631
    sd1              -0.700    0.011  -63.228    0.000   -0.700   -0.700
    sd2              -0.710    0.011  -66.102    0.000   -0.710   -0.710
    sd3               0.701    0.011   64.861    0.000    0.701    0.701
    sd4               0.694    0.011   62.012    0.000    0.694    0.694
    sd5              -0.687    0.011  -60.382    0.000   -0.687   -0.687
    sd6              -0.710    0.011  -66.615    0.000   -0.710   -0.710
    sd7               0.690    0.013   52.042    0.000    0.690    0.690
    sd8               0.683    0.012   58.091    0.000    0.683    0.683
  factor2 =~
    it6               0.705    0.011   64.358    0.000    0.705    0.705
    it7              -0.692    0.012  -57.695    0.000   -0.692   -0.692
    it8              -0.701    0.011  -62.954    0.000   -0.701   -0.701
    it9               0.690    0.012   55.503    0.000    0.690    0.690
    it10              0.702    0.012   59.851    0.000    0.702    0.702
    sd9               0.673    0.011   59.095    0.000    0.673    0.673
    sd10              0.698    0.011   61.611    0.000    0.698    0.698
    sd11             -0.719    0.012  -62.245    0.000   -0.719   -0.719
    sd12             -0.687    0.013  -54.227    0.000   -0.687   -0.687
    sd13              0.707    0.011   61.651    0.000    0.707    0.707
    sd14              0.703    0.013   55.083    0.000    0.703    0.703
    sd15             -0.692    0.013  -54.861    0.000   -0.692   -0.692
    sd16             -0.687    0.012  -58.792    0.000   -0.687   -0.687
  SD =~
    sd1               0.315    0.019   16.764    0.000    0.315    0.315
    sd2              -0.261    0.019  -13.479    0.000   -0.261   -0.261
    sd3               0.293    0.019   15.763    0.000    0.293    0.293
    sd4              -0.307    0.019  -16.490    0.000   -0.307   -0.307
    sd5               0.305    0.019   15.992    0.000    0.305    0.305
    sd6              -0.318    0.018  -17.535    0.000   -0.318   -0.318
    sd7               0.311    0.021   14.856    0.000    0.311    0.311
    sd8              -0.308    0.019  -16.326    0.000   -0.308   -0.308
    sd9               0.358    0.018   19.441    0.000    0.358    0.358
    sd10             -0.292    0.020  -14.596    0.000   -0.292   -0.292
    sd11              0.309    0.020   15.431    0.000    0.309    0.309
    sd12             -0.326    0.020  -16.032    0.000   -0.326   -0.326
    sd13              0.292    0.020   14.543    0.000    0.292    0.292
    sd14             -0.300    0.021  -14.325    0.000   -0.300   -0.300
    sd15              0.308    0.021   14.940    0.000    0.308    0.308
    sd16             -0.297    0.020  -14.918    0.000   -0.297   -0.297

Regressions:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  it1 ~
    SD                0.280    0.020   14.105    0.000    0.280    0.280
  it2 ~
    SD                0.413    0.020   20.933    0.000    0.413    0.413
  it3 ~
    SD                0.369    0.020   18.801    0.000    0.369    0.369
  it4 ~
    SD                0.280    0.019   15.057    0.000    0.280    0.280
  it5 ~
    SD                0.366    0.020   18.017    0.000    0.366    0.366
  it6 ~
    SD                0.400    0.018   21.915    0.000    0.400    0.400
  it7 ~
    SD                0.382    0.019   19.789    0.000    0.382    0.382
  it8 ~
    SD                0.215    0.020   10.530    0.000    0.215    0.215
  it9 ~
    SD                0.424    0.019   22.454    0.000    0.424    0.424
  it10 ~
    SD                0.496    0.017   29.103    0.000    0.496    0.496

Covariances:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  factor1 ~~
    SD                0.000                               0.000    0.000
  factor2 ~~
    SD                0.000                               0.000    0.000
  factor1 ~~
    factor2          -0.290    0.017  -17.440    0.000   -0.290   -0.290
```

```
Thresholds:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
    it1|t1           -1.657    0.034  -49.184    0.000   -1.657   -1.657
    it1|t2           -1.146    0.025  -45.186    0.000   -1.146   -1.146
    it1|t3           -0.670    0.022  -31.120    0.000   -0.670   -0.670
    it1|t4           -0.182    0.020   -9.133    0.000   -0.182   -0.182
    it2|t1           -1.793    0.037  -48.349    0.000   -1.793   -1.793
    it2|t2           -1.332    0.028  -48.013    0.000   -1.332   -1.332
    it2|t3           -0.834    0.023  -36.989    0.000   -0.834   -0.834
    it2|t4           -0.361    0.020  -17.793    0.000   -0.361   -0.361
    it3|t1           -1.680    0.034  -49.095    0.000   -1.680   -1.680
    it3|t2           -1.161    0.026  -45.489    0.000   -1.161   -1.161
    it3|t3           -0.700    0.022  -32.270    0.000   -0.700   -0.700
    it3|t4           -0.188    0.020   -9.417    0.000   -0.188   -0.188
    it4|t1           -1.012    0.024  -42.192    0.000   -1.012   -1.012
    it4|t2           -0.537    0.021  -25.718    0.000   -0.537   -0.537
    it4|t3           -0.029    0.020   -1.486    0.137   -0.029   -0.029
    it4|t4            0.468    0.021   22.674    0.000    0.468    0.468
    it5|t1           -1.991    0.043  -45.939    0.000   -1.991   -1.991
    it5|t2           -1.504    0.031  -49.218    0.000   -1.504   -1.504
    it5|t3           -1.006    0.024  -42.032    0.000   -1.006   -1.006
    it5|t4           -0.501    0.021  -24.136    0.000   -0.501   -0.501
    sd1|t1           -0.393    0.020  -19.267    0.000   -0.393   -0.393
    sd1|t2            0.106    0.020    5.343    0.000    0.106    0.106
    sd1|t3            0.608    0.021   28.678    0.000    0.608    0.608
    sd1|t4            1.088    0.025   43.998    0.000    1.088    1.088
    sd2|t1           -1.004    0.024  -41.979    0.000   -1.004   -1.004
    sd2|t2           -0.482    0.021  -23.297    0.000   -0.482   -0.482
    sd2|t3           -0.024    0.020   -1.233    0.218   -0.024   -0.024
    sd2|t4            0.478    0.021   23.110    0.000    0.478    0.478
    sd3|t1           -0.436    0.021  -21.270    0.000   -0.436   -0.436
    sd3|t2            0.051    0.020    2.561    0.010    0.051    0.051
    sd3|t3            0.550    0.021   26.244    0.000    0.550    0.550
    sd3|t4            1.057    0.024   43.289    0.000    1.057    1.057
    sd4|t1           -0.812    0.022  -36.260    0.000   -0.812   -0.812
    sd4|t2           -0.331    0.020  -16.380    0.000   -0.331   -0.331
    sd4|t3            0.174    0.020    8.754    0.000    0.174    0.174
    sd4|t4            0.705    0.022   32.451    0.000    0.705    0.705
    sd5|t1           -0.979    0.024  -41.332    0.000   -0.979   -0.979
    sd5|t2           -0.468    0.021  -22.705    0.000   -0.468   -0.468
    sd5|t3            0.029    0.020    1.486    0.137    0.029    0.029
    sd5|t4            0.497    0.021   23.981    0.000    0.497    0.497
    sd6|t1           -0.643    0.021  -30.055    0.000   -0.643   -0.643
```

```
sd6|t2            -0.124    0.020   -6.227    0.000   -0.124   -0.124
sd6|t3             0.358    0.020   17.636    0.000    0.358    0.358
sd6|t4             0.853    0.023   37.626    0.000    0.853    0.853
sd7|t1             0.385    0.020   18.922    0.000    0.385    0.385
sd7|t2             0.873    0.023   38.259    0.000    0.873    0.873
sd7|t3             1.395    0.029   48.613    0.000    1.395    1.395
sd7|t4             1.842    0.038   47.877    0.000    1.842    1.842
sd8|t1            -0.374    0.020  -18.389    0.000   -0.374   -0.374
sd8|t2             0.135    0.020    6.765    0.000    0.135    0.135
sd8|t3             0.633    0.021   29.688    0.000    0.633    0.633
sd8|t4             1.129    0.025   44.854    0.000    1.129    1.129
it6|t1            -0.466    0.021  -22.611    0.000   -0.466   -0.466
it6|t2             0.036    0.020    1.834    0.067    0.036    0.036
it6|t3             0.517    0.021   24.851    0.000    0.517    0.517
it6|t4             1.015    0.024   42.271    0.000    1.015    1.015
it7|t1            -0.077    0.020   -3.857    0.000   -0.077   -0.077
it7|t2             0.413    0.020   20.206    0.000    0.413    0.413
it7|t3             0.884    0.023   38.602    0.000    0.884    0.884
it7|t4             1.402    0.029   48.665    0.000    1.402    1.402
it8|t1            -0.140    0.020   -7.049    0.000   -0.140   -0.140
it8|t2             0.358    0.020   17.636    0.000    0.358    0.358
it8|t3             0.881    0.023   38.488    0.000    0.881    0.881
it8|t4             1.398    0.029   48.639    0.000    1.398    1.398
it9|t1             0.039    0.020    1.960    0.050    0.039    0.039
it9|t2             0.540    0.021   25.842    0.000    0.540    0.540
it9|t3             1.049    0.024   43.108    0.000    1.049    1.049
it9|t4             1.583    0.032   49.325    0.000    1.583    1.583
it10|t1           -0.950    0.023  -40.539    0.000   -0.950   -0.950
it10|t2           -0.467    0.021  -22.643    0.000   -0.467   -0.467
it10|t3            0.023    0.020    1.138    0.255    0.023    0.023
it10|t4            0.532    0.021   25.502    0.000    0.532    0.532
sd9|t1            -0.666    0.022  -30.968    0.000   -0.666   -0.666
sd9|t2            -0.136    0.020   -6.859    0.000   -0.136   -0.136
sd9|t3             0.352    0.020   17.385    0.000    0.352    0.352
sd9|t4             0.882    0.023   38.516    0.000    0.882    0.882
sd10|t1           -1.352    0.028  -48.222    0.000   -1.352   -1.352
sd10|t2           -0.857    0.023  -37.742    0.000   -0.857   -0.857
sd10|t3           -0.344    0.020  -17.008    0.000   -0.344   -0.344
sd10|t4            0.148    0.020    7.460    0.000    0.148    0.148
sd11|t1           -1.447    0.030  -48.965    0.000   -1.447   -1.447
sd11|t2           -0.969    0.024  -41.060    0.000   -0.969   -0.969
sd11|t3           -0.475    0.021  -22.985    0.000   -0.475   -0.475
sd11|t4            0.001    0.020    0.032    0.975    0.001    0.001
```

```
    sd12|t1          -1.852    0.039  -47.766    0.000   -1.852   -1.852
    sd12|t2          -1.300    0.027  -47.649    0.000   -1.300   -1.300
    sd12|t3          -0.777    0.022  -35.081    0.000   -0.777   -0.777
    sd12|t4          -0.282    0.020  -14.021    0.000   -0.282   -0.282
    sd13|t1           0.043    0.020    2.182    0.029    0.043    0.043
    sd13|t2           0.524    0.021   25.130    0.000    0.524    0.524
    sd13|t3           1.032    0.024   42.693    0.000    1.032    1.032
    sd13|t4           1.553    0.031   49.315    0.000    1.553    1.553
    sd14|t1           0.344    0.020   17.008    0.000    0.344    0.344
    sd14|t2           0.852    0.023   37.598    0.000    0.852    0.852
    sd14|t3           1.344    0.028   48.143    0.000    1.344    1.344
    sd14|t4           1.822    0.038   48.081    0.000    1.822    1.822
    sd15|t1           0.203    0.020   10.175    0.000    0.203    0.203
    sd15|t2           0.716    0.022   32.873    0.000    0.716    0.716
    sd15|t3           1.195    0.026   46.097    0.000    1.195    1.195
    sd15|t4           1.728    0.035   48.838    0.000    1.728    1.728
    sd16|t1          -1.570    0.032  -49.325    0.000   -1.570   -1.570
    sd16|t2          -1.053    0.024  -43.186    0.000   -1.053   -1.053
    sd16|t3          -0.548    0.021  -26.183    0.000   -0.548   -0.548
    sd16|t4          -0.039    0.020   -1.960    0.050   -0.039   -0.039

Variances:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
    SD                1.000                               1.000    1.000
    factor1           1.000                               1.000    1.000
    factor2           1.000                               1.000    1.000
   .it1               0.637                               0.637    0.637
   .it2               0.322                               0.322    0.322
   .it3               0.432                               0.432    0.432
   .it4               0.703                               0.703    0.703
   .it5               0.468                               0.468    0.468
   .sd1               0.411                               0.411    0.411
   .sd2               0.427                               0.427    0.427
   .sd3               0.423                               0.423    0.423
   .sd4               0.424                               0.424    0.424
   .sd5               0.436                               0.436    0.436
   .sd6               0.395                               0.395    0.395
   .sd7               0.427                               0.427    0.427
   .sd8               0.438                               0.438    0.438
   .it6               0.344                               0.344    0.344
   .it7               0.375                               0.375    0.375
   .it8               0.462                               0.462    0.462
   .it9               0.344                               0.344    0.344
   .it10              0.262                               0.262    0.262
   .sd9               0.419                               0.419    0.419
   .sd10              0.428                               0.428    0.428
   .sd11              0.387                               0.387    0.387
   .sd12              0.422                               0.422    0.422
   .sd13              0.415                               0.415    0.415
   .sd14              0.416                               0.416    0.416
   .sd15              0.425                               0.425    0.425
   .sd16              0.439                               0.439    0.439
```

We see that the fit index was adequate, all factor loadings were significant and all desirability regressions for the extra items were significant.

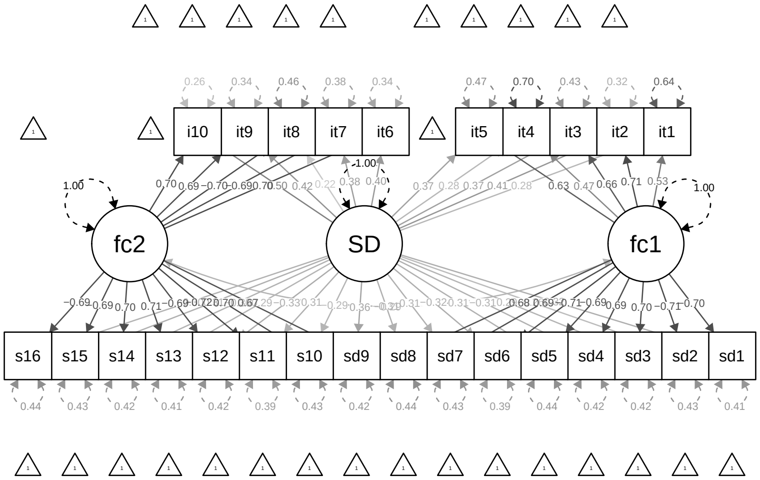
To extract factor scores to use for other analyses, simply use the following code.

```
data_with_scores <- lavPredict(sem.fit,
                                type = "lv",
                                method = "EBM",
                                label = TRUE,
                                append.data = TRUE,
                                optim.method = "bfgs"
                                )
```

We see that in the variable `data_with_scores`, the factor scores of each subject were calculated and these scores were added to their database.

Let's see an image representation of the model using the code below.

```
semPlot::semPaths(object = sem.fit,
                  layout = "tree2",
                  rotation = 3,
                  whatLabels = "std",
                  edge.label.cex = 0.5,
                  what = "std",
                  edge.color = "black")
```



8.7 References

Bastos, R. V. S., Valentini, F. (2023). *Simulations for two theoretically sound controls for social desirability: MIMIC and Forced-Choice.* (Publication No. 157.932 B33s). Master's thesis, Universidade São Francisco.

Connelly, B. S., & Chang, L. (2016). A meta-analytic multitrait multirater separation of substance and style in social desirability scales. *Journal of Personality*, *84*(3), 319-334. https://doi.org/10.1111/jopy.12161

Edwards, A. L. (1953). The relationship between the judged desirability of a trait and the probability that the trait will be endorsed. *Journal of Applied Psychology*, *37*(2), 90. https://doi.org/10.1037/h0058073

Edwards, A. L. (1957). *The social desirability variable in personality assessment and research.* Dryden Press.

Edwards, A. L. (1967). The social desirability variable: A broad statement. In I. A. Berg (Ed.), *Response set in personality assessment* (pp. 32–47). Aldine.

Epskamp S (2022). *semPlot: Path Diagrams and Visual Analysis of Various SEM Packages' Output.* R package. https://CRAN.R-project.org/package=semPlot

Ferrando, P. J. (2005). Factor analytic procedures for assessing social desirability in binary items. *Multivariate Behavioral Research*, *40*(3), 331-349. https://doi.org/10.1207/s15327906mbr4003_3

Ferrando, P. J., Lorenzo-Seva, U., & Chico, E. (2009). A general factor-analytic procedure for assessing response bias in questionnaire measures. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(2), 364-381. https://doi.org/10.1080/10705510902751374

Graziano, W. G., & Tobin, R. M. (2002). Agreeableness: Dimension of personality or social desirability artifact?. *Journal of Personality*, *70*(5), 695-728. https://doi.org/10.1111/1467-6494.05021

Greenblatt, R. L., Mozdzierz, G. J., & Murphy, T. J. (1984). Content and response-style in the construct validation of self-report inventories: A canonical analysis. *Journal of clinical psychology*, *40*(6), 1414-1420. https://doi.org/10.1002/1097-4679(198411)40:6/%3C1414::AID-JCLP2270400624/3.0.CO;2-K>

Griffith, R., Malm, T., English, A., Yoshita, Y., & Gujar, A. (2006). Applicant faking behavior: Teasing apart the influence of situational variance, cognitive biases, and individual differences. In R. L. Griffith & M. H. Peterson (Eds.), *A closer examination of applicant faking behavior* (pp. 151 – 178). Information Age.

Hebert, J. R., Ma, Y., Clemow, L., Ockene, I. S., Saperia, G., Stanek, E. J., Merriam, P. A., & Ockene, J. K. (1997). Gender differences in social desirability and social approval bias in dietary self-report. *American Journal of Epidemiology*, *146*(12), 1046–1055. https://doi.org/10.1093/oxfordjournals.aje.a009233

King, M. F., & Bruner, G. C. (2000). Social desirability bias: A neglected aspect of validity testing. *Psychology & Marketing*, *17*(2), 79-103. https://doi.org/10.1002/(SICI)1520-6793(200002)17:2%3c79::AID-MAR2%3e3.0.CO;2-0

Lange, F., & Dewitte, S. (2019). Measuring pro-environmental behavior: Review and recommendations. *Journal of Environmental Psychology*, *63*, 92-100. https://doi.org/10.1016/j.jenvp.2019.04.009

Lanz, L., Thielmann, I., & Gerpott, F. H. (2022). Are social desirability scales desirable? A meta-analytic test of the validity of social desirability scales in the context of prosocial behavior. *Journal of Personality*, *90*(2), 203-221. https://doi.org/10.1111/jopy.12662

Leite, W. L., & Cooper, L. A. (2009). Detecting social desirability bias using factor mixture models. *Multivariate Behavioral Research*, *45*(2), 271–293. https://doi.org/10.1080/00273171003680245

Li, A., & Bagger, J. (2006). Using the BIDR to distinguish the effects of impression management and self-deception on the criterion validity of personality measures: A meta-analysis. *International Journal of Selection and Assessment*, *14*(2), 131-141. https://doi.org/10.1111/j.1468-2389.2006.00339.x

Malhotra, N. K. (1988). Some observations on the state of the art in marketing research. *Journal of the Academy of Marketing Science*, *16*(1), 4-24. https://doi.org/10.1177/009207038801600102

McFarland, L. A., & Ryan, A. M. (2000). Variance in faking across noncognitive measures. *Journal of Applied Psychology*, *85*(5), 812-821. https://doi.org/10.1037/0021-9010.85.5.812

Mirowsky, J., & Ross, C. E. (1991). Eliminating Defense and Agreement Bias from Measures of the Sense of Control: A 2 X 2 Index. *Social Psychology Quarterly*, *54*(2), 127. https://doi.org/10.2307/2786931

Navarro-Gonzalez D, Vigil-Colet A, Ferrando PJ, Lorenzo-Seva U, Tendeiro JN (2021). *vampyr: Factor Analysis Controlling the Effects of Response Bias.* https://CRAN.R-project.org/package=vampyr

Nederhof, A. J. (1985). Methods of coping with social desirability bias: A review. *European Journal of Social Psychology*, *15*, 263–280. https://doi.org/10.1002/ejsp.2420150303

Ones, D. S., Viswesvaran, C., & Reiss, A. D. (1996). Role of social desirability in personality testing for personnel selection: The red herring. *Journal of applied psychology*, *81*(6), 660. https://doi.org/10.1037/0021-9010.81.6.660

Paulhus, D. L. (1981). Control of social desirability in personality inventories: Principal-factor deletion. *Journal of Research in Personality*, *15*(3), 383–388. https://doi.org/10.1016/0092-6566(81)90035-0

Paulhus, D. L. (1984). Two-component models of socially desirable responding. Journal of Personality and Social Psychology, 46(3), 598–609. https://doi.org/10.1037/0022-3514.46.3.598

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), Measures of social psychological attitudes: Vol. 1. Measures of personality and social psychological attitudes (pp. 17–59). Academic Press. https://doi.org/10.1016/B978-0-12-590241-0.50006-X

Paulhus, D. L., & John, O. P. (1998). Egoistic and moralistic biases in self-perception: The interplay of self-deceptive styles with basic traits and motives. Journal of Personality, 66(6), 1025–1060. https://doi.org/10.1111/1467-6494.00041

Peabody, D. (1967). Trait inferences: Evaluative and descriptive aspects. *Journal of Personality and Social Psychology*, *7*(4, Pt.2), 1-18. https://doi.org/10.1037/h0025230

Peterson, R. A., & Kerin, R. A. (1981). The quality of self-report data: review and synthesis. *Review of marketing*, 5-20.

Pettersson, E., Mendle, J., Turkheimer, E., Horn, E. E., Ford, D. C., Simms, L. J., & Clark, L. A. (2014). Do maladaptive behaviors exist at one or both ends of personality traits? *Psychological Assessment*, *26*(2), 433-446. https://doi.org/10.1037/a0035587

Pettersson, E., Turkheimer, E., Horn, E. E., & Menatti, A. R. (2012). The General Factor of Personality and Evaluation. *European Journal of Personality*, *26*(3), 292-302. https://doi.org/10.1002/per.839

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *88*(5), 879-903. https://doi.org/10.1037/0021-9010.88.5.879

R Core Team (2024). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Raymark, P. H., & Tafero, T. L. (2009). Individual differences in the ability to fake on personality measures. *Human Performance*, *22*(1), 86–103. https://doi.org/10.1080/08959280802541039

Riggio, R. E., Salinas, C., & Tucker, J. (1988). Personality and deception ability. *Personality and Individual Differences*, *9*(1), 189–191. https://doi.org/10.1016/0191-8869(88)90050-5

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2), 1-36. https://doi.org/10.18637/jss.v048.i02

Saucier, G., Ostendorf, F., & Peabody, D. (2001). The non-evaluative circumplex of personality adjectives. *Journal of Personality*, *69*(4), 537-582. https://doi.org/10.1111/1467-6494.694155

Snell, A. F., Sydell, E. J., & Lueke, S. B. (1999). Towards a theory of applicant faking: Integrating studies of deception. *Human Resource Management Review*, *9*(2), 219–242. https://doi.org/10.1016/S1053-4822(99)00019-4

Ten Berge, J. M. F., & Kiers, H. A. L. (1991). A numerical approach to the approximate and the exact minimum rank of a covariance matrix. *Psychometrika*, *56*, 309–315. https://doi.org/10.1007/BF02294464

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin*, *133*(5), 859–883. https://doi.org/10.1037/0033-2909.133.5.859

Uziel, L. (2010). Rethinking social desirability scales: From impression management to interpersonally oriented self-control. *Perspectives on Psychological Science*, *5*(3), 243–262. https://doi.org/10.1177/1745691610369465

Vecina, M. L., Chacón, F., & Pérez-Viejo, J. M. (2016). Moral absolutism, self-deception, and moral self-concept in men who commit intimate partner violence: A comparative study with an opposite sample. *Violence Against Women*, *22*(1), 3–16. https://doi.org/10.1177/1077801215597791

de Vries, R. E., Zettler, I., & Hilbig, B. E. (2014). Rethinking trait conceptions of social desirability scales: Impression management as an expression of honesty-humility. *Assessment*, *21*(3), 286–299. https://doi.org/10.1177/1073191113504619

Williams, E. A., Pillai, R., Lowe, K. B., Jung, D., & Herst, D. (2009). Crisis, charisma, values, and voting behavior in the 2004 presidential election. *The Leadership Quarterly*, *20*(2), 70–86. https://doi.org/10.1016/j.leaqua.2009.01.002

Zickar, M. J., & Robie, C. (1999). Modeling faking good on personality items: An item-level analysis. *Journal of Applied Psychology*, *84*(4), 551. https://doi.org/10.1037/0021-9010.84.4.551

Ziegler, M., & Buehner, M. (2009). Modeling socially desirable responding and its effects. *Educational and Psychological Measurement*, *69*(4), 548-565. https://doi.org/10.1177/0013164408324469

Ziegler, M., Maccann, C., & Roberts, R. D. (2011). *New perspectives on faking in personality assessment.* Oxford University Press.

9 Acquiescence Bias

Numerous studies in psychology, education, and marketing involving human subjects are conducted through questionnaires (Bruner et al., 2001). It is assumed that participants will truthfully respond to the items in such research, thus accurately representing their behaviors, thoughts, and feelings with minimal measurement errors. However, it is known that this type of research comes with a host of issues, such as response biases or method effects (e.g., Weijters et al., 2010). In 1942, Cronbach proposed that participants respond to a true-or-false test. From his data, he observed that some respondents tended to choose the "true" option more frequently than others. This style of responding to a questionnaire is termed **acquiescence** and is commonly defined as **the positive endorsement of the item, regardless of its content** (Robinson et al., 1973), while disagreement is endorsing the item negatively. Thus, those who are more acquiescent tend to mark higher response options on the questions. Table 9.1 provides an example of responding to an extroversion questionnaire. In this example, note that the respondent tends to mark closer to the extremes (indicated in bold), which would indicate an acquiescent person.

Table 9.1: Hypothetical Acquiescence Responses in Extroversion Items

Item	Totally Disagree	Partially Disagree	Partially agree	I totally agree
I am a communicative person	1	2	3	***4***
I like interacting with people	1	2	***3***	4
I don't feel energized when I have large social interactions	1	2	3	***4***

9.1 Acquiescence: trait or state?

Acquiescence is a response style correlated with individual and cultural variables. Literature suggests that individuals with high levels of acquiescence tend to be young, non-depressed, and have a high sense of coherence (Hinz et al., 2007), as well as possessing lower educational levels (Soto et al., 2008). A study investigating whether acquiescence is an inherited trait found no relationship between acquiescence and genetic sharing among monozygotic and dizygotic siblings, suggesting it is also influenced by environmental factors (Kam et al., 2013). Further evidence of its environmental influence includes research indicating that respondents from collectivist cultures tend to be more compliant than those from individualistic cultures (Chen et al., 1995).

Studies suggest that a portion of acquiescence remains stable over time (Billiet & Davidov, 2008). However, employing a latent state-trait modeling approach, Danner et al. (2015) uncovered that acquiescence exhibits trait-like characteristics (i.e., stable over time) as well as state-like features (i.e., subject to situational changes), indicating that both individual traits, such as cognitive ability or personality, and situational factors, such as fatigue, should be taken into account when investigating acquiescence. Some critics of acquiescence research, such as Ferrando et al. (2004), argue that acquiescence should not represent a personality trait because this latent trait cannot be measured through scales. This notion is flawed, and we will delve into this further later. Nonetheless, the authors found that acquiescence is consistent across different domains, present both in studies of personality factors and attitudes (Ferrando et al., 2004).

9.2 Problems Related to Acquiescence: Control Through Inverted Items

Acquiescence poses a myriad of challenges for analyses pertaining to a psychological instrument, even when it explains little of the data variance (Danner et al., 2015; Savalei & Falk, 2014). In a simulation study with random intercepts, it was found that acquiescence affects validity coefficients, overestimating positive regressions and underestimating negative regressions (Valentini and Hauck Filho, 2020). Furthermore, through simulation studies, the influence of acquiescence on factor analysis was observed, leading to poor performance of extraction methods with data uncontrolled for acquiescence (Valentini, 2017). The same acquiescence issue can be encountered in studies with real data. In studies employing personality and attitude scales, it was found that not only was the scale's structural factor affected, but also the dimensionality and magnitude of relationships (Kam et al., 2012). Thus, the importance of attempting to control or eliminate this type of bias from analyses becomes evident.

The most common and recommended method for controlling acquiescence bias is to compose the scale with items in the direction of the construct, also known as positive items (e.g., “I am depressed”) and inverted items (e.g., by negation: “I am not depressed” or by using an antagonistic adjective: "I am happy"; Baumgartner and Steenkamp, 2001). There are three considerations to bear in mind when conducting item inversion. The first is that inverted items are known to slow response speed and compel participants to read and process items more carefully (Podsakoff et al., 2003). The second, often overlooked, is to verify whether positive and inverted items are complementary and measure the same construct at different levels (Marsh, 1986), which is not always the case (Chang, 1995). The third consideration arises from the process of recoding inverted items to obtain the total scale score. This process assumes that the two extremes (e.g., "Strongly Disagree" and "Strongly Agree") have the same score (with one being the inverse of the other) and carry the same semantic meaning (Suárez-Alvez et al., 2018). However, agreeing with an item is not the same as disagreeing with its inverted counterpart (Enos, 2000), furthermore, some constructs (e.g., resilience) are conceptually unsuitable for inversion as they are of a positive nature (Luthar and Zigler, 1991). Thus, the inclusion of inverted items depends greatly on the feasibility of manipulation and how this manipulation is related to the response scale and the construct.

Item inversion alters how people respond to that item (Pilotte & Gable, 1990). This may occur because participants' cognitive processing for each type of item is not necessarily the same (Suárez-Alvez et al., 2018), especially when individuals' reading skills are lower (Marsh, 1996). Therefore, if understanding inverted items requires better linguistic ability, then these items favor respondents with better verbal skills (Suárez-Alvez et al., 2018), and the construct being measured may be contaminated by other variables that have little relation to the study's objective, such as lack of attention and confusion when responding to the item (Van Sonderen et al., 2013). Additionally, in some cases, participants respond inconsistently to scales that contain inverted items with antagonistic adjectives (Zhang et al., 2016). This occurs because participants may not interpret antonyms used in inverted items as contradictory to the construct of interest, thus they may agree with both the positive item and the inverted item (Weijters & Baumgartner, 2012).

Some studies suggest that the combination of positive and inverted items does not reduce acquiescence bias (Sauro & Lewis, 2011; cf. Primi et al., 2020), and that the proportion of extreme responses for both types of items is similar (Sauro & Lewis, 2011). Salazar's study (2015) demonstrated that inverted items do not alter the response pattern of positive items when combined on the same scale, and that the data better fit the theorized factorial structure using only positive items. Additionally, inverted items bring about methodological issues, such as the emergence of different factors for positive and inverted items (Knight et al., 1988). Woods (2006) showed that if at least 10% of participants respond carelessly to 10 inverted items (on a scale of 23 items), researchers are likely to reject the unidimensional model. Meanwhile, Hughes (2009), through a simulation study, found that even a small percentage of incorrect responses to inverted items leads to significant differences in scale means, thus altering subsequent analyses. Other issues include positive items correlating more with each other than inverted items (Hinz et al., 2007), scales with only positive items providing more precise descriptions both practically and statistically than mixed or solely inverted item scales (Schriesheim & Hill, 1981), inverted items decreasing instrument accuracy (Schriesheim & Hill, 1981), increasing interpretation problems in cross-cultural studies (Wong et al., 2003), contaminating the covariance structure of the data (Savalei & Falk, 2014), and reducing model fit (Essau, 2012).

Controlling acquiescence may seem entirely detrimental so far. However, acquiescence is a response bias, introducing measurement error into responses. Thus, if a person is acquiescent and responds to an instrument solely with positive items, it will not be possible to detect levels of acquiescence and control it. This could be one of the reasons for a better model fit with only positive items, given that the error associated with acquiescent response in one item correlates with that of another item. Therefore, it is necessary to weigh the options carefully, especially considering the size of the instrument and the sample to be collected.

9.3 The Removal of Acquiescence: Statistics and Design

Acquiescence can be removed in two ways: through statistical analyses that eliminate acquiescence from the covariance structure of the data or through research design. Regarding analyses, acquiescence should be addressed prior to conducting any covariance-based analysis, such as reliability analysis, factor analysis, and structural equation modeling (Billiet and Mc-Clendon, 2000; Cambré et al., 2002; Kam et al., 2012; Lorenzo-Seva et al., 2016). To eliminate acquiescence from the covariance structure of the data, it is generally necessary to make two assumptions (Savalei and Falk, 2014). The first assumption is that the acquiescence of each item is independent of the latent factor being measured, meaning this case should be carefully examined in each analysis, as it may not always hold true (Ferrando et al., 2003). The second assumption is that acquiescence bias behaves like a latent factor, affecting different items in different ways (Billiet and McClendon, 2000), and should also be critically examined in each case.

Despite the possibility of controlling acquiescence through scale score composition, it cannot be controlled within the factorial structure of the scale (Savalei and Falk, 2014). To address this issue, some strategies are employed in research design to mitigate this bias. The study by Weijters et al. (2010) demonstrates that individuals exhibit higher levels of acquiescence if the questionnaire labels all response levels (e.g., ranging from “Strongly Disagree” to “Strongly Agree”) and includes a midpoint (e.g., “Neither Agree nor Disagree”). Additionally, adding more gradations of agreement and disagreement does not affect the level of acquiescence, meaning a 5-point scale does not show less or more acquiescence than a 7-point scale (Weijters et al., 2010). Barnette (2000) found in their research that reversing half of the response options, the anchors, leads to higher levels of accuracy and observed variance.

Fribourg et al. (2006) employed a different research design compared to others, comparing Likert scales with semantic differential scales. The study results indicate that semantic differential data are more suitable to the model than Likert format, and they exhibit clearer unidimensionality. Furthermore, the semantic differential scale did not correlate with measures of social desirability, further reducing response falsification (Friborg et al., 2006). Additionally, a semantic differential response scale showed no acquiescence bias in another study (Lewis, 2018).

Finally, Zhang & Savalei (2016) explored an alternative version that enhances the factorial structure of psychological scales, termed the expanded format. The expanded format involves writing one item for each variation of the response scale, meaning if it's a four-point scale, the researcher must write one item representing each level of the latent trait. The participant selects which of these four items best represents them. The expanded format yielded a lower number of dimensions in an exploratory factor analysis (closer to the previously theorized number), better model fit indices, and improved reliability indices (Zhang & Savalei, 2016).

9.4 How to Control Acquiescence in R

9.4.1 Controlling Acquiescence with Ferrando et al. (2009)

To run with the analysis by Ferrando et al. (2009), we first have to install the *vampyr* (Navarro-Gonzalez et al., 2021) package to run the analyses.

```
install.packages("vampyr")
```

And tell the program that we are going to use the functions of these packages.

```
library(vampyr)
```

To run the analyses, we will use a dataset from the package itself. Let's see what the database looks like.

```
summary(vampyr::vampyr_example)
```

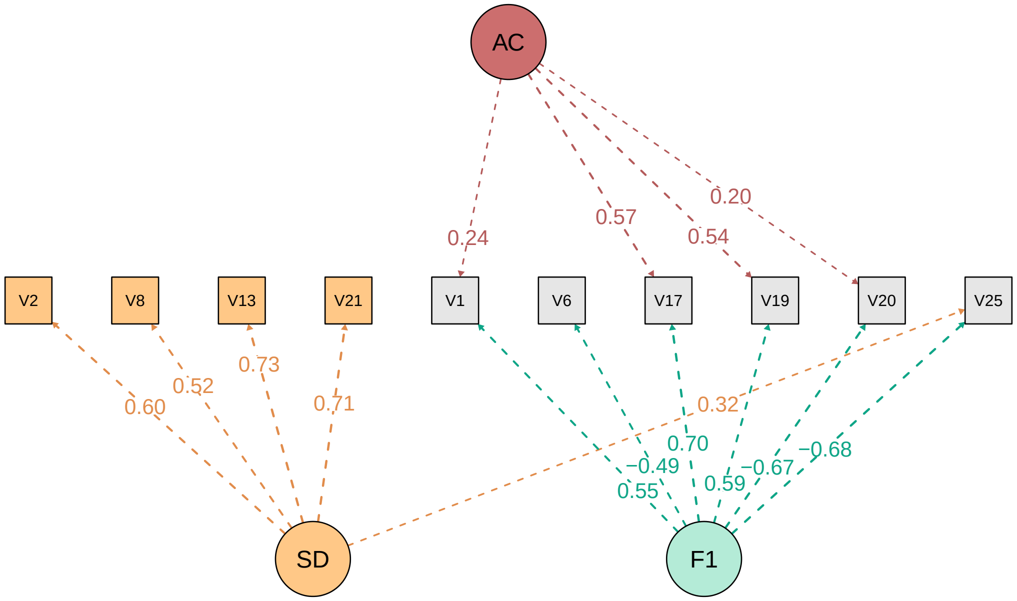
```
       V2              V8             V13             V21       
 Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000  
 1st Qu.:3.000   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:2.000  
 Median :4.000   Median :4.000   Median :2.000   Median :3.000  
 Mean   :3.667   Mean   :3.263   Mean   :2.317   Mean   :2.947  
 3rd Qu.:5.000   3rd Qu.:4.000   3rd Qu.:3.000   3rd Qu.:4.000  
 Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000  
       V1              V6             V17            V19             V20       
 Min.   :1.000   Min.   :1.000   Min.   :1.00   Min.   :1.000   Min.   :1.000  
 1st Qu.:3.000   1st Qu.:1.000   1st Qu.:3.00   1st Qu.:3.000   1st Qu.:1.000  
 Median :4.000   Median :2.000   Median :4.00   Median :4.000   Median :2.000  
 Mean   :3.643   Mean   :2.467   Mean   :3.71   Mean   :3.493   Mean   :1.997  
 3rd Qu.:5.000   3rd Qu.:3.000   3rd Qu.:5.00   3rd Qu.:5.000   3rd Qu.:3.000  
 Max.   :5.000   Max.   :5.000   Max.   :5.00   Max.   :5.000   Max.   :5.000  
```

```
      V25
 Min.   :1.000
 1st Qu.:1.000
 Median :1.000
 Mean   :1.687
 3rd Qu.:2.000
 Max.   :5.000
```

According to the package, we have a database with 300 observations and 10 variables, where 6 items measure physical aggression and we have 4 markers of social desirability. Items 1, 2, 3, and 4 are markers of DS ("pure" measures of DS), and the remaining 6 items measure physical aggression. Items 5, 7 and 8 are in the positive pole of the target construct and items 6, 9 and 10 are written in the negative pole of the target construct.

To perform the analysis controlling both desirability and acquiescence, simply use the following code.

```
res <- ControlResponseBias(vampyr_example,
                           content_factors = 1,
                           SD_items = c(1,2,3,4),
                           corr = "Polychoric",
                           contAC = TRUE,
                           unbalanced_items = c(),
                           rotat = "promin",
                           PA = FALSE,
                           factor_scores = FALSE,
                           path = TRUE
                           )
```



```
DETAILS OF ANALYSIS

Number of participants                        :   300
Number of items                               :    10
Items selected as SD items                    :  1, 2, 3, 4
Dispersion Matrix                             : Polychoric Correlations
Method for factor extraction                  : Unweighted Least Squares (ULS)
Rotation Method                               : none

-----------------------------------------------------------------------

Univariate item descriptives

Item       Mean        Variance    Skewness     Kurtosis (Zero centered)

Item   1   3.667       1.260      -0.555       -0.566
Item   2   3.263       1.760      -0.379       -1.005
Item   3   2.317       1.695       0.601       -0.880
Item   4   2.947       1.924      -0.033       -1.284
Item   5   3.643       1.374      -0.565       -0.535
Item   6   2.467       1.802       0.487       -0.967
Item   7   3.710       1.678      -0.652       -0.716
Item   8   3.493       1.629      -0.411       -0.862
Item   9   1.997       1.515       1.041       -0.011
Item  10   1.687       0.925       1.293        0.838

Polychoric correlation is advised when the univariate distributions of ordinal items are
asymmetric or with excess of kurtosis. If both indices are lower than one in absolute value,
then Pearson correlation is advised. You can read more about this subject in:

Muthen, B., & Kaplan D. (1985). A Comparison of Some Methodologies for the Factor Analysis o
Non-Normal Likert Variables. British Journal of Mathematical and Statistical Psychology, 38,

Muthen, B., & Kaplan D. (1992). A Comparison of Some Methodologies for the Factor Analysis o
Non-Normal Likert Variables: A Note on the Size of the Model. British Journal of Mathematical
and Statistical Psychology, 45, 19-30.

-----------------------------------------------------------------------

Adequacy of the dispersion matrix

Determinant of the matrix     = 0.047816437916934
Bartlett's statistic          =   896.4 (df =    45; P = 0.000000)
Kaiser-Meyer-Olkin (KMO) test = 0.76664 (fair)

-----------------------------------------------------------------------
EXPLORATORY FACTOR ANALYSIS CONTROLLING SOCIAL DESIRABILITY AND ACQUIESCENCE
-----------------------------------------------------------------------

Robust Goodness of Fit statistics

          Root Mean Square Error of Approximation (RMSEA) = 0.032

 Robust Mean-Scaled Chi Square with 23 degrees of freedom = 30.146

              Non-Normed Fit Index (NNFI; Tucker & Lewis) = 0.989
                             Comparative Fit Index (CFI) = 0.994
                             Goodness of Fit Index (GFI) = 0.977

-----------------------------------------------------------------------

                  Root Mean Square Residuals (RMSR) = 0.0452
Expected mean value of RMSR for an acceptable model = 0.0578 (Kelley's criterion)
```

```
-----------------------------------------------------------------------

Unrotated loading matrix

         Factor SD Factor AC Factor 1
Item   1   0.60258   0.00000  0.00000
Item   2   0.51525   0.00000  0.00000
Item   3   0.72711   0.00000  0.00000
Item   4   0.71129   0.00000  0.00000
Item   5  -0.07851   0.23765  0.54830
Item   6   0.27520   0.00216 -0.49050
Item   7  -0.16413   0.57416  0.70136
Item   8  -0.14320   0.54069  0.59103
Item   9   0.26560   0.19594 -0.66798
Item  10   0.31733   0.06247 -0.68222
```

This analysis allows controlling the effects of two response biases: Social Desirability and Acquiescence, extracting the variance due to these factors before extracting the content variance. If you don't have or want to control acquiescence, just remove the `SD_items = c(1,2,3,4)` argument.

We do not always have an instrument that is completely balanced, that is, we do not always have the same number of positive and negative items in an instrument. This must be said to the function, just put the column position of the items in your database in the `unbalanced_items = c()` argument. For example, if the items in columns 10, 11, and 17 of your database are items that do not have an opposite pole, you would put the argument as follows: `unbalanced_items = c(10,11,17)`. The items you place in this argument will not be used in the calculation.

We see that Bartlett's test of sphericity and KMO were calculated before proceeding with Exploratory Factor Analysis. Furthermore, the model fit indices were calculated. We also see that items 6, 9 and 10 have even high loadings on the desirability factor ("Factor SD"), and items 5, 7 and 8 on the acquiescence factor ("Factor AC").

The function allows you to calculate people's factor scores. Factor scores work like when you calculate the mean scores of an instrument to correlate with others, but calculating averages has certain assumptions, while factor scores have others. So, to calculate the factor scores while controlling the DS and acquiescence biases, simply leave the factor scores argument as TRUE (`factor_scores = TRUE`) and save the result in some variable. In our case, we save the results in the `res` variable.

To save only the factor scores, simply extract the scores from the `res` list.

```
scores <- res$Factor_scores
```

This way, just put this column of factor scores together with your data (using “cbind()”) and then calculate whatever analysis you want.

9.4.2 Controlling Acquiescence with Random Intercepts

First, we have to install the *lavaan* (Rosseel, 2012) package for the analyzes and the *EGAnet* (Golino & Christensen, 2023) package for the dataset.

```
install.packages("lavaan")
install.packages("EGAnet")
```

Next, we tell R that we are going to use the functions from the packages.

```
library(lavaan)
library(EGAnet)
```

Then, we must have information on which model we should test. In other words, we have to know the theory behind some instrument: how many factors we have, which items represent which factors, whether or not the factors are correlated, etc.

Let’s use the *EGAnet* package as an example (i.e., *Wiener Matrizen-Test 2*), which has 2 factors and items on the positive and negative pole.

```
model_RI <- '
              factor1 =~ NA*wmt1 + wmt2 + wmt3 + wmt5 + wmt11 +
              wmt12 + wmt13 + wmt15 + wmt16 + wmt17 + wmt18

              factor2 =~ NA*wmt4 + wmt6 + wmt7 + wmt8 +
              wmt9 + wmt10 + wmt14

              # Random Intercepts
              acquiescence =~ 1*wmt1 + 1*wmt2 + 1*wmt3 + 1*wmt5 +
              1*wmt11 + 1*wmt12 + 1*wmt13 + 1*wmt15 + 1*wmt16 +
              1*wmt17 + 1*wmt18 + 1*wmt4 + 1*wmt6 + 1*wmt7 +
              1*wmt8 + 1*wmt9 + 1*wmt10 + 1*wmt14

              factor1 ~~ 0*acquiescence
              factor2 ~~ 0*acquiescence

              acquiescence ~~ acquiescence

              factor1 ~~ 1*factor1
              factor2 ~~ 1*factor2
              '
```

Now let's calculate the internal structure controlling for acquiescence.

```
sem.fit <- lavaan::sem(model = model_RI,
                      data = EGAnet::wmt2[,7:24],
                      estimator = 'WLSMV',
                      ordered = TRUE
                      )

lavaan::summary(sem.fit,
                fit.measures=TRUE,
                standardized=TRUE
        )
```

```
lavaan 0.6.17 ended normally after 43 iterations

  Estimator                                       DWLS
  Optimization method                           NLMINB
  Number of model parameters                        38

  Number of observations                          1185

Model Test User Model:
                                              Standard      Scaled
  Test Statistic                               232.896     285.231
  Degrees of freedom                               133         133
  P-value (Chi-square)                           0.000       0.000
  Scaling correction factor                                  0.873
  Shift parameter                                           18.557
    simple second-order correction

Model Test Baseline Model:

  Test statistic                             12385.490    7849.254
  Degrees of freedom                               153         153
  P-value                                        0.000       0.000
  Scaling correction factor                                  1.589

User Model versus Baseline Model:
  Comparative Fit Index (CFI)                    0.992       0.980
  Tucker-Lewis Index (TLI)                       0.991       0.977

  Robust Comparative Fit Index (CFI)                         0.922
  Robust Tucker-Lewis Index (TLI)                            0.910

Root Mean Square Error of Approximation:

  RMSEA                                          0.025       0.031
  90 Percent confidence interval - lower         0.020       0.026
  90 Percent confidence interval - upper         0.030       0.036
  P-value H_0: RMSEA <= 0.050                    1.000       1.000
  P-value H_0: RMSEA >= 0.080                    0.000       0.000

  Robust RMSEA                                               0.065
  90 Percent confidence interval - lower                     0.054
  90 Percent confidence interval - upper                     0.076
  P-value H_0: Robust RMSEA <= 0.050                         0.011
  P-value H_0: Robust RMSEA >= 0.080                         0.012

Standardized Root Mean Square Residual:

  SRMR                                           0.052       0.052

Parameter Estimates:

  Parameterization                               Delta
  Standard errors                           Robust.sem
  Information                                 Expected
  Information saturated (h1) model        Unstructured

Latent Variables:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  factor1 =~
    wmt1              0.233    0.065    3.580    0.000    0.233    0.233
    wmt2              0.607    0.066    9.213    0.000    0.607    0.607
    wmt3              0.449    0.061    7.410    0.000    0.449    0.449
    wmt5              0.314    0.062    5.054    0.000    0.314    0.314
    wmt11             0.019    0.074    0.260    0.795    0.019    0.019
    wmt12             0.061    0.074    0.828    0.408    0.061    0.061
    wmt13             0.110    0.069    1.603    0.109    0.110    0.110
    wmt15             0.136    0.070    1.947    0.052    0.136    0.136
    wmt16             0.126    0.071    1.772    0.076    0.126    0.126
    wmt17            -0.044    0.078   -0.568    0.570   -0.044   -0.044
    wmt18            -0.339    0.098   -3.452    0.001   -0.339   -0.339
  factor2 =~
    wmt4              0.300    0.056    5.328    0.000    0.300    0.300
    wmt6              0.504    0.052    9.750    0.000    0.504    0.504
    wmt7              0.352    0.055    6.452    0.000    0.352    0.352
    wmt8              0.269    0.057    4.695    0.000    0.269    0.269
    wmt9              0.393    0.054    7.292    0.000    0.393    0.393
    wmt10             0.477    0.054    8.910    0.000    0.477    0.477
    wmt14             0.227    0.060    3.817    0.000    0.227    0.227
  acquiescence =~
    wmt1              1.000                               0.580    0.580
    wmt2              1.000                               0.580    0.580
    wmt3              1.000                               0.580    0.580
    wmt5              1.000                               0.580    0.580
    wmt11             1.000                               0.580    0.580
    wmt12             1.000                               0.580    0.580
    wmt13             1.000                               0.580    0.580
    wmt15             1.000                               0.580    0.580
    wmt16             1.000                               0.580    0.580
    wmt17             1.000                               0.580    0.580
    wmt18             1.000                               0.580    0.580
    wmt4              1.000                               0.580    0.580
    wmt6              1.000                               0.580    0.580
    wmt7              1.000                               0.580    0.580
    wmt8              1.000                               0.580    0.580
    wmt9              1.000                               0.580    0.580
    wmt10             1.000                               0.580    0.580
    wmt14             1.000                               0.580    0.580

Covariances:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  factor1 ~~
    acquiescence      0.000                               0.000    0.000
  factor2 ~~
    acquiescence      0.000                               0.000    0.000
  factor1 ~~
    factor2           0.591    0.078    7.602    0.000    0.591    0.591

Thresholds:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
    wmt1|t1          -0.475    0.038  -12.521    0.000   -0.475   -0.475
    wmt2|t1          -0.881    0.042  -20.956    0.000   -0.881   -0.881
    wmt3|t1          -0.651    0.039  -16.544    0.000   -0.651   -0.651
    wmt5|t1          -0.475    0.038  -12.521    0.000   -0.475   -0.475
    wmt11|t1          0.447    0.038   11.833    0.000    0.447    0.447
    wmt12|t1          0.471    0.038   12.406    0.000    0.471    0.471
    wmt13|t1          0.195    0.037    5.311    0.000    0.195    0.195
    wmt15|t1          0.445    0.038   11.776    0.000    0.445    0.445
    wmt16|t1          0.412    0.038   10.972    0.000    0.412    0.412
    wmt17|t1          0.815    0.041   19.787    0.000    0.815    0.815
    wmt18|t1          0.641    0.039   16.320    0.000    0.641    0.641
    wmt4|t1          -0.158    0.037   -4.325    0.000   -0.158   -0.158
    wmt6|t1          -0.355    0.037   -9.533    0.000   -0.355   -0.355
    wmt7|t1          -0.208    0.037   -5.658    0.000   -0.208   -0.208
    wmt8|t1           0.116    0.037    3.164    0.002    0.116    0.116
    wmt9|t1          -0.158    0.037   -4.325    0.000   -0.158   -0.158
    wmt10|t1         -0.280    0.037   -7.569    0.000   -0.280   -0.280
    wmt14|t1          0.128    0.037    3.513    0.000    0.128    0.128

Variances:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
    acquiescence      0.337    0.016   21.029    0.000    1.000    1.000
    factor1           1.000                               1.000    1.000
    factor2           1.000                               1.000    1.000
   .wmt1              0.609                               0.609    0.609
   .wmt2              0.295                               0.295    0.295
   .wmt3              0.462                               0.462    0.462
   .wmt5              0.565                               0.565    0.565
   .wmt11             0.663                               0.663    0.663
   .wmt12             0.659                               0.659    0.659
   .wmt13             0.651                               0.651    0.651
   .wmt15             0.645                               0.645    0.645
   .wmt16             0.647                               0.647    0.647
   .wmt17             0.661                               0.661    0.661
   .wmt18             0.548                               0.548    0.548
   .wmt4              0.573                               0.573    0.573
   .wmt6              0.409                               0.409    0.409
   .wmt7              0.539                               0.539    0.539
   .wmt8              0.591                               0.591    0.591
   .wmt9              0.508                               0.508    0.508
   .wmt10             0.436                               0.436    0.436
   .wmt14             0.611                               0.611    0.611
```

We can calculate from people's factor scores, just use the following code.

```
scores <- lavaan::lavPredict(
                      sem.fit,
                      type = "lv",
                      method = "EBM",
                      label = TRUE,
                      append.data = TRUE,
                      optim.method = "bfgs"
                      )
```

We see that in the variable “scores” the factor scores of each subject were calculated and these scores were added to their database.

9.4.3 Controlling Acquiescence with Random Intercepts Exploratory Graph Analysis

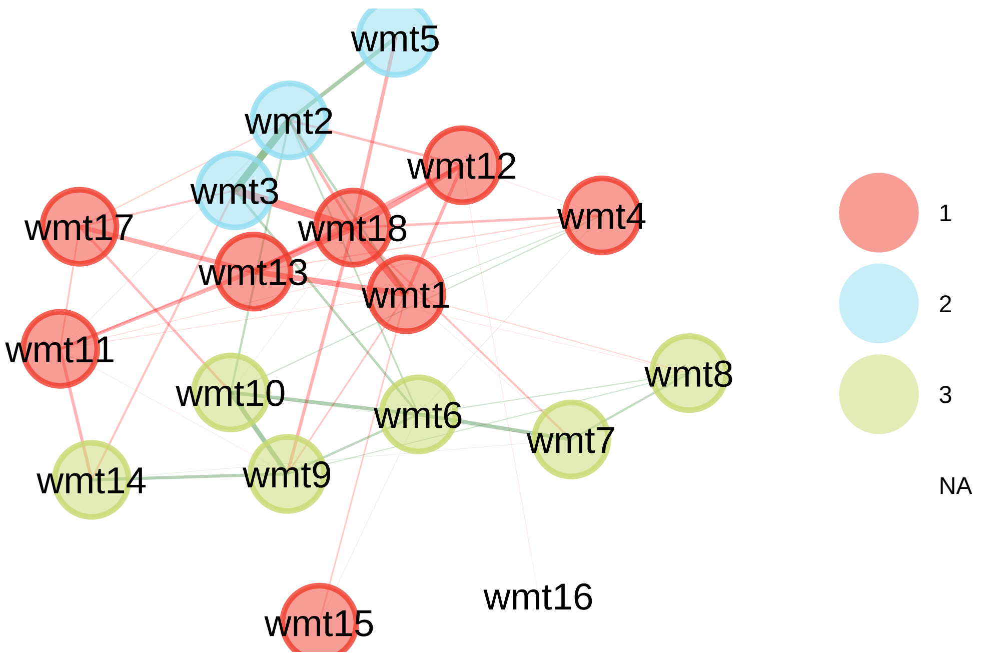
First, we have to install the *EGAnet* (Golino & Christensen, 2023) package for the analyzes and *lavaan* (Rosseel, 2012) for the fit indices.

```
install.packages("EGAnet")
install.packages("lavaan")
```

Next, we tell R that we are going to use the functions from the packages.

```
library(EGAnet)
library(lavaan)
```

```
EGA_RI<- EGAnet::riEGA(data = EGAnet::wmt2[,7:24])
```



We can also bootstrap controlling for acquiescence.

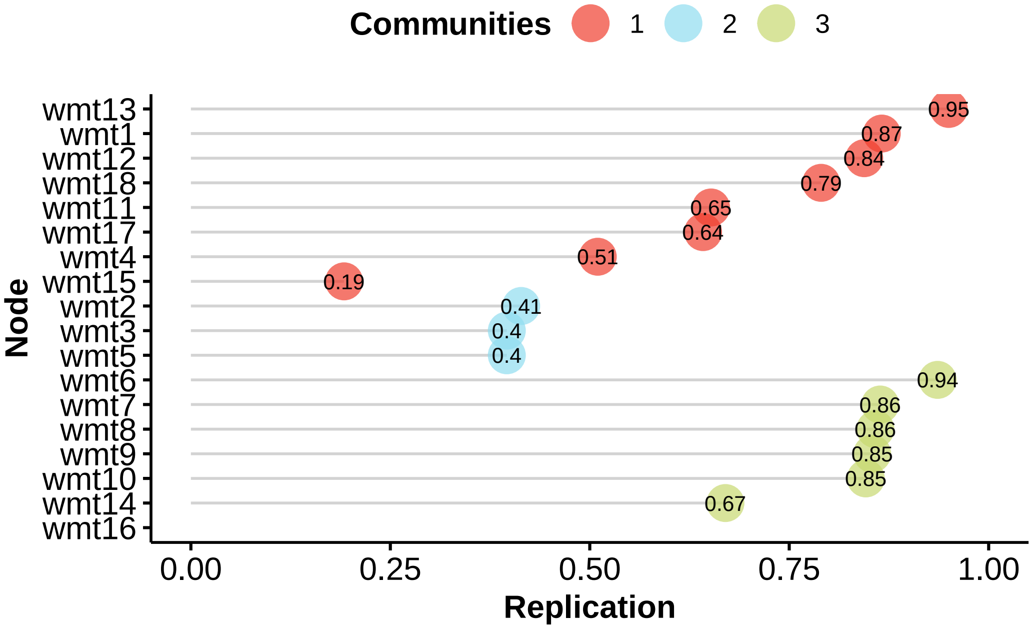
To get a summary of the results, just take the bootstrap output.

```
summary(boot.ri)
```

```
Model: GLASSO (EBIC)
Correlations: auto
Algorithm:  Walktrap
Unidimensional Method:  Louvain

----

EGA Type: riEGA
Bootstrap Samples: 500 (Parametric)

                1     2     3     4    5     6
Frequency:  0.006 0.286 0.444 0.198 0.06 0.006

Median dimensions: 3 [1.24, 4.76] 95% CI
```

Additionally, we can take the stability output of the items.

```
EGAnet::itemStability(boot.ri)
```



```
EGA Type: riEGA
Bootstrap Samples: 500 (Parametric)

Proportion Replicated in Dimensions:

 wmt1  wmt2  wmt3  wmt4  wmt5  wmt6  wmt7  wmt8  wmt9 wmt10 wmt11 wmt12 wmt13
0.866 0.414 0.396 0.510 0.396 0.936 0.864 0.858 0.854 0.846 0.652 0.844 0.950
wmt14 wmt15 wmt16 wmt17 wmt18
0.670 0.192    NA 0.642 0.790
```

We can see network loadings (similar to factor loadings), with the code:

```
Network_loadings <- EGAnet::net.loads(EGA_RI)

print(Network_loadings$std)
```

```
                 1            2           3        NA
wmt13  0.310676334  0.000000000  0.006782463 0.0000000
wmt18  0.291274702  0.293579539  0.099702419 0.0000000
```

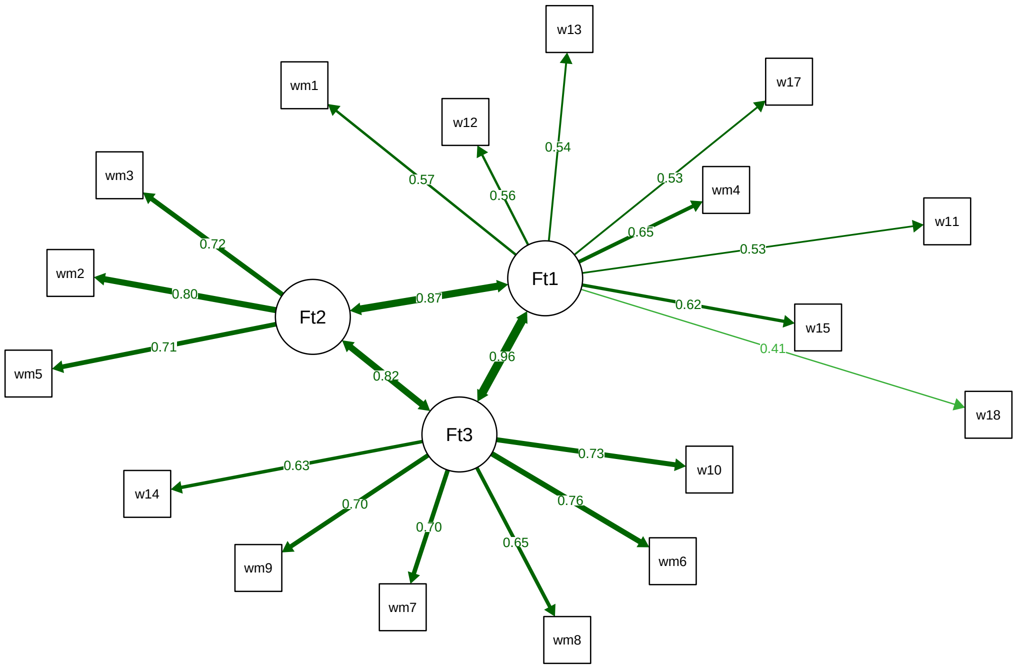
```
wmt1   0.243167322 -0.046997589  0.052806787 0.0000000
wmt12  0.153284729  0.050799771  0.000000000 0.1115768
wmt11  0.102052291  0.008739402  0.062516602 0.0000000
wmt17  0.080001971  0.063525600  0.045647999 0.0000000
wmt4   0.074492247  0.000000000 -0.030165523 0.0000000
wmt15  0.019148038  0.000000000 -0.005967561 0.0000000
wmt2   0.117715634  0.230607341  0.069198724 0.0000000
wmt3   0.123121237  0.151225610  0.081589490 0.0000000
wmt5   0.044845930  0.079381732  0.000000000 0.0000000
wmt9   0.066910306  0.000000000  0.196554454 0.0000000
wmt6  -0.010125599  0.086182827  0.191423280 0.0000000
wmt10  0.051388034  0.043413062  0.152658672 0.0000000
wmt7   0.029578222  0.000000000  0.117856996 0.0000000
wmt8   0.017213035  0.000000000  0.077519918 0.0000000
wmt14  0.038274764 -0.042494612  0.061525090 0.0000000
wmt16 -0.006992994  0.000000000  0.000000000 0.0000000
```

This step by step must be repeated (removing items with low stability or factor loadings in the wrong dimensions) until the stability of the items is above 70% or 75%.

We were also able to obtain the fit through a Confirmatory Factor Analysis by *EGAnet.*

```r
fit <- EGAnet::CFA(EGA_RI$EGA,
                   data = EGAnet::wmt2[,7:24],
                   estimator = "WLSMV",
                   plot.CFA = TRUE,
                   layout = "spring"
                  )
```

```
[1] "wmt1"  "wmt4"  "wmt11" "wmt12" "wmt13" "wmt15" "wmt17" "wmt18"
[1] "wmt2" "wmt3" "wmt5"
[1] "wmt6"  "wmt7"  "wmt8"  "wmt9"  "wmt10" "wmt14"
```



To request fit indices we can use *lavaan*.

```
lavaan::fitMeasures(fit$fit, fit.measures = "all")
```

```
                  npar                          fmin
                37.000                         0.089
                 chisq                            df
               209.834                       116.000
                pvalue                  chisq.scaled
                 0.000                       294.191
             df.scaled                 pvalue.scaled
               116.000                         0.000
  chisq.scaling.factor                baseline.chisq
                 0.741                     11168.058
           baseline.df               baseline.pvalue
               136.000                         0.000
 baseline.chisq.scaled            baseline.df.scaled
              7229.787                       136.000
baseline.pvalue.scaled baseline.chisq.scaling.factor
                 0.000                         1.555
                   cfi                           tli
                 0.991                         0.990
```

```
                  cfi.scaled                   tli.scaled
                       0.975                        0.971
                  cfi.robust                   tli.robust
                       0.906                        0.889
                        nnfi                          rfi
                       0.990                        0.978
                         nfi                         pnfi
                       0.981                        0.837
                         ifi                          rni
                       0.992                        0.991
                 nnfi.scaled                   rfi.scaled
                       0.971                        0.952
                  nfi.scaled                  pnfi.scaled
                       0.959                        0.818
                  ifi.scaled                   rni.scaled
                       0.975                        0.975
                 nnfi.robust                   rni.robust
                       0.889                        0.906
                       rmsea             rmsea.ci.lower
                       0.026                        0.020
             rmsea.ci.upper              rmsea.ci.level
                       0.032                        0.900
               rmsea.pvalue              rmsea.close.h0
                       1.000                        0.050
       rmsea.notclose.pvalue           rmsea.notclose.h0
                       0.000                        0.080
               rmsea.scaled       rmsea.ci.lower.scaled
                       0.036                        0.031
       rmsea.ci.upper.scaled         rmsea.pvalue.scaled
                       0.041                        1.000
rmsea.notclose.pvalue.scaled                 rmsea.robust
                       0.000                        0.075
       rmsea.ci.lower.robust       rmsea.ci.upper.robust
                       0.064                        0.085
         rmsea.pvalue.robust rmsea.notclose.pvalue.robust
                       0.000                        0.211
                         rmr                   rmr_nomean
                       0.050                        0.053
                        srmr                 srmr_bentler
                       0.053                        0.050
         srmr_bentler_nomean                         crmr
                       0.053                        0.053
                 crmr_nomean                   srmr_mplus
                       0.056                           NA
           srmr_mplus_nomean                        cn_05
                          NA                      803.022
                       cn_01                          gfi
                     871.893                        0.985
                        agfi                         pgfi
                       0.980                        0.746
                         mfi                         wrmr
                       0.961                        1.171
```

We can calculate from people's factor scores, just use the following code.

```
fe <- lavaan::lavPredict(fit$fit,
                         type = "lv",
                         method = "EBM",
                         label = TRUE,
                         append.data = TRUE,
                         optim.method = "bfgs"
                         )
```

9.5 References

Barnette, J. J. (2000). Effects of stem and likert response option reversals on survey internal consistency: If you feel the need, there is a better alternative to using those negatively worded stems. *Educational and Psychological Measurement*, *60*(3), 361–370. https://doi.org/10.1177/00131640021970592

Baumgartner, H., & Steenkamp, J. (2001). Response styles in marketing research: A cross national investigation. *Journal of Marketing Research*, *38*(2), 143–156. https://doi.org/10.1509/jmkr.38.2.143.18840

Billiet, J. B., & Davidov, E. (2008). Testing the stability of an acquiescence style factor behind two interrelated substantive variables in a panel design. *Sociological Methods Research*, *36*(4), 542–562. https://doi.org/10.1177/0049124107313901

Billiet, J. B., & McClendon, M. J. (2000). Modeling acquiescence in measurement models for two balanced sets of items. *Structural Equation Modeling*, *7*, 608–628. https://doi.org/10.1207/S15328007SEM0704_5

Bruner, G. C., James, K. E., & Hensel, P. J. (2001). *Marketing scales handbook. A compilation of multi item measures*, volume iii. American Marketing Association.

Cambré, B., Welkenhuysen-Gybels, J., & Billiet, J. (2002). Is it content or style? An evaluation of two competitive measurement models applied to a balanced set of ethnocentrism items. *International Journal of Comparative Sociology*, *43*, 1–20. https://doi.org/10.1177/002071520204300101

Chang, L. (1995). Connotatively consistent and reversed connotatively inconsistent items are not fully equivalent: Generalizability study. *Educational and Psychological Measurement*, *55*, 991–997. https://doi.org/10.1177/0013164495055006007

Chen, C., Shin-ying, L., & Stevenson, H. W. (1995). Response style and cross-cultural comparisons of rating scales among east asian and north american students. *Psychological Science*, *6*, 170–175. https://doi.org/10.1111/j.1467-9280.1995.tb00327.x

Cronbach, L. J. (1942). Studies of acquiescence as a factor in the true-false test. *Journal of Educational Psychology*, *33*, 401–415. https://doi.org/10.1037/h0054677

Danner, D., Aichholzer, J., & Rammstedt, B. (2015). Acquiescence in personality questionnaires: Relevance, domain specificity, and stability. *Journal of Research in Personality*, *57*, 119–130. https://doi.org/10.1016/j.jrp.2015.05.004

Enos, M. M. (2000). Just say no!: The impact of negation in survey research. *Popular Measurement*, *3*(1), 34–39.

Essau, e. a., C. A. (2012). Psychometric properties of the strength and difficulties questionnaire from five european countries. *International Journal of Methods in Psychiatric Research*, *21*(3), 232–245. https://doi.org/10.1002/mpr.1364

Ferrando, P. J., Condon, L., & Chico, E. (2004). The convergent validity of acquiescence: An empirical study relating balanced scales and separate acquiescence scales. *Personality and individual differences*, *37*(7), 1331–1340. https://doi.org/10.1016/j.paid.2004.01.003

Ferrando, P. J., Lorenzo-Seva, U., & Chico, E. (2003). Unrestricted factor analytic procedures for assessing acquiescent responding in balanced, theoretically unidimensional personality scales. *Multivariate Behavioral Research*, *38*(2), 353–374. https://doi.org/10.1207/S15327906MBR3803_04

Friborg, O., Martinussen, M., & Rosenvinge, J. H. (2006). Likert-based vs. semantic differential-based scorings of positive psychological constructs: A psychometric comparison of two versions of a scale measuring resilience. *Personality and Individual Differences*, *40*(5), 873-884. https://doi.org/10.1016/j.paid.2005.08.015

Golino, H., & Christensen, A. P. (2023). *EGAnet: Exploratory Graph Analysis – A framework for estimating the number of dimensions in multivariate data using network psychometrics.* R package.

Hinz, A., Michalski, D., Schwarz, R., & Herzberg, P. Y. (2007). The acquiescence effect in responding to a questionnaire. GMS Psycho-Social Medicine, 4.

Hughes, G. D. (2009). The impact of incorrect responses to reverse-coded survey items. *Research in the Schools*, *16*(2).

Kam, C., Schermer, J. A., Harris, J., & Vernon, P. A. (2013). Heritability of acquiescence bias and item keying response style associated with the HEXACO Personality Scale. *Twin Research and Human Genetics*, *16*(4), 790-798.

Kam, C., Zhou, X., Zhang, X., & Ho, M. Y. (2012). Examining the dimensionality of self-construals and individualistic–collectivistic values with random intercept item factor analysis. *Personality and Individual Differences*, *53*(6), 727-733. https://doi.org/10.1016/j.paid.2012.05.023

Knight, R. G., Chisholm, B. J., Marsh, N. V., & Godfrey, H. P. (1988). Some normative, reliability, and factor analytic data for the revised UCLA Loneliness Scale. *Journal of Clinical Psychology*, *44*(2), 203-206. https://doi.org/10.1002/1097-4679(198803)44:2%3C203::AID-JCLP2270440218%3E3.0.CO;2-5

Lewis, J. R. (2018). Comparison of item formats: Agreement vs. item-specific endpoints. *Journal of Usability Studies*, *11*(1).

Lorenzo-Seva, U., Navarro-González, D., & Vigil-Colet, A. (2016). How response bias affects the factorial structure of personality self-reports.

Luthar, S. S., & Zigler, E. (1991). Vulnerability and competence: A review of research on resilience in childhood. *American Journal of Orthopsychiatry*, *6*(1), 6–12. https://doi.org/10.1037/h0079218

Marsh, H. W. (1986). Multidimensional Self Concepts: Do Positively and Negatively Worded Items Measure Substantively Different Components of Self.

Marsh, H. W. (1996). Positive and negative global self-esteem: A substantively meaningful distinction or artifactors?. *Journal of personality and social psychology*, *70*(4), 810-819. https://doi.org/10.1037/0022-3514.70.4.810

Navarro-Gonzalez, D., Vigil-Colet, A., Ferrando, P. J., Lorenzo-Seva, U., & Tendeiro, J. N. (2021). *vampyr: Factor Analysis Controlling the Effects of Response Bias*. https://CRAN.R-project.org/package=vampyr.

Pilotte, W. J., & Gable, R. K. (1990). The impact of positive and negative item stems on the validity of a computer anxiety scale. *Educational and Psychological Measurement*, *50*(3), 603–610. https://doi.org/10.1177/0013164490503016

Podsakoff, P.M., MacKenzie, S.B., Lee, J.Y., & Podsakoff, N.P. (2003). Common method biases in behavioral research: a critical review of the literature and recommended remedies. *Journal of applied psychology*, *88*(5), 879-903. https://doi.org/10.1037/0021-9010.88.5.879

Primi, R., De Fruyt, F., Santos, D., Antonoplis, S., & John, O. P. (2020). True or false? Keying direction and acquiescence influence the validity of socio-emotional skills items in predicting high school achievement. *International Journal of Testing*, *20*(2), 97-121. https://doi.org/10.1080/15305058.2019.1673398

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2), 1-36. https://doi.org/10.18637/jss.v048.i02

Robinson, J. P., Shaver, P. R., and Wrightsman, L. S. (1991). *Measures of social psychological attitudes* (Vol. 1. Measures of personality and social psychological atitudes). Academic Press.

Salazar, M. S. (2015). The dilemma of combining positive and negative items in scales. *Psicothema*, *27*(2), 192–199. https://doi.org/10.7334/psicothema2014.266

Sauro, J., & Lewis, J. (2011). When designing usability questionnaires, does it hurt to be positive? Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2215–2224. https://doi.org/10.1145/1978942.1979266

Savalei, V., & Falk, C. F. (2014). Recovering substantive factor loadings in the presence of acquiescence bias: A comparison of three approaches. *Multivariate behavioral research*, *49*(5), 407–424. https://doi.org/10.1080/00273171.2014.931800

Schriesheim, C. A., & Hill, K. D. (1981). Controlling acquiescence response bias by item reversals: The effect on questionnaire validity. *Educational and psychological measurement*, *41*(4), 1101–1114. https://doi.org/10.1177/001316448104100420

Soto, C. J., John, O. P., Gosling, S. D., & Potter, J. (2008). The developmental psychometrics of big five self-reports: Acquiescence, factor structure, coherence, and differentiation from ages 10 to 20. *Journal of personality and social psychology*, *94*(4), 718-737. https://doi.org/10.1037/0022-3514.94.4.718

Suárez-Alvarez, J., Pedrosa, I., Lozano Fernández, L. M., García-Cueto, E., Cuesta, M., & Muñiz, J. (2018). Using reversed items in Likert scales: A questionable practice. *Psicothema*, *30*(2), 149-158. https://doi.org/10.7334/psicothema2018.33

Valentini, F. (2017). Influência e controle da aquiescência na análise fatorial [Influence and control of acquiesncence in factor analysis]. *Avaliação Psicológica* [Psychological Assessment], *16*(2), 6–12. https://doi.org/10.15689/ap.2017.1602.ed

Valentini, F., & Hauck Filho, N. (2020). O impacto da aquiescência na estimação de coeficientes de validade [Acquiescence impact in the estimation of validity coefficients]. *Avaliação Psicológica* [Psychological Assessment], *19*(1), 1–3. http://dx.doi.org/10.15689/ap.2020.1901.ed

Van Sonderen, E., Sanderman, R., & Coyne, J. C. (2013). Ineffectiveness of reverse wording of questionnaire items: Let's learn from cows in the rain. *PloS one*, *8*(7), e68967. https://doi.org/10.1371/journal.pone.0068967

Weijters, B., & Baumgartner, H. (2012). Misresponse to reversed and negated items in surveys: A review. *Journal of Marketing Research*, *49*(5), 737-747. https://doi.org/10.1509/jmr.11.0368

Weijters, B., Cabooter, E., & Schillewaert, N. (2010). The effect of rating scale format on response styles: The number of response categories and response category labels. *International Journal of Research in Marketing*, *27*(3), 236-247. https://doi.org/10.1016/j.ijresmar.2010.02.004

Woods, C. M. (2006). Careless responding to reverse-worded items: Implications for confirmatory factor analysis. *Journal of Psychopathology and Behavioral Assessment*, *28*(3), 186–191. https://doi.org/10.1007/s10862-005-9004-7

Wong, N., Rindfleisch, A., & Burroughs, J. E. (2003). Do reverse-worded items confound measures in cross-cultural consumer research? The case of the material values scale. *Journal of consumer research*, *30*(1), 72-91. https://doi.org/10.1086/374697

Zhang, X., Noor, R., & Savalei, V. (2016). Examining the effect of reverse worded items on the factor structure of the need for cognition scale. *PloS one*, *11*(6), e0157795. https://doi.org/10.1371/journal.pone.0157795

Zhang, X., & Savalei, V. (2016). Improving the factor structure of psychological scales: The expanded format as an alternative to the likert scale format. *Educational and psychological measurement*, *76*(3), 357–386. https://doi.org/10.1177/0013164415596421

10 Measurement Invariance/Equivalence

If you are familiar with human research, you may have seen some examples of group comparisons. For example, if we want to test the effectiveness of an antidepressant, we may want to compare gender differences (since depression can differ based on gender). However, are you sure that the instrument you are using (e.g., Beck Depression Inventory) has the same structure for men or women?

Another example is in cross-cultural research. We could measure differences in subjective well-being between countries and see if some countries are happier than others. But how can you infer that your results are accurate if you don't know whether you can compare the latent variable scale scores?

> One criterion for comparing scale scores is that the latent variable is understood and measured equivalently across groups/countries. This property is known as measurement invariance or lack of bias. (Svetina et al., 2020).

So, whenever you want to compare test scores between groups in some way, it would be interesting to do an invariance analysis on the scale first. You may have already come across results about IQ, saying that some countries or cultures have higher or lower IQ than others, without even testing measurement invariance.

A test is invariant if participants belonging to different groups, who have the same score on the factor underlying the test, have on average the same score on an observed item (Lubke & Muthén, 2004).

A common statistical method for establishing evidence of measurement invariance is through Multigroup Confirmatory Factor Analysis (CFA-MG). In CFA-MG, we use a hierarchical test set to impose constraints on the parameters of interest across groups.

10.1 Defining Measurement Invariance/Equivalence

A general factor model is defined as

$$\Sigma = \Lambda \Phi \Lambda' + \Theta$$

Where Σ represents the covariance matrix of the observed variables (or items); Λ is the matrix of factor loadings that expresses the degree of association between the vector of latent variables (with associated covariance matrix Φ) to the observed variables; Θ represents the covariance matrix of measurement errors for the observed variables.

Consider that v is the mean structure. Then, the observed variables' means can be represented by $E(Y) = E(v + \Lambda\eta + \epsilon)$, where η is the vector of latent variables. This model has the assumption that $E(\epsilon) = 0$ and $E(\eta) = k = 0$, then $E(X) = v$, where X is the vector of observed variables. This model generalizes to multiple population by permitting separate covariance matrix for each population or group, i.e., $\Sigma^{(g)}$ with mean structure $v^{(g)}, g = 1, ..., G$.

To establish measurement invariance is that if the null hypothesis is $H_0 : \Sigma^{(1)} = \Sigma^{(2)} = ... = \Sigma^{(G)}$ is rejected, a series of tests follows. I list them below.

1. Configural invariance: tests whether the factorial structure is the same across groups (i.e. number of factors, and whether items percent on the same factor). Thus, the number and pattern of parameters are assumed to be equal across groups. Nevertheless, the values of the parameters are assumed to differ withing identification constraints. If configural invariance is not found, this means that the items load in different factors for different groups.

2. Metric invariance: tests whether the factor loadings of items are equal across groups. The null hypothesis states that the pattern and value of factor loadings are equivalent across groups, i.e., $H_0 : \Lambda^{(1)} = \Lambda^{(2)} = ... = \Lambda^{(G)}$. If metric invariance is not found, it means that one or more items of the instrument are being answered with bias by one or more groups. Therefore, inferences of differences between groups may be biased. Although this is an indicator of response bias in the items, the next step is necessary to assume equal comparison between groups.

3. Scalar invariance: in addition to equal factor loadings, it tests whether the intercepts are equal between groups. Thus, the null hypothesis is $H_0 : \Lambda^{(1)} = \Lambda^{(2)} = ... = \Lambda^{(G)}$, and $v^{(1)} = v^{(2)} = ... = v^{(G)}$. If scalar invariance is not found, any differences found between groups are not related to the latent trait, but to the non-equivalence of measurement of the instrument parameters.

4. Strict Invariance: implies that, in addition to loadings and intercepts, residual variances are similar between groups. Residual variance is simply item variance that is not associated with the latent variable you are measuring.

Some authors advocate the need for strict invariance as a condition for comparing group means (Lubke & Dolan, 2003). In practice, this level of invariance is rarely achieved, given that scalar invariance supports comparisons between groups of manifest (or latent) variable means on the latent variable of interest (Hancock, 1997; Svetina et al., 2020, Thompson & Green, 2006). Thus, most authors suggest achieving scalar invariance instead, in order to conclude invariance (Svetina et al., 2020).

10.2 Measurement Invariance for Categorical Variables

Previously, I described the concept of invariance where the distribution of observed variables is assumed to be multivariate normal. However, in Psychology, most of the surveys are binary of ordinal. If we use the multivariate distribution to categorical variables, we might have consequences on parameters, model fit, and cross-group comparisons (Beauducel & Herzberg, 2006; Lubke & Muthén, 2004; Muthén & Kaplan, 1985). To surpass this, we can use other estimators, like the diagonally weighted least squares (DWLS) family of estimators. The categorical measurement invariance goes as follows (Svetina et al., 2020).

Imagine we have a p X 1 vector of observed variables Y, which take ordered values 0, 1, 2, ..., C. For each observed variable Y_j, with $j = 1, 2, ..., p$, it is assumed that there is an underlying continuous latent response variable Y_j^*, which has the value of that determines the observed category of the observed variable Y_j. Y_j^* is related to Y_j by a set of $C+1$ threshold parameters $\tau_j = (\tau_{j0}, \tau_{j1}, ..., \tau_{jC+1})$, where $\tau_{j0} = -\infty$ and $\tau_{jC+1} = \infty$. Thus, the probability that $Y_j = c$ is:

$$P(Y_j = c) = P(\tau_{jc} \leq Y_j^* \leq \tau_{jc+1})$$

For $c = 0, 1, ..., C$. The model for the vector of latent response variables is:

$$Y^* = v + \Lambda\eta + \epsilon$$

where factor loadings and residuals are defined the same as before, v is a vector of latent intercept parameters. In addition, mean and covariance structure of this model is the same: $E(Y^*) = v$, $Cov(Y^*) = \Sigma^* = \Lambda\Phi\Lambda' + \Theta$, where $E(Y^*) = v$ is assumed to be zero for identification purposes.

In the typical scenario, the categorical factor model can be expanded to encompass multiple groups by accommodating distinct thresholds and covariance matrices for the latent response variables within each population. These are denoted as $\tau^{(k)}$) and $\Sigma^{*(k)}$, where k ranges from 1 to K (with $v^{(k)} = 0$ for all k). In a similar vein, for ordinal data, assessments are made for "baseline" invariance, equivalent slopes, and equal slopes and thresholds, which mirror configural, metric, and scale invariance, respectively. To ascertain the viability of these invariance assumptions, both overall and difference chi-square tests are employed.

10.3 Measurement Invariance in R

To run a Multigroup Confirmatory Factor Analysis, we must first install the *lavaan* (Rosseel, 2012) and *semTools* (Jorgensen et al., 2022) packages for analyses, and *psych* (Revelle, 2023) for the database.

```r
install.packages("lavaan")
install.packages("semTools")
install.packages("psych")
```

And tell the program that we are going to use the functions of these packages

```r
library(lavaan)
```

```
This is lavaan 0.6-17
lavaan is FREE software! Please report any bugs.
```

```r
library(semTools)
```

```
###############################################################################

This is semTools 0.5-6

All users of R (or SEM) are invited to submit functions or ideas for functions.

###############################################################################
```

```r
library(psych)
```

```

Attaching package: 'psych'

The following objects are masked from 'package:semTools':

    reliability, skew

The following object is masked from 'package:lavaan':

    cor2cov
```

To run the analyses, we will use the BFI database (Big Five Personality Factors Questionnaire) that already exists in the *psych* package. We will differentiate between genders (1=Male and 2=Female).

```
dat<- psych::bfi
```

We will store the results of our models in an empty matrix called *results*, where we will extract chi-square, degrees of freedom, RMSEA, CFI and TLI.

```
results<-matrix(NA, nrow = 3, ncol = 6)
```

Let's do the analysis with the 5 BFI factors. First we place the model.

```
mod.cat <- "Agre =~ A1 + A2 + A3 + A4 + A5
            Con =~  C1 + C2 + C3 + C4 + C5
            Extr =~ E1 + E2 + E3 + E4 + E5
            Neur =~ N1 + N2 + N3 + N4 + N5
            "
```

10.3.1 Baseline Model

First, let's make the base model (*baseline model*), where there are no restrictions (*constraints*) between the groups.

```
baseline <- measEq.syntax(configural.model = mod.cat,
                          data = dat,
                          ordered = TRUE,
                          parameterization = "delta",
                          ID.fac = "std.lv",
                          ID.cat = "Wu.Estabrook.2016",
                          group = "gender",
                          group.equal = "configural")
```

The function *measEq.syntax* from the semTools package automatically generates the lavaan syntax to perform a confirmatory factor analysis. As can be seen from the baseline model specification, items are treated as ordinals, delta parameterization and Wu and Estabrook's 2016 model identification are used.

Let's then fit the base model.

```
model.baseline <- as.character(baseline)

fit.baseline <- cfa(model.baseline,
                    data = dat,
                    group = "gender",
```

```r
                    ordered = TRUE)
```

Now let's save the results in the matrix we created initially.

```r
results[1,]<-round(
  data.matrix(fitmeasures(fit.baseline,
                          fit.measures = c("chisq.scaled",
                          "df.scaled",
                          "pvalue.scaled",
                          "rmsea.scaled",
                          "cfi.scaled",
                          "tli.scaled"))),
                      digits=3)
```

10.3.2 Thresholds Invariance Model

```r
prop4 <- measEq.syntax(configural.model = mod.cat,
                       data = dat,
                       ordered = TRUE,
                       parameterization = "delta",
                       ID.fac = "std.lv",
                       ID.cat = "Wu.Estabrook.2016",
                       group = "gender",
                       group.equal = c("thresholds"))

model.prop4 <- as.character(prop4)

fit.prop4 <- cfa(model.prop4,
                 data = dat,
                 group = "gender",
                 ordered = TRUE)

results[2,]<-round(data.matrix(
  fitmeasures(fit.prop4,
             fit.measures = c("chisq.scaled",
             "df.scaled",
             "pvalue.scaled",
             "rmsea.scaled",
             "cfi.scaled",
             "tli.scaled"))),
               digits=3)
```

To examine the relative fit of the model and compare the chi-square statistics between the baseline model and the model where threshold constraints are employed, we use function lavTestLRT.

```
lavTestLRT(fit.baseline,fit.prop4)
```

```
Scaled Chi-Squared Difference Test (method = "satorra.2000")

lavaan NOTE:
    The "Chisq" column contains standard test statistics, not the
    robust test that should be reported per model. A robust difference
    test is a function of two standard (not robust) statistics.

              Df AIC BIC  Chisq Chisq diff Df diff Pr(>Chisq)
fit.baseline 328         3894.5
fit.prop4    388         3918.2     56.415      60     0.6075
```

10.3.3 Thresholds and Factor Loadings Invariance Model

```
prop7 <- measEq.syntax(configural.model = mod.cat,
                       data = dat,
                       ordered = TRUE,
                       parameterization = "delta",
                       ID.fac = "std.lv",
                       ID.cat = "Wu.Estabrook.2016",
                       group = "gender",
                       group.equal = c("thresholds", "loadings"))

model.prop7 <- as.character(prop7)

fit.prop7 <- cfa(model.prop7,
                 data = dat,
                 group = "gender",
                 ordered = TRUE
)
```

```
results[3,] <- round(
  data.matrix(
  fitmeasures(fit.prop7,
  fit.measures = c("chisq.scaled",
  "df.scaled",
  "pvalue.scaled",
  "rmsea.scaled",
  "cfi.scaled",
  "tli.scaled"))),
  digits = 3)

colnames(results) <- c("chisq","df","pvalue","rmsea","cfi","tli")
rownames(results) <- c("baseline","thresholds","loadings")
```

Examining fit indices (*results*):

```
print(results)
```

```
              chisq  df pvalue rmsea   cfi   tli
baseline   4111.951 328      0 0.096 0.867 0.846
thresholds 4372.509 388      0 0.091 0.860 0.863
loadings   4120.539 404      0 0.086 0.869 0.877
```

we noticed that in general, the model fit improved as the models became more restricted by imposing equality of thresholds (prop4) and equality of factor loadings (prop7).

We can perform the chi-square difference test between the threshold invariance (prop4) and threshold + factor loadings (pro7) models to assess the feasibility of measurement invariance.

```
lavTestLRT(fit.prop4,fit.prop7)
```

```

Scaled Chi-Squared Difference Test (method = "satorra.2000")

lavaan NOTE:
    The "Chisq" column contains standard test statistics, not the
    robust test that should be reported per model. A robust difference
    test is a function of two standard (not robust) statistics.

           Df AIC BIC  Chisq Chisq diff Df diff Pr(>Chisq)
fit.prop4 388         3918.2
fit.prop7 404         3953.0      21.36      16     0.1651
```

10.4 Data Interpretation

For interpretation, you can see Table 1 of the study by Svetina et al., (2020). In it, there is a summary of several simulation studies and which cutoff points are recommended. For example, if you have an ordinal distribution, you are comparing two groups, with each group having 150/150 or 150/500 or 500/500 participants, having 2 to 4 factors, you should test the levels of measurement invariance through of the chi-square difference with $p < 0.05$ between each level of invariance. If you consider that you have data with normal distribution, you are comparing 2 groups, you have an N of 150, 250 or 500 per group, and you are comparing only 1 factor, the difference between each CFI invariance level should not be greater than or equal to 0.005, while the RMSEA difference must not be less than or equal to 0.010.

10.5 References

Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. Structural Equation Modeling: A Multidisciplinary Journal, 13(2), 186–203. https://doi.org/10.1207/s15328007sem1302_2

Hancock, G. R. (1997). Structural equation modeling methods of hypothoesis testing of latent variable means. *Measurement & Evaluation in Counseling & Development (American Counseling Association)*, *30*(2), 91–105. https://doi.org/10.1080/07481756.1997.12068926

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2022). *semTools: Useful tools for structural equation modeling.* R package. Retrieved from https://CRAN.R-project.org/package=semTools

Lubke, G. H., & Muthén, B. O. (2004). Applying multigroup confirmatory factor models for continuous outcomes to Likert scale data complicates meaningful group comparisons. *Structural equation modeling*, *11*(4), 514-534. https://doi.org/10.1207/s15328007sem1104_2

Muthén, B., & Kaplan, D. (1985). A Comparison of Some Methodologies for the Factor Analysis of Non-Normal Likert Variables. British Journal of Mathematical and Statistical Psychology, 38, 171–189. https://doi.org/10.1111/j.2044-8317.1985.tb00832.x

R Core Team (2023). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Revelle, W. (2023). *psych: Procedures for Psychological, Psychometric, and Personality Research.* Northwestern University, Evanston, Illinois. R package. https://CRAN.R-project.org/package=psych

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2), 1-36. https://doi.org/10.18637/jss.v048.i02/

Svetina, D., Rutkowski, L., & Rutkowski, D. (2020). Multiple-group invariance with categorical outcomes using updated guidelines: an illustration using M plus and the lavaan/semtools packages. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(1), 111-130. https://doi.org/10.1080/10705511.2019.1602776

11 Non-Parametric Item Response Theory

In psychometrics, with the advent of computers and new ways of carrying out statistical analyses, it has become a consensus to use techniques such as Factor Analysis and/or Item Response Theory to verify the validity of a given psychological instrument (Franco et al., 2022). Parametric tests are generally used, both Factor Analysis and Item Response Theory. But what are parametric tests and non-parametric tests?

11.1 (Non)-Parametric Tests

Parametric techniques are based on statistical models that place constraints. For example, in a Pearson correlation, we impose the restriction that the relationship between two variables is linear. The same occurs in a Factor Analysis, which assumes a linear relationship between the construct and the items. Therefore, if there is a relationship between the variables, but this relationship is not linear, the models will underestimate or overestimate the magnitudes of such relationships. That is why so-called non-parametric techniques emerged, which are techniques that do not make specific restrictions on the behavior of variables.

11.2 Mokken Scaling

Mokken Scale Analysis can be thought of as a Non-Parametric Item Response Theory technique as it does not assume the exact form of the item response function. It starts and **tests** the same three assumptions that are common to parametric Item Response Theory:

- Unidimensionality: only one latent trait of individuals interacts with a latent characteristic of items. In other words, only one latent trait explains the behavior of the items. In the realm of Item Response Theory (IRT), the principle of unidimensionality posits that a single latent trait, denoted as θ, suffices to explain the underlying structure of the data. The primary rationale behind this lies in the preference of practical researchers for measurement instruments that focus on capturing a single trait at a time. This approach streamlines the interpretation of test results, making them more manageable and comprehensible.

- Local independence: is the idea that when items are conditioned on the latent trait, they should not correlate. In other words, an individual's response to item i is not influenced by his or her responses to the other items in the same test. Local independence is as follows. Imagine that $X = (X_1, X_2, ..., X_k)$ is a vector that contains item scores variables, and $x = (x_1, x_2, x_k)$ is a realization of X. In cases we deal with dichotomous items, each $x_i = 0$ or $x_i = 1$. The probability of an individual to have a score x_i on item i, given the latent trait leval θ, is $P(X_i = x_i|\theta)$. Thus, local independence means that

$$P(X = x|\theta) = \prod_{i=1}^{k} P(X_i = x_i|\theta)$$

One implication of local independence is that when the latent trait θ is held constant, the covariance between items i and j is zero: $Cov(X_i, X_j|\theta) = 0$. However, in a group where θ varies, the covariance is positive $(Cov(X_i, X_j) > 0)$ because the items are measuring the same underlying trait. Nonetheless, this covariance disappears when θ is fixed because it is entirely accounted for by θ.

- Latent Monotonicity: when the individual has a greater latent trait, the greater the probability of giving the correct answer or giving the highest answer on a scale. In other words, the higher the level of knowledge in mathematics, the more likely a person is to get an SAT math question right. These assumptions posit that that the conditional probability $P_i(\theta)$ is monotonely nondecreasing in θ, which means:

$$P_i(\theta_a) \leq P_i(\theta_b)$$

- Non-intersection of Item Response Functions: means that item response functions should not intersect. A quick note is that, the success probability for a fixed item depends on a person's ability (or trait level) and is called its item response function (IRF); it is usually assumed that this IRF increases if the person has more of the latent trait. Thus, this assumption says that the k item response functions are non-intersecting across θ. Non-intersection means that all item response functions can be ordered and numbered such that:

$$P_1(\theta) \leq P_2(\theta) \leq ... \leq P_k(\theta), \text{for all } \theta.$$

As consequence for this formula, Item 1 is the most difficult item, followed by Item 2, and so on.

From these assumptions, two Mokken Scale Analysis models are derived.

1. Monotonic Homogeneity Model (Mokken, 1971): which respects the first three assumptions (unidimensionality, local independence and latent monotonicity). Thus, the Monotonic Homogeneity Model is an Item Response Theory model for measuring persons on an ordinal scale.

2. Double Monotonicity Model: which respects the four assumptions of unidimensionality, local independence, latent monotonicity and non-intersection. It is a special case of 1.

These two models have a difference. The first model (monotonic homogeneity) allows ordering only the respondents, while the second (double monotonicity) allows ordering both individuals and items (thus allowing the ordering of items by level of difficulty; Sijtsma & Molenaar, 2002).

11.3 Scalability Coefficient

To test the assumptions of the monotonic homogeneity and double monotonicity models, the main index used is the Loevinger scalability coefficient H (Loevinger, 1948). There are three scalability indices:

- the item pair index (H_{ij}):
$$H_{ij} = \frac{COV(X_I, X_j)}{COV(X_I, X_j)^{max}}$$
Where X_i is the sum score of item i, X_j is the sum score of item j, and the superscript *max* indicates the maximum covariance that the two items can have if they have no Guttman errors.
- the item index (H_j):
$$H_j = \frac{COV(X_j, R_{-j})}{COV(X_j, R_{-j})^{max}}$$
Where X_j is the sum score of item j, R_{-j} is the and the remainder score (*rest score*) when disregarding item j (i.e., the sum score of all the items minus item j), and the superscript *max* indicates the maximum covariance that the two items can have if they do not have Guttman errors.
- the overall test index (H):
$$H = \frac{\sum_{j=1}^{J} COV(X_j, R_{-j})}{\sum_{j=1}^{J} COV(X_j, R_{-j})^{max}}$$
Where X_j is the sum score of item j, R_{-j} is the and the remainder score (*rest score*) when disregarding item j (i.e., the sum score of all the items minus item j), and the superscript *max* indicates the maximum covariance that the two items can have if they do not have Guttman errors.

The H coefficient indices can vary between -1 or +1, and the assumptions of unidimensionality, local independence and latent monotonicity imply: $0 \leq H_{ij} \leq 1$, for all $i \neq j$; $0 \leq H_j \leq 1$, for all j; and $0 \leq H \leq 1$. Thus, if all assumptions are respected, the observed values of the H indices should not be less than 0. Of course, it is possible to observe negative values when the items are not suitable for the scale (Sijtsma & Molenaar, 2002). All this means that the calculation of Guttman scalability coefficients is both descriptive and also serves predictive purposes for the quality of the measurements, allowing for more robust inferences (Franco et al., 2022).

11.4 Mokken Scaling in R

To do this, we will use the *mokken* package (van der Ark, 2012). First, let's install the package on the computer.

```
install.packages("mokken")
```

Then, we will inform the program that we are going to use the functions of this package.

```
library(mokken)
```

```
Carregando pacotes exigidos: poLCA
```

```
Carregando pacotes exigidos: scatterplot3d
```

```
Carregando pacotes exigidos: MASS
```

We will use a database available in the *mokken* (Van der Ark, 2007) package, with responses to 12 dichotomous items administered to 425 children from grades 2 to 6 in the Netherlands (Verweij, Sijtsma & Koops, 1996). Each item is a transitive reasoning task about physical properties of objects, with two items used as pseudo-items (items 11 and 12), four items about length relations (items 01, 02, 07 and 09), five items about width relations (items 03, 04, 05, 08 and 10) and an item related to area relations (item 06).

```
data(transreas)

data <- transreas[,2:ncol(transreas)] # Select only the test items
```

11.4.1 Dimensionality Analysis in R

In Mokken Scale Analysis, we do not perform dimensionality analysis with techniques such as Parallel Analysis. This is done through the automatic item selection procedure (AISP, *Autometed Item Selection ProcedureI;* Mokken, 1971). The AISP uses the scalability coefficient H_i to select the most representative item of the dimension and then the item pair scalability coefficient to select the largest subset of items that measure the same construct (Mokken, 1971). Then, after selecting the best items for the first dimension, unselected items are tested to try to compose a second subscale, and so on, until it is no longer possible to allocate any item to any subscale. The scalability coefficient of pairs of items should not be less than 0.30 (Straat te al., 2013), and it is recommended to use the genetic algorithm (in the code, `search="ga"`). The following table presents all the items in the rows and the minimum values of the scalability coefficient (H_j) of the best item represented in the columns.

```
AISP <- aisp(data, # Items
             search="ga", # Genetic Algorithm
             lowerbound=seq(.3,.8,by=.05) # Which H to show
             )

# Print Results
print(AISP)
```

```
     0.3 0.35 0.4 0.45 0.5 0.55 0.6 0.65 0.7 0.75 0.8
T09L   1    1   1    1   1    1   2    1   1    2   1
T12P   0    0   0    0   0    0   0    0   0    0   0
T10W   1    1   1    1   1    2   1    0   0    0   0
T11P   0    0   0    0   0    0   0    0   0    0   0
T04W   2    2   0    0   0    0   0    0   0    0   0
T05W   0    0   0    0   0    0   0    0   0    0   0
T02L   2    2   0    0   0    0   0    0   0    0   0
T07L   1    1   1    1   1    1   0    2   2    1   0
T03W   1    1   1    1   1    2   0    0   0    0   0
T01L   1    1   1    1   0    0   0    0   0    0   0
T08W   1    1   1    1   1    1   2    1   1    2   1
T06A   1    1   1    1   1    1   1    2   2    1   0
```

It can be seen that the pseudo-items did not aggregate into any subscale. In general, AISP identified that, at most, two scales can be generated, represented by the numbers 1 and 2. The number 0 means that, given that coefficient H, the item does not form any scale.

Let's save Scale 1 formed by the coefficient H = 0.45, given that this subscale is constant up to the limit of 0.45. This means that a robust scale can be created using items 1, 3, 6, 7, 8, 9 and 10. This way, pseudo-items 11 and 12 would be discarded, in addition to items 02, 04 and 05, which probably present more Guttman errors than would be expected for one-dimensional items. The second scale does not present consistency when varying the limits of the scalability coefficient, which may indicate that it is a spurious scale.

Let's save the new scale in a variable for subsequent analyses.

```
scale1 <- data[,colnames(data)[which(AISP[,"0.45"] == 1)]]
```

11.4.2 Latent Monotonicity Analysis in R

Using only the items that were maintained after AISP, the following table presents the items, the scalability indices for each item (H_j), the number of active pairs (PA)—which represents the maximum possible number of tests of monotonicity for each item—, the number of monotonicity violations (Vi) that were identified for each item, the magnitude of the largest violation (MaxVi), the z value of this largest violation (Zmax) for inferential testing, and the number of violations which were significant in each item (Zsig).

```
MonLat <- check.monotonicity(scale1)

summary(MonLat)
```

```
     ItemH #ac #vi #vi/#ac maxvi sum sum/#ac zmax #zsig crit
T09L  0.50   1   0       0     0   0       0    0     0    0
T10W  0.52   3   0       0     0   0       0    0     0    0
T07L  0.51   3   0       0     0   0       0    0     0    0
T03W  0.53   3   0       0     0   0       0    0     0    0
T01L  0.46   3   0       0     0   0       0    0     0    0
T08W  0.55   1   0       0     0   0       0    0     0    0
T06A  0.59   0   0     NaN     0   0     NaN    0     0    0
```

We see that item 6 (T06A) does not present adequate variability in scores, so it has little information about the respondents' scores, so it should be excluded. Let's save the remaining items in a new variable.

```
scale2 <- scale1[,-7]
```

11.4.3 Local Independence in R

```
CA <- check.ca(scale2)

print(CA)
```

```
[[1]]
[1] TRUE TRUE TRUE TRUE TRUE TRUE
```

The result is a vector of booleans (TRUE or FALSE) with length equal to the number of items. If TRUE, it indicates that the item is still on the scale, if FALSE, it indicates that the item should be removed. None of the items were considered outliers, that is, all items were maintained.

11.4.4 Non-intersection Analysis of Item Response Functions

Using only the items that were maintained by the AISP and the monotonicity analysis, the analysis presented in the following table was performed.

```
NI <- check.pmatrix(scale2)
summary(NI)
```

```
     ItemH #ac #vi #vi/#ac maxvi sum sum/#ac zmax #zsig crit
T09L  0.49  20   0       0     0   0       0    0     0    0
T10W  0.51  20   0       0     0   0       0    0     0    0
T07L  0.49  20   0       0     0   0       0    0     0    0
T03W  0.53  20   0       0     0   0       0    0     0    0
T01L  0.47  20   0       0     0   0       0    0     0    0
T08W  0.55  20   0       0     0   0       0    0     0    0
```

No item was found to be in violation, so all were kept for the next analysis.

11.5 References

Franco, V. R., Laros, J. A., & Bastos, R. V. S. (2022). Theoretical and practical foundations of Mokken scale analysis in psychology. *Paidéia (Ribeirão Preto)*, *32*. https://doi.org/10.1590/1982-4327e3223

Loevinger, J. (1948). The technique of homogenous tests compared with some aspects of "scale analysis" and factor analysis. *Psychological Bulletin*, *45*(6), 507-530. https://doi.org/10.1037/h0055827

Mokken, R. J. (1971). *A theory and procedure of scale analysis*. De Gruyter. https://doi.org/10.1515/9783110813203

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. Sage. https://doi.org/10.4135/9781412984676

Straat, J. H., van der Ark, L. A., & Sijtsma, K. (2013). Comparing optimization algorithms for item selection in Mokken scale analysis. *Journal of Classification*, *30*, 72-99. https://doi.org/10.1007/s00357-013-9122-y

Van der Ark LA (2007). Mokken Scale Analysis in R. *Journal of Statistical Software*, *20*(11), 1-19. https://doi.org/10.18637/jss.v020.i11.

Verweij, A. C., Sijtsma, K., & Koops, W. (1996). A Mokken scale for transitive reasoning suited for longitudinal research. *International Journal of Behavioral Development*, *19*(1), 219-238. https://doi.org/10.1177/016502549601900115

12 How to Validate Psychological Tests: A Step by Step Guide

This chapter aims to help you understand the steps involved in constructing and validating psychological tests. As we know, psychological tests are important instruments for evaluating skills, behaviors and psychological traits. However, to ensure that tests are valid and reliable, it is necessary to follow a rigorous and systematic construction and validation process. This guide will provide an overview of the steps involved in creating and validating psychological tests. I hope this 11-step guide gives you a thorough understanding of the process of creating and validating psychological tests and helps you develop high-quality instruments to assess psychological skills and traits.

12.1 Define the Objective of the Test

The test objective must be clearly defined before the test construction process begins. To do this, you must have the following questions in mind: What do you want to evaluate? What are the behaviors, skills or psychological traits that we intend to measure? Why are you trying to measure this? What are the theoretical and practical implications of having a measure of this?

These questions are crucial to have guidance on where to start looking, researching, and immersing yourself in the subject. Sometimes we want to develop an instrument out of necessity. For example, in one research, I developed an instrument to measure how much people felt they suffered prejudice and discrimination. Why did I create this instrument? 1. I wanted to study the impact of this phenomenon on people's self-esteem and well-being, so, as this instrument didn't exist yet, I had to create it. 2. Because studying this subject was important to me. As I always say, we put a little bit of ourselves into the research, and the choice of the topic came from me for a reason. So, you can use some justifications for creating instruments. The need to measure why it will impact X, Y or Z and a shortage of instruments is the most common.

12.2 Literature Review

At this stage, a literature review must be carried out to verify which measurement instruments exist that assess the same construct or psychological trait, in order to avoid duplication of efforts. You can follow some tips, such as:

1. Select relevant databases: it is important to search for articles in relevant databases in the area of interest, such as PsycINFO, Scopus, Web of Science, Scielo, PubMed, among others.
2. Use appropriate keywords: it is important to use keywords appropriate to the area of interest to obtain more accurate results. Some examples of keywords are "psychological instrument", "validation", "psychometrics", "construction", "scale", among others.
3. Carry out a systematic search: it is important to carry out a systematic search for articles, following a pre-defined strategy, such as the use of Boolean operators (AND, OR) and the inclusion or exclusion of specific terms. In other words, when you search for a racism instrument, you will use "scale" AND "racism" OR "validation" AND "racism", and so on. Using these operators makes a filter in article search engines. They will select when the two words appear together (AND) or when there is one or the other word (OR).
4. Analyze the search results: it is important to analyze the articles found and identify the instruments already validated and used in the area of interest, as well as the gaps and limitations of existing instruments. Furthermore, it is important to read and evaluate the selected articles to see the methodological quality of the studies, the validity and reliability of the instruments used and the authors' conclusions.
5. Synthesize the information: it is important to synthesize the information obtained in a narrative review or in a table that presents the characteristics of the identified instruments, such as the construct assessed, the target population, the number of items, fit indices, number of factors , among other relevant information.

12.3 Should I Build or Adapt an Instrument?

The decision to build or adapt a psychological instrument depends on several factors, such as the research objectives, the characteristics of the target population, the construct or variable to be evaluated, the cultural and linguistic context, among others.

When to build a psychological instrument:

- When there is no validated instrument for the construct or variable of interest;
- When the target population you want to measure presents specific characteristics that are not covered by existing instruments;
- When the construct or variable of interest is multidimensional and existing instruments do not cover all relevant dimensions;
- When a specific approach or theory is sought to evaluate the construct or variable of interest.

When adapting a psychological instrument:

- When there is already a validated instrument for the theoretical construct you thought of or variable of interest, but it was developed in another language or culture;
- When the intention is to use an instrument validated in another language or culture, but it is necessary to make adaptations to guarantee semantic, conceptual, cultural and linguistic equivalence;

It is important to highlight that the adaptation of a psychological instrument requires specific methodological care to guarantee semantic, conceptual, cultural and linguistic equivalence, in addition to verifying the validity and reliability of the adapted instrument. The construction of a psychological instrument requires a solid theoretical and empirical foundation to guarantee the validity and reliability of the constructed instrument.

12.4 Which Model Should I Choose?

This is also a crucial step, which is almost never thought of when building a psychological instrument. When we do a test, we will apply this test to a specific model. For example, we have the factor model (which is most commonly used to develop instruments), principal components, the network model, the latent profile model, and so on.

I want you to think critically while building the instrument. Before applying a test to a model, you must think about which model you want to test your data on, and only then test the model.

12.5 Think About How the Participant Will Answer the Questionnaire

Another crucial step, which is almost never thought of when building a psychological instrument, is the questionnaire response scale. For example, when presenting the item, should the participant select only one option? Should you order the options? And so on.

We have some possible response scales that are important to mention.

1. Likert scales: on this scale, participants select the degree of agreement or disagreement with a given statement. For example, a Likert scale can be used to assess how much the person agrees with the item "I am a communicative person.", with the options: completely agree, partially agree, neither agree nor disagree, partially disagree, completely disagree.

2. Forced-Choice: The researcher offers the respondent, for example, three or four response options (items) organized into blocks. Then, the participant must order the items according to what is most frequent (within that block) and least frequent. Therefore, you will have a hierarchy of what represents that person most.

3. Expanded Format: It is a mix of the Likert format and the forced-choice format. You present, for example, 5 answer options, where the participant must select only one. However, instead of being in degree of agreement, each gradation is a specific item, which varies in degree of intensity. For example.

    1. I hate other people.
    2. I don't like other people.
    3. I do not like or dislike other people.
    4. I generally like other people.
    5. I love others'.

4. Frequency scale: on this scale, participants select the frequency with which a certain behavior, thought or feeling occurs. For example, a frequency scale can be used to assess how often someone has experienced a certain symptom in the last seven days, with the options: never, rarely, sometimes, often, always.

5. Semantic Differential: The format of a semantic differential scale is usually presented in the form of a vertical line, where the concept in question is placed in the middle of the line. At each end of the line, opposite adjectives are presented, such as "good" and "bad", "positive" and "negative", "strong" and "weak", among others. The participant is asked to mark a point on the vertical line that best represents their attitude towards the concept in question, according to the position of the opposing adjectives.

For example, in a psychological test that assesses attitudes towards school, a semantic differential scale can be used, presenting the word "school" in the middle of the line and the adjectives "fun" and "boring" at the ends. The participants would be asked to mark a point on the vertical line that best represents their opinion of the school.

12.6 Item Construction

Based on the definition of the test objective and the literature review, a series of items must be constructed that assess the psychological construct or trait that is intended to be measured. It is important that the items are clear, precise and relevant to the objective of the test. I'll talk about some item construction tips below.

1. Create items that can be answered by all participants you will collect: Items must be formulated in such a way that they can be answered by all participants, regardless of their educational or cultural background. Therefore, always keep in mind the sample you are going to collect, otherwise you may be biasing the instrument towards one group compared to another.

    a. Consider cultural sensitivity: When formulating items, take cultural differences into account and avoid including questions that may be offensive or inappropriate for certain groups.

2. Enter only one idea per item. Sometimes, I have dealt with items that wanted to measure extroversion and agreeableness at the same time, and this is very confusing when responding, in addition to confusing the interpretation of the scores. See the example of the item "I am a communicative person, who likes to help others.". If the person agrees with this item, they may agree with "being communicative", "helping others" or both. And how do we interpret this? There's not much of a way. This will also mess with the respondent's head. Anyway, avoid it!
3. Use clear and simple language: Items should be written in a clear and easy-to-understand way for participants, without complicated and technical words (such as, undoubtedly; etc.) or jargon. Use slang only if it is appropriate for the target population.

    a. Within this suggestion there is a suggestion that is a maxim of psychometrics. Avoid using the word "No" as much as possible (for example, instead of using "I'm not sad", use the antonym "I'm happy"). This is because a negative is more difficult to understand than a more direct item.

4. Avoid suggestive or biased questions: items should not suggest an answer or include words that could lead the participant to a specific answer.
5. Do not use adverbs of intensity, as this can make the item more difficult to agree or disagree with in an artificial way. For example, instead of using "I really like Black Sabbath", just use "I like Black Sabbath". See, if a person completely agrees with the item "I like Black Sabbath", it means they like it a lot. If she responds that she disagrees with the item "I really like Black Sabbath", she may still like Black Sabbath, but she doesn't agree that she likes it that much.
6. Include control items or "gotchas": To ensure that participants are paying attention and are not simply choosing random answers, include items that check whether they are carefully reading the questions. For example, use the item "Mark alternative 5 on the response scale".
7. Vary the content of the items: Items should cover a variety of content or aspects of the construct being measured in order to obtain a more comprehensive measure. So, when measuring extroversion, you don't want to measure just the number of times the person speaks, that is, that they are communicative. You can measure things like feeling better about social interactions, frequency of social interactions, and so on.

    a. Include negative items: Include items that assess both positive and negative behaviors in order to obtain a more complete and accurate measure of the construct being measured. In other words, if we think about extroversion: add items that the more the person agrees with, the greater their extroversion (such as "I like going out with large groups of people"); and also items that the more the person agrees with, the lower their extroversion (such as "I avoid leaving the house with people I don't know").

8. Perform an item analysis: After constructing the items, perform an item analysis to determine whether they are consistent with the underlying theory of the construct being measured and whether they are correlated with other items in a coherent manner.

9. Consider the extent of the test: take into account the extent of the test and the time required for its application, seeking to create an instrument that is sufficiently comprehensive and can be applied in a reasonable time.

12.7 Judges' Assessment and Item Writing

The evaluation of judges is an important stage in the process of constructing a psychological instrument. This assessment involves the review of the instrument's items by subject matter experts, with the aim of identifying possible problems or limitations of the items.

To carry out this assessment, it is recommended to follow the following steps:

1. Selection of judges: select a group of experts on the subject, who have theoretical knowledge and/or practical experience in the area of the construct being measured.
2. Submission of items: send the instrument items to the judges, along with instructions for evaluating and defining the construct you are measuring.
3. Item evaluation: ask the judges to evaluate the instrument's items, checking whether they are clear, objective, relevant and suitable for measuring the construct being evaluated.
4. Judges' feedback: request that judges provide detailed feedback on the instrument's items, indicating which items need to be modified or deleted, as well as which items could be added to improve the construct measurement.
5. Analysis of results: analyze the results of the judges' evaluation and use the information obtained to make adjustments to the instrument's items. There are some statistical analyzes that can be used to evaluate the agreement between judges in the evaluation stage of the items of a psychological instrument. Some of these analyzes include:

    1. Content Validity Coefficient (CVC): the CVC is a statistical index that measures the agreement between judges regarding the content validity of the items. This coefficient is calculated based on the judges' assessment of the following categories: (a) clarity of language (which consists of analyzing the language used in the items, taking into account the characteristics of the responding population); (b) practical relevance (which aims to assess whether the item is in fact important for the instrument); and (c) theoretical relevance, which involves analyzing the association between the item and the theory. It ranges from 0 to 1, with higher values indicating greater agreement between judges.
    2. Item Validity Index (IVI): IVI is an index that measures agreement between judges regarding the validity of items. It also varies from 0 to 1, with higher values indicating greater agreement between judges.
    3. Intraclass Correlation Coefficient (ICC): the ICC is an index that measures the agreement between judges in relation to the answers given by participants in the instrument. It ranges from -1 to 1, with higher values indicating greater agreement between judges.

4. Fleiss Kappa Coefficient: Fleiss Kappa is an index that measures the agreement between judges in relation to the answers given by participants in categorical items. It ranges from 0 to 1, with higher values indicating greater agreement between judges.

It is important that you, as the responsible researcher, carry out a careful analysis of the results of the judges' evaluation, comparing the judges' responses and identifying the points where there is agreement or disagreement. Furthermore, you must take into account the feedback provided by the judges and make the necessary adjustments to the instrument, always considering the ethical and professional guidelines in force. When possible, rewrite the items as recommended by the evaluators.

12.8 Target Population Assessment

The evaluation of a psychological instrument with the target population is a fundamental step in the instrument validation process. This evaluation aims to test whether the instrument is understandable, relevant and reliable for the population it is intended to evaluate.

To carry out this assessment, I suggest the following steps:

1. Define the sample: define the sample of participants who will be invited to participate in the study. This sample must be representative of the target population of the instrument and must be large enough to guarantee the statistical validity of the results.
2. Select participants: select participants according to the previously established inclusion and exclusion criteria. It is important to ensure that selected participants are able to understand and respond to the instrument. To do this, you can show the instrument during an interview, which will make you available to answer any questions you may have.
3. Ask the representative sample questions about the phenomenon you want to build an instrument for. Create an interview guide that seeks to verify the suitability of that construct for that target population. For example, you may want to measure racism suffered by black people, which is probably different from the racism suffered by indigenous people, and so on. Therefore, try to better understand how that construct works for the target population you thought of. If nothing is in line with what you thought, I recommend you review your theory and your instrument. Always listen to your participants!
4. Ask participants if they understood the items, and if there is any doubt, flag the item, explain what you meant and ask what the best way would be to ask that question to that population.

12.9 Internal Structure Assessment: Training Stage

1. This step will collect new data, as it aims to test whether the theoretical structure holds in your data. The evaluation of the internal structure of a psychological instrument is a process that aims to identify and verify the consistency of the dimensions or factors underlying the construct or variable of interest. This is a crucial step in validating an instrument.

Here, you will send your instrument to your sample, and ask them to respond to the scale you built. Of course, this step serves some specific models, including the unrestricted common factor model, principal component analysis and latent profile analysis.

1. First, you will estimate the dimensionality: either through parallel analysis, Exploratory Graph Analysis, or Eigenvalue > 1 (I don't recommend the latter). In latent profile analysis you will adhere to other methods and heuristics. They usually do trial and error, adding a factor and testing the fit of the model.
2. Once you know the number of dimensions, you will apply Exploratory Factor Analysis or other Exploratory Graph Analysis methods, or Latent Profile Analysis. Here, the aim is to see the relationship of the item with the construct (i.e., to see the factor or component loadings and where the items clustered [in the case of the factor and principal components model]; or to see the probability of the item belong to a certain class [in the case of latent profiles]).
3. The third stage involves refining the instrument. To do this, you will delete bad items:

a. Those who presented low loads/probabilities of belonging;
b. exclude items that have high response bias, if you have collected data to do so;
c. exclude item that has many loadings or crossed probabilities, that is, it has loadings or probabilities of belonging to more than one factor at the same time
d. Exclude items that were not grouped into factors in a pertinent way.

4. Finally, you evaluate the possibility of rewriting any bad items

12.10 Internal Structure Assessment: Validation Stage

In this step, you will collect new data to verify that the results of the training step replicate in a new sample, the validation sample. See, in this case, training and validation is a term used in other areas of knowledge, such as machine learning. So, the term validation here is not the same as validation in psychology.

Data collection should be in the same way as the previous step, with some minimal adjustments. In general, adjustments should be made to remove more noise from the data. Such as making small adjustments to bad items when possible, or improving the way you collect.

Anyway, here you will apply Confirmatory Factor Analysis, with the theoretical and empirical structure in mind. If the empirical structure from the previous step is different from the structure you theorized, you can see the difference between the fit indices. Just be careful because if the model is not nested, the models are not directly comparable.

In the case of other models, such as principal components, latent profiles and networks, this sample serves the same purpose: checking whether the results are replicated. You just won't apply a Confirmatory Factor Analysis.

12.11 Other Evidence of Validity

Here, you will collect new data to check three types of validity (you can do 1 study for the following validities, or 1 study for each validity below. It's up to you).

1. Convergent Validity: Convergent validity refers to the extent to which an instrument is correlated with other measures that theoretically should be related to the construct being measured. To do this, you will collect data to check if your instrument correlates with other scales that measure THE SAME THING as you; and collect data from other instruments that measure things that should be correlated but are different constructs. Example: Consider an instrument that was developed to measure social anxiety in adolescents. To assess convergent validity, you can apply the instrument together with other previously validated social anxiety measures, in addition to applying Life Satisfaction and Neuroticism instruments. If scores on the new instrument are highly correlated with scores on already validated measures, this suggests that the instrument has high convergent validity.
2. Divergent Validity: Divergent validity refers to the extent to which an instrument is not correlated with other measures that theoretically should not be related to the construct being measured. In other words, the lower the correlations between the instrument and other unrelated measures, the greater the divergent validity. Continuing with the example above, the researcher can assess the divergent validity of the instrument by measuring its correlation with measures that should not be related to social anxiety, such as the professional interests scale or the racism scale. If scores on the new instrument are not correlated with scores on these measures, this suggests that the instrument has high divergent validity.

12.12 Extra Observations

Equity: It is important to ensure that the test is equitable in relation to differences between some groups, such as cultural, gender, education, socioeconomic equity, etc. This may include adapting items for different groups or carefully evaluating the effects of sensitive group items.

One way to do this is to exclude items that may be more difficult for a given gender or to check the equity of scores between genders through invariance analysis. Another example is using items that do not depend on specific prior knowledge or financial resources. Furthermore, when the test is administered in different languages, it is important to ensure that the translation is accurate and that linguistic differences between languages do not affect the results.

Reliability: The reliability of the test must be rigorously evaluated, so that the results are consistent and reliable, whether over time, or between different evaluators, or even in a single application. Examples of single-application reliability analysis are Cronbach's Alpha, McDonald's Omega, Greatest Lower Bound, H Coefficient, and so on.

Accessibility: Accessibility of psychological tests is an important concern, and it is necessary to ensure that tests are accessible to people with physical or sensory disabilities.

Examination of response bias: Participants may respond in a biased or desirable manner on a psychological test, which can affect the validity of the results. It is important to examine response bias in a test and take steps to minimize it, such as including control items or reversed items.

Well, I must point out that there may be more data collection involved, especially if the structure of the instrument is not so easy to find or if you want to consider these extra points. Additionally, you can further refine the instrument, if necessary, through Item Response Theory (when this model applies, among others).