Factor analysis and item analysis

Jules L. Ellis

Contents

Preface

The aim of this book is to give bachelor students in the behavioral sciences an introduction to basic models for latent variables, so that the student can use these models in the construction of scales from items. I discuss successively the following topics: factor analysis, internal consistency reliability (removed: IRT).

It is questionable to use factor analysis for item analysis, but nevertheless this is the most common technique for item analysis in psychology. Moreover, some important psychological theories are based on factor analysis. Therefore, factor analysis must still be discussed.

A step-by-step description is given that focuses on practical application. The format of an 'basic report' and 'concise report' (= ‘short report’) is followed, which was also used in the earlier books of the series. On the other hand, theoretical backgrounds and problems are also discussed, such as the problem of determining the number of factors and the relatively large subjective component of the conclusions in factor analysis.

The book tries to be a step towards the use of structural equation models. For example, the use of the chi-square test and fit sizes such as RMSEA is discussed, as well as the limitations thereof.

I have opted for a level such that the student can actually apply it on real data. On the other hand, the study load had to be limited. That is why the range of topics is not particularly wide.

This book is based on a Dutch book that has been used and improved for several years in the course ‘Psychometrics and Decision Theory’ of the second year of the bachelor's program in Psychology at the Radboud University.

Jules L. Ellis
Nijmegen, December 2017

Part VII

Factor analysis and item analysis

1 Constructing tests, scales and questionnaires

1.1 Use of the terms test and scale

First, let us try to stipulate these terms. A **measuring instrument** is any method that will lead to quantitative data. A **test** is a measuring instrument consisting of multiple components called **items**, from which a single **total score** is derived for each individual. A **questionnaire** can also be a test, if a total score is calculated. A test is not necessarily a *performance test*; it can also be a measuring instrument for attitudes, emotions and social behaviour. To prevent confusion, the word **scale** is often used instead of test. Both test and scale are also used for measuring instruments that consist of multiple, coherent tests, which are then called **subscales** or **subtests**. However, the word scale is also used to designate a test whose items go together well. A scale that consists of subscales should be called a measuring instrument, while its subscales should be called scales. The long and short of it is: everyone uses their own nomenclature. Why would we break with this great tradition?

1.2 Phases in a validation study

When constructing a test, the first step is to study its validity and reliability. Obviously, this should be done before the test is actually used. We will call this kind of study a **validation study**. In a validation study, the following steps are often addressed. The main points (1, 2, …) are in chronological order, while the order of subpoints (a, b, …) within a main point is of lesser importance. One could write an entire book on each of these points.

1 Preparation

a *Choice of the kind of properties that will be measured.* In principle, a separate subscale of multiple items should be made for each psychological property.

b *Exploration of the domain using literature and interviews.* Let's say you want a scale for aggression, for example. You could investigate whether such a scale already exists in literature and whether the scales that exist can be used in your case. The use of pre-existing scales increases the comparability of your study. If your conclusion is that you need to create a new scale after all, you would do well to delve into what properties are often seen as expressions of aggression. This will serve as the starting point to come up with the items. Additionally, it will not hurt to talk with persons from the target group beforehand, to get an idea of what is going on in that area.

2 Formulating the items

During this process, there should be a constant evaluation of the extent to which it may be expected that the items are suitable on substantive grounds. In this process, the following facets will be addressed, among others.

a *Content of the individual items.* Each item should clearly fit within the chosen domain. If you create a scale for aggression, you should not add an item that depends on characteristics other than someone's aggression, such as 'when things are not going well at work, I have a higher tendency to make hurtful remarks toward my partner'. The answers to that question will also depend on the situation of a person in whether they have a job and a partner.

b *Representativity of the collection of items.* Together, the items should form a good representation of the domain. A scale for aggression should not only contain questions about aggression during going out, but also in school, at work and at home. Please note that this could conflict with point (a).

c *Number of items.* Each subscale should have enough items, whilst keeping in mind that 30% of the items might get dropped during analysis. It is hard to say exactly when there are 'enough items', but a scale with less than 10 items is not impressive. Consider that even an exam with 40 multiple-choice items with 4 answer categories is still quite unreliable, since someone with a true score of 5 on a scale of 0 to 10 has 95% chance for a score between 3 and 7 (Ellis, 2004, p. 236). As the test becomes more important – it could lead to important decisions on the individual, for example – higher requirements should be set for the number of items. A practical limitation is that an overly large number of items could cause the test takers to answer the items less seriously.

d *Precise formulation of the items.* The items should not be open to interpretation. The formulation should be adapted to the target group and the aim of the measurement in terms of intelligibility and language. A well-known example is the rule that one should avoid double negatives. Whether that is a good rule is debatable. If you interviewed random school children in the village of Groesbeek about their experience with the local snack bar, it probably would be. In the case of an exam for law students, however, this would be a foolish rule, as juridical texts are full of quintuple negatives, and the exam is meant to gauge whether the students understand such formulations.

e *Content and number of answer categories.* The answer categories should be chosen in such a way that a reasonable variation can be expected. Items to which most people will provide the same answer are not informative. If you asked the student population In Rotterdam 'how often do you go to church', the answer categories 'less than once a week / once a week / twice a week or more' would probably be useless, whereas the same might be distinctive in a strict Christian village. A small number of answer categories will lead to the answers containing little information. The usually recommended number of answer categories is between 5 and 7.

f *Expert opinion.* At this stage, it is advisable to ask a panel of experts for their opinion on the items. As an example: in 2007, I was involved in the construction of a questionnaire to measure the innovative strength of healthcare groups. Several organisations and advisors in the Netherlands have specialised in the stimulation and guidance of innovation. Such people have been working with innovations for years, so you might expect they would know something about it. Therefore, a part of the construction was inviting a group of experts in the field of innovation to discuss the items. The central question in this discussion was whether the items captured the notion of 'innovation'.

3 Planning the first administration

The first administration will yield data based upon which the scale can be adjusted, if necessary. During the planning, thought should be put into the way those data will be analysed and which kind of data will be needed for the analyses. The following points should be taken into consideration.

a *Make use of multiple raters.* If observers or raters are involved, the inter-rater reliability should also be determined. To achieve this, it is necessary to use multiple raters for a single **subject** (= test taker).

b *Other variables that must be measured.* These include background variables and related measuring instruments.

c *Required number of subjects*. The number of subjects required depends in part on the statistical properties of the data that are acquired (MacCallum, Browne & Sugawara, 1996). When applied on scale construction, a minimum of 100 subjects would be necessary. However, Barrett (2007) states that each article with a structural equation model (the kind of model that is also used in scale construction) with less than 200 test subjects should be rejected anyhow. Wirth and Edwards (2007) also suggest that a minimum of 200 test subjects is needed. Some analyses, however, require a minimum of 1000 subjects (Flora & Curran, 2004).

4 First administration of the scale

During this step, the first data are collected that will be analysed in the following steps.

5 Analysis of data of individual items

This is actually akin to a prewash or preselection, during which the worst items are removed.

a *Inter-rater reliability of items*. If the items are based on observations or assessments, then it should be determined that different observers have a high degree of agreement. Because if they cannot agree on the elementary data, you might as well drop the item. The agreement is usually determined using Cohen's kappa or with an intraclass correlation.

b *Variance of the items*. Items with a lower variance are usually less suitable as they contribute little to the differentiation of subjects. If I were to ask an exam question such as 'What is 1 + 1?', everyone would give the correct answer (variance 0) and I might just as well not ask the question. A question that everyone will get wrong is equally uninformative. Items with a low variance will often make a small or even negative contribution to the reliability. As a limit, it is suggested that items that are scored in integers have a variance of at least 1. However, this is not a hard-set limit, and no single limit is fully justifiable. My advice is to remove items if their variance is zero, as well as based on factor analysis, IRT analysis or reliability analysis, but not based on their variance.

c *Skewness and unimodality of the distribution of the items*. Large discrepancies in the skewness of items have an influence on the correlations and thus on the results of the factor analysis. In most forms of factor analysis, the assumption is made that the items follow a normal distribution. As items usually have a small number of discrete answer categories (such as 0-1-2-3-4), these cannot be normally distributed. Nevertheless, it is useful to keep the deviations as small as possible by removing items that strongly deviate from symmetry and unimodality. (Unimodal means that the histogram looks like a bell, not like a bathtub.)

6 Analysis of the relations between the items

The question here is whether the items measure the same property. This is also called *unidimensionality* or *homogeneity*. This is important to justify summarising the items scores into a single total score per subject.

a *Correlations between the items*. Items of the same scale should correlate positively (with the added note that items may be mirrored based on their content). This is because they essentially should measure the same trait.

b *Factor analysis of the items*. This analysis is a more detailed study of whether the correlations between the items justify believing that the items measure the same trait. If the items turn out not to be unidimensional, the scale might have to be split into subscales, or items might have to be removed.

c *Analysis of the internal consistency reliability*. In this analysis, it is investigated whether the scale contains enough items to allow the sum score to be seen as a reliable measurement. How many items are required also depends on the strength of the correlation between the items. In addition, items with a negative contribution to the reliability are removed from the scale.

7 Establish norms

In this step, you investigate what the averages, standard deviations and percentiles are for the groups to be differentiated. This is done to describe which scores are ‘normal’ in the studied population.

8 Analysis of the relations of the test scores to other variables
This is often only done in later administrations of the test. It is a continuous process of elaboration of the information about the test.
a *Test-retest reliability*. With this, the stability of the scores over time is investigated. The scores do not always need to be stable; that depends on the construct and the theories on its subject. For example: a test that aims to measure the state of mind on a specific day will not necessarily yield the same result a year later. But it is generally useful to know the stability.
b *Criterion validation*. With this, the extent to which the test can predict other variables is investigated. The aim hereof is to substantiate the practical use of the test.
c *Construct validation*. With this, it is investigated whether the test has the theoretically expected relations. This is mainly about the theoretical interpretability of the test.

In the following chapters, we will only consider the aspects in point 6. A central notion therein is the 'unidimensionality', which we will now introduce.

1.3 Unidimensionality

In simple words, **unidimensionality** means that items measure the same thing. This can be seen as a part of construct validity. Construct validity means that the items measure the intended construct, and this implies that they all measure the same. The kind of data needed to assess unidimensionality is more akin internal consistency reliability, however. Unidimensionality also has an influence on that reliability.

As was noted in the previous section, unidimensionality is needed to justify that item scores are summarised into a single total score per subject. This requires some elaboration. It might seem fully obvious to add up item scores. But how do you know *which* scores you can add up? Can you add depression items to intelligence items? No, of course you cannot. Not any more than you can add up apples and pears. But then, how do you know that none of the existing tests are adding up apples and pears? As a test constructor, you might think that items fit well together, but do the data reflect that? Would you still think two items measure the same thing if they turn out to have a negative correlation?

We use tests to quantify human behaviour, and it is not at all obvious that this is meaningful. Some people are deeply offended when you try to compress their beautiful feelings into nasty numbers. And they might be right, because whence comes the idea that it is reasonable to do so? It is true that we have been summing up scores in various tests for the past century, but that does not mean it is sensible.

The crucial question is to what extent the sum score presents a good summary of the behaviour that the subject showed during the test. In other words: to what extent the item scores are summarised adequately. For example, assume a questionnaire consists

of 50 yes/no questions, which we code with a 0 (no) and 1 (yes). Someone with a sum score of 25 therefore answered 'yes' 25 times. But without additional information, we still do not know to *which* questions the subject answered yes. These could be the first 25 questions just as well as the 25 last questions. Or only the odd-numbered questions. It is therefore conceivable that, in a group of subjects with the same total score, two subgroups exist that show an exactly opposite behaviour: the questions that one group answers with a 'yes', are answered with a 'no' in the other group. In such a case, the sum score is not a good summary of the item scores. If the concerned questionnaire would be about aggression, it might mean that there is not just one type of aggression, but at least two different types, for example direct and indirect aggression. The consequence of this is that one should not use a single test score per subject, but at least two different test scores. Research into unidimensionality therefore forces psychologists to differentiate their concepts (in this example: aggression into direct and indirect aggression) if the data gives reason to do so. This also prevents the all too easy addition of items that are actually different.

In the above paragraph, you saw an example in which a sum score was not a good summary of the item scores. So, when *is* the sum score a good summary of the item scores? That is the case when subjects with the same sum score generally have about the same item scores, apart from a random noise. The models used to investigate unidimensionality describe this in great detail. Most of these models make the following assumptions:

1. *Unidimensionality*. Each person can be characterised by a single value that indicates to what extent that person has the trait that we want to measure. This value is generally unknown and is therefore called the latent trait. For an intelligence test, this would be someone's unknown true intelligence, and for a depression scale, it is someone's unknown true depression. That value is generally indicated by θ (the Greek letter theta), but in some cases, it is also indicated by τ (the Greek letter tau, for ‘true score’).
2. *Monotonicity*. The expected score for an item increases with θ. If the score for item i is indicated by X_i and the expected value by E (for expectation), the following then holds:

   $$E(X_i \mid \theta) = f_i(\theta)$$

   where f_i is an increasing function. Here, $E(X_i \mid \theta)$ denotes the average of X_i in the subpopulation of subjects with score θ on the latent trait.
3. *Local independence*. Within a group of subjects with the same value of θ, the items are not correlated. In the total population, high correlations between items are allowed.

For example: someone who has a low aggression should have a low score *on all items used* (except for noise), and if that subject's aggression increases, this should show in all items. Another example is a written exam. It is desirable that good students have a better chance on all questions. An exam question on which good students score poorly – that is, a question on which one scores worse the better one knows the subject matter – is undesirable.

Unidimensionality and monotonicity cannot be distinguished empirically. If you would only assume unidimensionality, without monotonicity or something alike, it would not be testable (Sijtsma & Junker, 2006, p. 86). The same applies to local independence. For this reason, the term unidimensionality is often used for the three assumptions jointly. The word therefore has two meanings.

In the text above, it was suggested that unidimensionality, monotonicity and local independence can be tested together. How can that be done? A simple prediction following from the aforementioned assumptions is: all items of the scale should have non-negative correlations (Mokken, 1971; Mokken & Lewis, 1982; Holland & Rosenbaum, 1986; Ellis & Junker, 1997; Junker & Ellis, 1997; Junker & Sijtsma, 2001b).

What is the difference between unidimensionality and internal consistency reliability? Unidimensionality implies that the items measure the same trait, save for noise, but it does not say anything about the size of the noise component. Internal consistency reliability says something about the size of the noise in the total score, but does not give an answer to the question whether the items measure the same trait. For example: the arithmetic items '3 + 4 = ?' and '5 + 2 = ?' are unidimensional, but their total score is not reliable because there are only two items. Conversely, the total score of an IQ test usually has a high reliability, but intelligence is not unidimensional, because there are various types of intelligence (such as fluid intelligence and crystallized intelligence).

2 Conducting and reporting factor analysis

2.1 Background

Factor analysis is a statistical method mainly developed by psychologists to study the empirical patterns of psychological test scores. The method was first developed by Spearman (1904a, 1927) in his research into intelligence. Later, the method was extended by Thurstone (1931, 1938), also in intelligence research. In personality theory, factor analysis has become known through the 16-factor theory of Cattell (1950; Cattell, Eber & Tatsuoka, 1970) and the now popular Big Five theory (McCrae & Costa, 1987). Nowadays, factor analysis is also used in other scientific areas, such as economics.

In psychology there are a number of fairly different goals for which factor analysis is used:

1. The first possibility is to describe theoretically a specific domain, for example performance on cognitive tasks. The outcomes of factor analysis are then part of a theory. In this context, a so-called confirmatory factor analysis will usually be used, which means that the theory will be tested.
2. The second option is to use factor analysis to gain insight into a domain, without any theory in advance. In that case, one speaks of exploratory factor analysis. There is no theory in advance, but the goal is to develop a theory.
3. The third option is to use factor analysis to limit the number of variables. In that case there is no theory in advance and there is no goal to develop a theory. In this case factor analysis is a form of data reduction. This can be important, for example, if the number of variables is so large that the overview is lost, or if other analyses lose a lot of power. For example, MANOVA uses a kind of factor analysis to limit the number of dependent variables.

In this book we will mainly focus on the use of factor analysis in test validation. Factor analysis is then used to examine the construct validity. This can be both confirmatory and explorative. In addition, it can be done on two levels of data:

a. Level one is to examine the relation of the **total score** of the test with the total scores of other tests. For example: the relationships between the subtests of an intelligence test such as the WAIS can be investigated in order to check whether they correspond with the Cattell-Horn-Carroll theory of intelligence.
b. Level two is to examine the relationships between the **item scores** within the test. This is done in the construction phase of the test. The question here is

whether the items are unidimensional. In other words, the question is whether the items fit together, and whether the data can be adequately summarised by using only one test score per person. These analyses logically precede analyses of the total scores.

In this chapter we will mainly focus on this last level, the use of factor analysis in item analysis in test construction. However, it is far from undisputed to use factor analysis for item analysis. Many psychometricians (which I define for convenience as people who publish in the journal Psychometrika) believe that Item Response Theory (IRT) should be used for this purpose and that factor analysis is an inferior and outdated method for this purpose. The reasons to still teach factor analysis in test construction are

- In most publications in psychology in which a new measuring instrument is developed, factor analysis is still used (Ten Holt, Van Duijn & Boomsma, 2010).
- Factor analysis is available in the widespread package SPSS, while IRT is not. IRT is available in the package R, but that might require more effort to learn.

And of course this is a vicious circle. Psychologists do not want to use IRT because nobody does it because it is not in SPSS because psychologists do not want to use it.

In any case, in the following chapters you will learn what is currently customary in test validation, and then you are actually about half a century behind (Borsboom, 2006, p. 425). Factor analysis is in itself an excellent technique and not outdated, but as a method for item analysis it is dubious. We will, however, also deal with other applications in which factor analysis is adequate. In later chapters we will also deal with IRT.

Incidentally, factor analysis and IRT are conceptually largely the same. In both cases, the question is whether the items fit together and measure the same. The main difference is that factor analysis is based on normally distributed variables and linear relationships, while in IRT the starting point is categorical variables with non-linear relationships.

2.2 Learning objectives of this chapter

After studying this chapter, you can apply a factor analysis on items in a validation study in order to

- assess unidimensionality, and if necessary
- divide the items into subscales, and
- identify unusable items.

In particular you can

- indicate whether an exploratory or confirmatory factor analysis is desirable;
- explain the purpose of that analysis;
- perform the factor analysis with SPSS as far as possible;

- thereby motivating the extraction method and the rotation method;
- draw conclusions about the number of factors on the basis of the output and motivate that conclusion;
- give a substantive interpretation of the factor pattern based on the output;
- on this basis draw a conclusion about the unidimensionality of a scale;
- use the factor pattern to divide the items into subscales and identify which items are useless.

In addition, you can make a more or less extensive report, which will be referred to as a 'basic report' and a 'concise report' respectively. Finally, it is desirable that you have some insight into the relation between factor loading and correlations, and you can

- on the basis of a loading plot with two orthogonal factors, estimate what the reproduced correlations of the manifest variables are ('visualise').

2.3 Definition of an basic report of a factor analysis

As in the statistics books on which this book elaborates (Ellis, 2003a, 2003b), we say that a basic report is an extensive point-by-point report of the analysis. In the case of a factor analysis, this should contain the same parts as in the previous books:

Design
Degree of control
Aggregate data
Hypotheses
Analysis method
Estimates
Test statistics
Decision
Interpretation

The most important parts of this are reflected in a concise report. This will be discussed at the end of this chapter. The above parts will be discussed below. However, we first start by discussing an example that will be used throughout the chapter.

2.4 Running example

As an example, in this chapter we take the research of Diesfeldt (1997). It examines a Dutch questionnaire that is intended to measure depression in the elderly in psychogeriatric centers. Of course there are more questionnaires to measure depression, such as the BDI (Beck's Depression Inventory). Why this list? Diesfeldt motivates this as follows:

> When using self-assessment scales, researchers should keep in mind that the limitations of language comprehension, insight and memory impede reliable

> interpretation of the questionnaire.[14,15] However, none of the questionnaires is designed specifically for use in elderly people with dementia. On further inspection, the questionnaires appear to contain complex sentences that place high demands on concentration and understanding. Some questions refer directly to cognitive symptoms that already occur in dementia, but do not in themselves have to lead to mood changes.
>
> (...)
>
> An already existing but little known method meets some of the above-mentioned drawbacks. It is the Depression List of the psychiatrist L.A. Cahn, which has been specially formulated for use in people with dementia. The terms used are simple and understandable. References to cognitive symptoms are avoided. (Diesfeldt, 1997, pp. 113-114; translated)

Diesfeldt therefore focuses in the article on a psychometric examination of the Depression List. In the Method section it is described as follows by Diesfeldt:

> The Depression List consists of 15 keywords that are shown one by one on separate charts (see Appendix 1).[20] The keywords are derived in part from the DSM-III criteria for the clinical diagnosis 'depression'.[21] They have been chosen to to gauge feelings of respondents about themselves, their environment and their future. The researcher supports the keyword with a simple question (for example: 'Do you feel satisfied?'), waits for the answer and then classifies it on a scoring form under one of the three pre-printed answer alternatives. The scores per item vary from 0 to 2, the sum score over the 15 items varies between 0 and 30. (Diesfeldt, 1997, p.114; translated)

Diesfeldt then discusses a study that suggests that the items have a high value of kappa. This means that in cases where two observers were present, they usually agreed on the scoring. In observation scales it is customary to examine this first, because an item with a low inter-rater reliability is in any case not suitable for a measuring instrument and therefore does not need to be investigated further. However, in the current chapter this analysis is not a learning objective and that is why we will not discuss it anymore.

The Diesfeldt research sample consisted of 197 people (that is rather small for a factor analysis), which were examined in the day center of a psycho-geriatric nursing home. The Depression List was thus administered to each of those persons. In addition, other measuring instruments were used, but these are not important in this chapter.

A part of Diesfeldt's psychometric research consisted of a factor analysis on the items. That will be discussed in this chapter. The question is: Are these items suitable for measuring depression?

Data

Because the items must be analysed in psychometric research, it is important that the data are collected at item level. It is therefore not enough to store only the sum score of

each person. The scores of each person on each item must be kept. The data matrix therefore looks schematically as Table 2.1.

Reading question. Why is Diesfeldt investigating the Depression List? What is the question of this research?

Table 2.1

Person	*v1*	*v2*	*v3*	*v4*	*v5*	*v6*	*v7*	*v8*	*v9*	*v10*	*v11*	*v12*	*v13*	*v14*	*v15*
A	1	1	1	1	1	2	2	0	2	1	2	2	1	1	0
B	2	1	0	2	2	2	1	0	1	1	1	1	1	1	2
C	0	0	1	0	0	2	2	1	1	1	1	0	0	0	1
D	1	2	2	1	0	0	1	0	2	2	2	0	2	1	2
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…

These variables include the keywords that are shown in table 2.2. The symbol (-) behind a keyword means that this item is mirrored. A high score is then awarded to a negative answer.

Tabel 2.2

Variable	*Original keyword*	*Translation*
v1	Tevreden (-)	Satisfied (-)
v2	Slapen (-)	Sleeping (-)
v3	Eten (-)	Eating (-)
v4	Gezond (-)	Healthy (-)
v5	Moe	Tired
v6	Oud	Old
v7	Eenzaam	Lonely
v8	Vrienden (-)	Friends (-)
v9	Bezoek (-)	Visitors (-)
v10	Somber	Gloomy
v11	Verveling	Boredom
v12	Opgewekt (-)	Excited (-)
v13	Hulpeloos	Helpless
v14	Zwak	Weak
v15	Toekomst (-)	Future (-)

2.5 Design

In the design section you specify the **manifest variables** of the research question. A manifest variable is an observable variable. When analysing test items, each column with scores of a test item is a manifest variable.

Explanation
We use factor analysis to explain the correlations between a large number of manifest variables using a small number of common factors. The manifest variables can be viewed here as dependent variables and the factors as independent variables. Characteristic of a factor analysis is, however, that the independent variables are **latent.** That is, they are not observed and often not even *observable.* Therefore, we will specify the factors in the hypotheses.

Example
The manifest variables are the items of the questionnaire. The design is therefore:

Manifest variables: the 15 items from the Depression List.

In a factor analysis there are usually a lot of manifest variables, so it is not very inspiring to write the names of them completely. Nevertheless, the description must be unambiguous.

Note that we have silently changed the meaning of the word ‘item’. Originally we meant by an item a part of the test (in the running example, a keyword). From now on, the corresponding columns of the data matrix will also be called items. For example, column *v1* of Table 2.1 may also be referred to as the item Satisfied.

2.6 Degree of control

This is almost always passively observed.

Explanation
A factor analysis is generally done with manifest variables that have been passively observed. The 'independent' variables are latent in a factor analysis, and it is difficult to see how you could manipulate them if you do not even know how to observe them.

2.7 Aggregated data

The aggregated data consist of the **correlation matrix** of manifest variables. If the items are ordered response categories, it is best to use **polychoric** correlations. The report must state what type of correlation has been used.

Explanation
1 Why are these the aggregated data?
Factor analysis aims to explain the correlations between the manifest variables. These therefore form the input on which the rest of the analysis is based. Sometimes a factor analysis is done on the covariance matrix, and then the covariance matrix forms the aggregated data. By default, however, the correlation matrix is used. Otherwise, the results will depend on the measurement units of the manifest variables.

2 Report the correlation matrix!

The correlation matrix is, unfortunately, often omitted in articles because it requires so much space. That is a bad practice. In factor analysis, subjective assessments play a relatively important role, and other researchers should be enabled to interpret the data differently (Henson & Roberts, 2006). This is possible only if the correlation matrix is reported. Especially now that many journals offer the possibility to add material to articles online, there is no excuse to omit the correlation matrix.

3 Why tetrachoric or polychoric correlations?

If the manifest variables are items of a questionnaire, they are usually not continuous, but discrete, ordered categories, such as ratings (e.g., 1-2-3-4). Such items are called **polytomous** (or sometimes polychotomous but see Weiss, 1995). Items with two response categories (for example, right / wrong) will be referred to as **dichotomous.** Pearson, who along with Galton invented the correlation coefficient, realized that the simple correlation coefficient sometimes gives a distorted picture of the relationship between two dichotomous variables, and developed an alternative that came to be known under the name **tetrachoric correlation** (Pearson, 1900). The generalization of this to polytomous items is called the **polychoric correlation** (Olsson, 1979 Digby, 1983; Drasgow, 1988; Hutchinson, 1993). In order to distinguish the correlations, the ordinary correlations are referred to as the ‘Pearson product-moment correlation'. Note that the addition of the name Pearson contributes little to the clarification of the type of correlation, because the Pearson product-moment correlation was probably first conceived by Galton, while the tetrachoric correlation was first conceived by Pearson.

In the 1940s researchers started to realize that there was a problem with conducting factor analysis on the product moment correlation of dichotomous items. In his Presidential Address to the Psychometric Society Carroll (1961) noted that it could be better to utilize tetrachoric correlations. The reason why items with ordered categories are better summarised with polychoric or tetrachoric correlations is that with an ordinary factor analysis on product moment correlations it is assumed that the variables are normally distributed and linearly related. These assumptions cannot be fulfilled with ordered categories, and a common factor analysis on product moment correlations can lead to artificial factors (McDonald & Ahlawat, 1974; Bernstein & Teng, 1989; Waller, Tellegen, McDonald & Lykken, 1996).

For a long time the prevailing opinion was therefore that with polytomous items it is better not to use the usual correlations, but rather the **polychoric correlations** (Carroll, 1961; Christoffersson, 1975 Muthén, 1978, 1984; Knol & Berger, 1991; Wirth & Edwards, 2007), although this does not always lead to different results (Parry & McArdle, 1991) and a disadvantage is that a larger sample is needed (Finch & West, 1997). Nevertheless, this recommendation was widely ignored for decades (e.g., Forrest, Lewis & Shevlin, 2000), probably because polychoric correlations are not implemented in SPSS. However, there are several other programs

in which they can be calculated (including LISREL, Mplus). Another solution to this problem is Item Response Theory (these chapters were not translated).

Example

Diesfeldt fortunately reports the correlation matrix. This allows us to replicate his analysis and perform alternative analyses. If in the following we sometimes come to a different conclusion than Diesfeldt, then one should remember that this is only possible because Diesfeldt neatly reported the correlation matrix.

The correlations reported by Diesfeldt are shown in Table 2.3.

Table 2.3

	v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11	v12	v13	v14	v15
v1	1.0	.31	.24	.40	.22	.23	.37	.17	.20	.38	.27	.50	.24	.26	.42
v2	.31	1.0	.18	.22	.31	.27	.13	.06	.08	.09	.25	.17	.03	.13	.17
v3	.24	.18	1.0	.04	.03	.17	.32	.25	.19	.18	.09	.17	.07	-.12	.17
v4	.40	.22	.04	1.0	.30	.34	.26	.11	.24	.34	.19	.30	.29	.52	.32
v5	.22	.31	.03	.30	1.0	.26	.16	.03	.06	.24	.11	.22	.20	.36	.24
v6	.23	.27	.17	.34	.26	1.0	.22	.07	.13	.34	.18	.39	.19	.29	.37
v7	.37	.13	.32	.26	.16	.22	1.0	.21	.52	.42	.26	.38	.23	.14	.30
v8	.17	.06	.25	.11	.03	.07	.21	1.0	.33	.29	.20	.22	.10	.06	.11
v9	.20	.08	.19	.24	.06	.13	.52	.33	1.0	.41	.29	.32	.14	.23	.19
v10	.38	.09	.18	.34	.24	.34	.42	.29	.41	1.0	.28	.52	.38	.37	.41
v11	.27	.25	.09	.19	.11	.18	.26	.20	.29	.28	1.0	.38	.19	.06	.35
v12	.50	.17	.17	.30	.22	.39	.38	.22	.32	.52	.38	1.0	.32	.31	.52
v13	.24	.03	.07	.29	.20	.19	.23	.10	.14	.38	.19	.32	1.0	.22	.22
v14	.26	.13	-.12	.52	.36	.29	.14	.06	.23	.37	.06	.31	.22	1.0	.20
v15	.42	.17	.17	.32	.24	.37	.30	.11	.19	.41	.35	.52	.22	.20	1.0

Diesfeldt reports them in a different order, which is based on the results of the factor analysis. I did not copy this, just to show you that without factor analysis, it's hard recognise any pattern in it. Do you agree? Let us now look ahead to these correlations if the variables are sorted on the basis of the factor analysis. Furthermore, the display

may be somewhat simplified by using the following principles, which have nothing to do with factor analysis:

- It is sufficient to display only the triangle at the lower left or top right of the correlation matrix because the correlation matrix is symmetric.
- The diagonal can be left blank because everyone knows that these correlations are exactly equal to 1.
- The 0. in each number may be omitted because everyone knows that correlations must be between -1 and +1.

The correlation matrix reported by Diesfeldt, is shown in Table 2.4. The **order of the variables** here is based on the factor analysis, and so is the use of bold font. The shading is added by me, also based on the factor analysis.

Table 2.4

	v1	v6	v10	v11	v12	v13	v15	v4	v5	v14	v3	v7	v8	v9	v2
v1		**23**	**38**	**27**	**50**	**24**	**42**	40	22	26	24	37	17	20	31
v6			**34**	**18**	**39**	**19**	**37**	34	26	29	17	22	7	13	27
v10				**28**	**52**	**38**	**41**	34	24	37	18	42	29	41	9
v11					**38**	**19**	**35**	19	11	6	9	26	20	29	25
v12						**32**	**52**	30	22	31	17	38	22	32	17
v13							**22**	29	20	22	7	23	10	14	3
v15								32	24	20	17	30	11	19	17
v4									**30**	**52**	4	26	11	24	22
v5										**36**	3	16	3	6	31
v14											-12	14	6	23	13
v3												**32**	**25**	**19**	18
v7													**21**	**52**	13
v8														**33**	6
v9															8
v2															

The pattern that you should recognise here, is that there are three groups of variables. The variables of the same group are always next to each other. The correlations within

a group are in bold text, and those cells are also shaded. A different shade has been used for each group. *Within a group of variables, the correlations are generally higher than between different groups.*

Consequently, factor analysis can be considered as **a glorified way of grouping variables based on their correlations.** But what does that mean? Well, the idea is that the variables within such a group apparently have more in common than variables of different groups. In factor analysis it is assumed that this happens because such variables partially have a common cause. This cause is known as a **common factor.** In this example, Diesfeldt called the factor of the first group *Spirited*, the second factor *Health*, and the third factor *Social contacts.* Those are just names. If you look at the keywords of the items, you will see that they are pretty good names, which describe what kind of items fall into the relevant group.

The common factors are not measured themselves; only their *manifestations* are measured. These manifestations are the items. The factors explain why some items correlate more with each other than other items.

Reading Question. How can you see that items probably belong to the same factor?

2.8 Hypotheses

In the hypotheses section, first indicate whether the factor analysis is **exploratory** or **confirmatory.** In an exploratory factor analysis **(EFA)** you have no hypothesis about the amount and nature of the factors. In that case, you use factor analysis to gain insight into the data, which may then lead to a theory. In a confirmatory factor analysis **(CFA)** you have a hypothesis about the amount and nature of the factors. That hypothesis is based on a substantive theory. A third possibility, which is rarely important in psychology is that factor analysis is used for **data reduction.** In that case there is no hypothesis and no hypothesis is formed. The only goal is to reduce the number of variables with as little loss of information as possible.

In a confirmatory factor analysis, proceed by specifying **how many common factors** are assumed. If more than one common factor is assumed, you should also indicate whether a **simple structure** is expected. A simple structure means that each manifest variable belongs to exactly one common factor. If possible, specify **which manifest variables belong to which factors.** Furthermore, state whether you expect that the **common factors** are **uncorrelated.**

Although you do not have to write this in the hypotheses, it can be helpful in a confirmatory factor analysis to state whether the alternative hypothesis is **fixed-domain** or **fixed-model.** In a fixed-domain analysis, rejection of the null hypothesis will lead to a change of the model, while the domain (the collection of variables that is investigated) remains the same. The model will then be changed by, for example, assuming more factors or by assuming that some variables (also) belong to a different factor. In a fixed-model analysis, rejection of the null hypothesis leads to a change of

the domain, while the model remains the same. The variables that are responsible for the violation of the model are then removed from the analysis.

In the investigation of unidimensionality of items that one wants to include in a scale (test), one has to use a confirmatory factor analysis with one factor per subscale and a fixed model. If this hypothesis is rejected, and if one proceeds by seeking a better division into subscales, these subsequent analyses should be considered explorative.

Names of factors will be displayed from now on in italics. In both latent and manifest variables, the name will begin with a capital letter to distinguish them from the same non-technical concept. This is a deliberate departure from the APA guidelines.

Explanation

1 Explorative versus confirmatory

In the literature, often a distinction is made between exploratory and confirmatory factor analysis. In exploratory factor analysis, you actually have no theory. You just have a lot of manifest variables of which you don't understand very much, and you hope that the factor analysis will create structure in this chaos. You do not know how many factors there will be, and you do not know what kind of factors can be expected and which manifest variables will group together. With confirmatory factor analysis, on the other hand, you have a hypothesis about all these points. For example, you would have the hypothesis that the items contain the factors *Emotion, Social* and *Physical*, and moreover you know in advance which item belongs to which factor.

2 Deviant meaning of confirmatory

In the literature, the term 'confirmatory factor analysis usually used in a narrower sense than in this book. In this stricter sense it concerns factor analyses in which a single statistical test is obtained for a hypothesis that describes how many factors there are and which manifest variables belong to which factors. Such analyses can normally not be done with SPSS, but only with specialized *structural equation modeling* (SEM) programs like LISREL. The exception to this is that, if the hypothesis contends that there is only a single factor, then SPSS can be used to do a confirmatory factor analysis in the strict sense of the term.

In this book we have chosen to use the term 'confirmatory' in a broader sense, indicating that there is a hypothesis that specifies which variables belong to which factor, even if a program is ultimately used that only partially tests this hypothesis. Please be aware that this usage of the word deviates from the usage elsewhere in the factor analytic literature – but it agrees with the usage in other methodological literature. This will be defended in Section 2.18.3.

3 Correlated versus uncorrelated factors

It used to be common practice to assume that factors are uncorrelated, mainly because of the mathematical simplicity that this assumption entails. Nowadays there is almost consensus that in both explorative and confirmatory factor analysis it is better not to assume in advance that the factors are uncorrelated. If the factors are correlated, it will automatically emerge during the analysis. Only if the goal is just data reduction, it may make more sense to define factors that are uncorrelated.

4 Fixed-domain versus fixed-model

The terms fixed-domain and fixed-model are not common in the literature. I use them to describe two different goals of factor analysis, not explicitly recognised in the literature, but nonetheless important. It is not a mandatory part of an basic report; it is only recommended.

In general you have to assume a fixed domain. You have a number of variables of which you want to explain the correlations, and if that does not work then you cannot simply throw away the disruptive part of the data and then pretend that you have succeeded. For example, Spearman (1904a, 1927), the psychologist who developed the first factor analysis, found that all cognitive tasks are highly correlated with each other. He explained this by assuming that all these cognitive tasks rely on the same common factor, which he called 'general intelligence'. If, after a while, it turns out that mathematics and philosophy have a negative correlation, then he cannot get rid of this by ignoring mathematics and philosophy. His theory was about *all* cognitive tasks, and it is clear that mathematics and philosophy belong to this domain too.

The important exception to this is a psychometric analysis of test items. In a test, the item scores of a person are eventually summarised in one test score, and that is only useful if the items are governed by one factor. If an item does not meet this, it should be removed from the test, and then it does not need to conform to the model anymore. In other words, the model is not *descriptive* but *prescriptive* here. In practice, one is often not sure which items belong to a test, and factor analysis is then used to provide a definitive answer. For example, suppose that the test is meant to measure depression. Nobody knows with certainty what depression is and what the symptoms are. When a given item does not fit well with the other items according to the factor analysis, then that is a reason to believe that this item might not be such a good indication of depression. There can be many reasons for this. One reason may be that the persons have misunderstood the question.

Examples

1 According to the above rules one should use a confirmatory factor analysis with one factor for the Depression List, with a fixed model. You could represent this hypothesis with Figure 2.1.

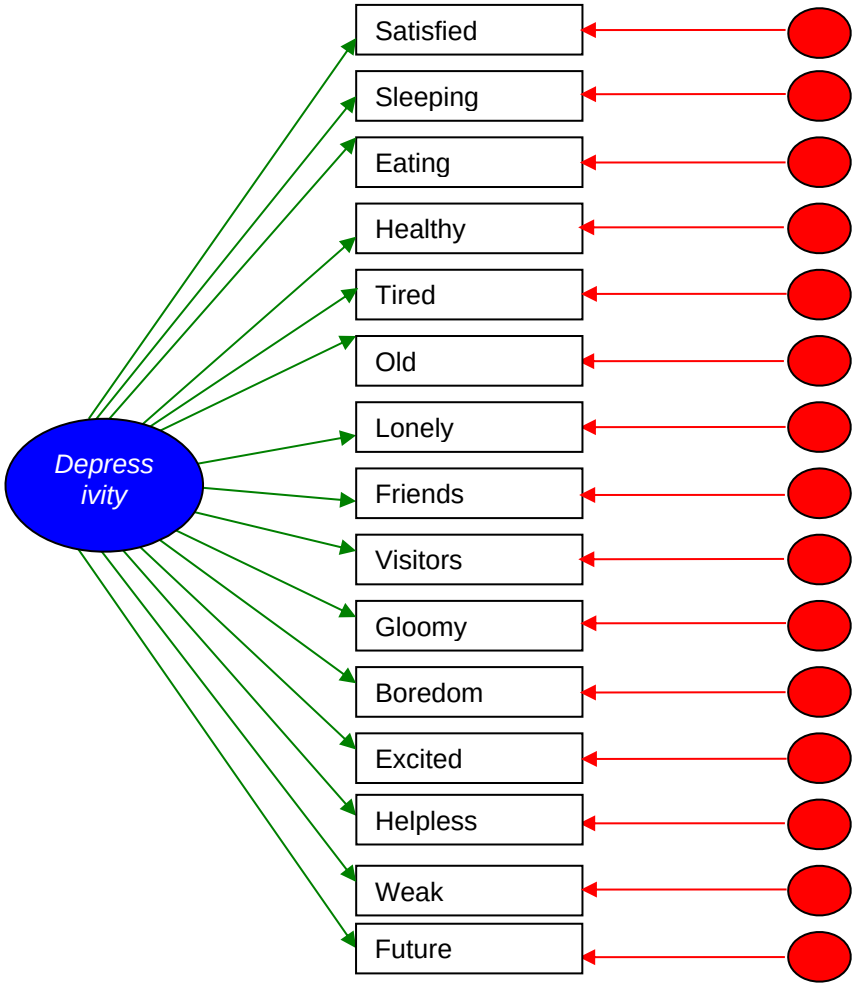

Figure 2.1

The left side of the figure indicates that all of the items (manifest variables: rectangles) depend on a common factor (latent variable: oval), which is referred to as *Depressivity*. The right side indicates that any overt variable additionally depends on a unique factor (latent variable), which is specific to that item. These may be noise and include measurement errors and are the reason that the item correlations are less than 1. These unique factors have not been defined in more detail.

2 In reality Diesfeldt did an exploratory factor analysis. The results suggested that there were several factors. This does not exclude the possibility that a model with one factor may also fit with the data. As a result, after this exploratory analysis it is not

clear to what extent it is possible to summarise the Depression List with a single score per person. Only the above discussed confirmatory analysis provides an answer to this.

The fact that Diesfeldt did something else than prescribed above does not mean that his research is wrong. I just think it would have been better to first show with a confirmatory factor analysis that a one-factor model does not match the data, and then proceed with an exploratory factor analysis to determine how many factors are needed. But this is nowadays much easier than it was in 1997. Later, Diesfeldt (2004) carried out a follow-up study in which confirmatory analyses were conducted on the basis of the results of the 1997 exploratory analysis.

3 The BPS (Van Loveren-Huyben, van der Bom & Bronts, 1988) is another tool for the elderly, based on observation of behavior by caregivers. There are 33 items, which are divided into three scales: *Cognition, Mood* and *Contacts.* In the scoring, one does not take the sum over all 33 items, but each person gets three sumscores, for *Cognition, Mood,* and *Contacts,* respectively. These scales are based on an exploratory factor analysis at the time of construction. In later years, it has been examined several times whether this structure of three scales is still adequate. The hypotheses are then as follows:

Confirmatory factor analysis
Three factors: Cognition, Mood and Contacts
Simple structure expected
The items belong to the following factors (see Table 2.5).

Table 2.5

Item	*Cognition*	*Mood*	*Contacts*
Cognition item 1	X		
Cognition item 2	X		
Cognition item 3	X		
...	...		
Mood item 1		X	
Mood item 2		X	
Mood item 3		X	
...		...	
Contacts item 1			X
Contacts item 2			X
Contacts item 3			X
...			...

The factors can be correlated.
The alternative hypothesis is fixed-model.

Reading Question. What is the hypothesis when factor analysis is used to analyse items of a test? What should you do if the hypothesis is not correct?

2.9 Analysis method

In the analysis method section, specify the **extraction method** and the **rotation method.** The extraction method determines how the factor pattern is sought initially. The most relevant options are **principale components analysis** (PCA) (Kelley, 1928; Hotelling, 1933, 1936) and **maximum likelihood** (ML) (Lawley, 1940; Jöreskog, 1967, 1969; Jöreskog & Lawley, 1968 Jöreskog & Sörbom, 1996, 1999, 2006). In psychology ML is almost always the better choice, or sometimes *principal axis factor* (PAF) (not to be confused with PCA) if there are deviations from the normal distribution (Finch & West, 1997; Fabrigar, Wegener & MacCallum Strahan, 1999. Reise, Waller & Comrey., 2000; Stewart et al, 2001; Costello & Osborne, 2005).

As noted in the section on aggregated data, factor analysis may be based on polychoric correlations. In that case, an *asymptotic distribution free* (ADF) method is recommended (Browne, 1974, 1984; Muthén, 1978, 1984; Jöreskog, 1990, 1994; Finch & West, 1997; Flora & Curran, 2004; Wirth & Edwards, 2007). These are not yet available in SPSS. Within SPSS, *unweighted least squares* (ULS) can be used in this case (ULS). This method will not yield a *p*-value, but it is to be expected that it leads to a good solution. The value of this method is underestimated (Jöreskog, 2003).

The rotation method determines how the initially extracted factor pattern is converted into a factor pattern with a simple structure (Thurstone, 1947, 1954). The most relevant choices are Varimax (Kaiser, 1958) and Promax (Hendrickson & White, 1964). **Varimax** is used if factors are assumed to be uncorrelated. This is called an **orthogonal** rotation. **Promax** can be used if the factors may be correlated. This is called an **oblique** rotation.

Explanation

1 The difference between extraction and rotation

The main output of factor analysis is the factor pattern. This is a table that describes how the manifest variables depend on factors (the exact meaning of a factor pattern will be explained in the next section, hence you may want to read the present section again after reading the next section). In an exploratory factor analysis, the pattern factor is searched in two phases (Mulaik, 1972). The first stage is the extraction. Herein, one searches a factor pattern that explains the correlation matrix as good as possible, assuming uncorrelated common factors. The factors come in the order of their importance: the most important factor, which explains the majority of the correlations, is found first. The second phase is the rotation. Here, the found factor pattern is converted into a factor pattern that explains the correlations equally well, but that

complies as much as possible to simple structure (Thurstone, 1954). In addition, these factors may be correlated.

In a confirmatory factor analysis, this division in phases is not used. A factor pattern is estimated that explains the correlations as good as possible and that moreover meets the specifications of the hypothesis (Jöreskog, 1969). In SPSS, it is not possible to do a confirmatory factor analysis in that way, unless the hypothesis is that there is one factor.

2 Extraction: PCA vs. ML

As mentioned earlier, especially PCA and ML are important in psychology. The reason for not recommending PCA is that PCA can be viewed as a special case of factor analysis in which the manifest variables do not contain unique factors. This implies that the variables would all have perfect reliability (Mulaik, 1965). This is an unrealistic assumption, which leads to a distortion of the outcomes if it has been violated. See Fabrigar, Wegener, MacCallum and Strahan (1999) for a more detailed discussion.

The other extraction methods are mostly of historical interest, or theoretically, and usually produce results that are very similar to the ML method. Only if there are deviations from a normal distribution they might be better.

Although I generally discourage PCA, it is for several reasons still important to discuss PCA in this book. The first reason is banal and teaches hopefully something about psychologists: an important reason why PCA is used so often, is that it is the default in SPSS (Reise, Waller & Comrey, 2000; Costello & Osborne, 2005) (a default is an option that the program chooses if you don't specify which option you want). Since many users of factor analysis have no idea what they are doing, they let it go, in the naive hope that the creators of SPSS picked the best option as the default. The makers of SPSS picked not the best but the simplest option as the default. PCA is a commonly used method, but that says nothing about its suitability. Borsboom (2006) writes about this:

> The reason that, say, Cronbach's alpha and principal components analysis are so popular in psychology is not that these techniques are appropriate to answer psychological research questions, or that they represent an optimal way to conduct analyses of measurement instruments. The reason for their popularity is that they are default options in certain mouse-click sequences of certain popular statistics programs. (Borsboom, 2006, p. 433)

The second reason is that PCA is by far the simplest form of the factor analysis, and the only one in which the factor scores can be calculated exactly from the manifest variables. There is therefore no need to rely on latent variables. Some authors therefore believe that PCA should not be called factor analysis. That no latent variables are involved, is seen as an advantage by some sort of rigorous empiricists who I do not understand. You cannot see the back of the moon either; do they not believe in it?

Nevertheless, this reason for choosing PCA is more rational than the fact that it is the default.

The third reason is that PCA is the best choice if one requires only a reduction of the number of variables without theoretical context (Fabrigar, Wegener & MacCallum Strahan, 1999).

3 Rotation: Varimax versus Promax

In both Varimax and Promax a factor pattern is searched that approximates simple structure as close as possible. In Varimax it is ensured that the factors remain uncorrelated, in Promax not. Varimax rotation is mainly used for exploratory factor analysis. The other rotation methods are very similar to either Varimax or Promax and usually give comparable results.

The statement that Varimax is often used for exploratory factor analysis, is an observation, and not an instruction to imitate it. In my opinion an oblique rotation is generally better, because it will more often discover a simple structure. That is not just my opinion; it was also the opinion of Thurstone (Abdi, 2003) and Cattell (1971, pp. 16-20). Also, recent authors call predominantly for oblique rotation (Fabrigar, Wegener & MacCallum Strahan, 1999. Reise, Waller & Comrey, 2000; Stewart et al., 2001; Costello & Osborne, 2005). Despite the fact that it has been advised for half a century, this recommendation is widely ignored.

The interpretation of the factor pattern is often easier with oblique rotation. In terms of content, there is usually little reason why the factors should be uncorrelated. For example, why should fluid and crystallized intelligence be uncorrelated?

A problem for orthogonal rotation is also that one often wishes that the model applies not only in the entire population, but also in various subpopulations (e.g., men and women). This principle is called **factorial invariance** (Millsap, 1997, 2007a) or subpopulation invariance (Ellis, 1993; Ellis & Junker, 1997). In general, if two variables are uncorrelated, they can very well be correlated in subpopulations. That also applies to the factors. Therefore, there is little reason to believe that there is a model with orthogonal factors that also holds in subpopulations. The problem is even greater because some subpopulations may be overrepresented in the sample, and that could lead to the assumption of orthogonal factors being violated. For this reason, oblique rotation should be preferred.

Instead of Promax one can also use Oblimin. The literature is neutral as regards the choice of Promax and Oblimin. In this book, Promax is used most of the time.

4 Running in SPSS with raw data as input

If the data file contains the **raw data** (not the correlation matrix!), you can control the factor analysis from the **menu:**

Analyze > Dimension Reduction > Factor ...

Put the manifest variables under Variables

Click the Extraction... button
Select the extraction Method: Principal components or Maximum likelihood
- choose Fixed number of factors, in a confirmatory analysis; also indicate the number of factors
- 'Based on Eigenvalue, Eigenvalues greater than: 1' is often chosen in an exploratory analysis, but is not recommended

Continue
Click the Rotation ... button
Choose the rotation Method: Varimax or Promax
At Display, tick the Rotated solution and possibly Loading plot (s).
Continue
Click the Options ... button
- Choose 'Sorted by size' under Coefficient Display Format

Continue, OK

It is often convenient to select at Options under Coefficient Display Format 'Suppress small values, Absolute value below:', and to specify the value .30, or .50. This means that loadings of less than 0.30 or 0.50 will not be displayed, which often makes the pattern more clear. For the final report in an article all factor loadings have to be displayed (Henson & Roberts, 2006).

5 Running in SPSS with correlation matrix as input

If the data file contains the **correlation matrix,** you can control the factor analysis not from the menu but you have to use **syntax:**

Perform the above steps except the last one (OK).
Click Paste instead of OK.
It opens a new window with the so-called syntax.

For example, the content of the syntax window looks like this:

```
FACTOR
  / VARIABLES v1 v2 v3 v4 v5 v6 v7 v8 v9 v10 v11 v12 v13 v14 v15
  / MISSING LISTWISE
  / ANALYSIS v1 v2 v3 v4 v5 v6 v7 v8 v9 v10 v11 v12 v13 v14 v15
  / PRINT INITIAL EXTRACTION ROTATION
  / CRITERIA MINEIGEN (1) ITERATE (25)
  / EXTRACTION PC
  / CRITERIA ITERATE (25)
  / ROTATION VARIMAX
  / METHOD = CORRELATION .
```

A brief explanation of this: The **command** is FACTOR and after each / is a **subcommand.** A sub-command is, for example, / EXTRACTION, and a **keyword** in this subcommand is PC. The point behind CORRELATION is the closing command and although it seems unimportant, it can go wrong if you omit it – but you'll find out.

Remove the sub-commands VARIABLES, MISSING and ANALYSIS.
Add on that place the subcommand / MATRIX = IN (COR = *)

The above syntax then becomes:

```
FACTOR
  / MATRIX = IN (COR = *)
  / PRINT INITIAL EXTRACTION ROTATION
  / CRITERIA MINEIGEN (1) ITERATE (25)
  / EXTRACTION PC
  / CRITERIA ITERATE (25)
  / ROTATION VARIMAX
  / METHOD = CORRELATION .
```

Now you can execute this syntax:
Put the cursor in the command.
Click the play-button (or choose Run > Current).

Examples
1 In the running example (Diesfeldt's data of the Depression List) we choose ML extraction. Only one factor is hypothesised, therefore no rotation is possible.

2 However Diesfeldt actually conducted a PCA extraction with Varimax rotation. In an exploratory factor analysis this is not unusual. Still, my preference in an exploratory analysis would be to use ML extraction with promax rotation. The reason why I prefer ML extraction is that PCA assumes that there are no unique factors.

3 In the above example of the BPS, we have a quite specific hypotheses about the factor pattern. In principle we would therefore like to conduct a confirmatory factor analysis. Because we limit ourselves to SPSS, we cannot really conduct a confirmatory factor analysis with multiple factors, but we will try to approach it as closely as possible. We therefore choose ML extraction with Promax rotation.

Reading Question. Which extraction method and which rotation method would you use in an item analysis of a test?

2.10 Estimates

The main estimates are the **factor loadings,** which are put together in a table called the **factor pattern**. In addition, one reports the communalities, the eigenvalues and the correlations between the factors.

In an analysis of the correlation matrix (not the covariance matrix), the loading of a manifest variable on a factor is equal to the regression weight of that manifest variable on that factor. We will assume henceforth that each factor is standardised. If there is only one factor or if factors are uncorrelated, then the factor loading is equal to the **correlation between the manifest variable and the factor**.

The **communality** of a manifest variable is equal to the squared correlation multiple of that variable on the common factors. This indicates the percentage of the variance of the manifest variable that is explained by the common factors. If the factors are uncorrelated, this is equal to the **sum of the squared loadings of the manifest variable**.

The **eigenvalue** of an unrotated factor equals the **sum of the squared loadings on that factor.** For rotated but uncorrelated factors such sums of squares can be computed, and are then sometimes called eigenvalue, but that term is not correct in this case. Nevertheless, the computed quantity indicates how much variance the factor explains in the manifest variables. The maximum value is equal to the number of variables. The sum of squares divided by the number of variables is interpreted as the percentage variance that the factor explains in the manifest variables.

In SPSS, factor loadings must be looked up after rotation in the table **Rotated** Component Matrix or Rotated Factor Matrix or **Pattern Matrix**. Do not confuse this with the Factor Matrix and the Structure Matrix, which look very similar but have different meanings. If no rotation was performed, factor loadings are given in the Factor Matrix or the Component Matrix. The other tables are not displayed.

In SPSS, communalities should be looked up in the Extraction column, not in the Initial column. For the eigenvalues, it is the other way around: these are given in the Initial column. The sums of squared loadings may be important too, and these are displayed in the Rotation column.

Explanation

Historically, many types of factor analysis start the computational process by estimating the communalities and / or the eigenvalues (Mulaik, 1972). Therefore, these coefficients always get much attention in texts on factor analysis. However, for substantive interpretation they are much less important than the factor pattern.

Example 1

In this example, a confirmatory factor analysis was done on the correlations of Diesfeldt, with one factor and ML extraction. The factor pattern, communalities, the eigenvalues and factor correlations are found in the output. In this case they are (including most SPSS-output) shown in Table 2.6 to 2.9.

```
FACTOR
    / MATRIX = IN (COR = *)
    / PRINT INITIAL EXTRACTION
    / CRITERIA FACTORS (1) ITERATE (25)
    / ML EXTRACTION
    / ROTATION NORO TATE
    / METHOD = CORRELATION .
```

Table 2.6

Communalities

	Initial	Extraction
V1 Satisfied (-)	0.415	0.387
V2 Sleeping (-)	0.248	0.094
V3 Eating (-)	0.229	0.074
V4 Healthy (-)	0.406	0.296
V5 Tired	0.244	0.141
V6 Old	0.298	0.248
V7 Lonely	0.412	0.316
V8 Friends (-)	0.185	0.096
V9 Visitors (-)	0.397	0.228
V10 Gloomy	0.464	0.487
V11 Boredom	0.265	0.205
V12 Excited (-)	0.498	0.532
V13 Helpless	0.207	0.184
V14 Weak	0.417	0.209
V15 Future (-)	0.383	0.377

Extraction Method: Maximum Likelihood.

Table 2.7

Total Variance Explained

Factor	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4.555	30.369	30.369	3.874	25.829	25.829
2	1.603	10.684	41.053			
3	1.237	8.245	49.299			
4	1.018	6.790	56.088			
5	0.914	6.093	62.181			
6	0.815	5.436	67.617			
7	0.795	5.299	72.916			
8	0.758	5.051	77.968			
9	0.684	4.557	82.525			
10	0.567	3.782	86.306			
11	0.484	3.228	89.535			
12	0.453	3.017	92.552			
13	0.423	2.822	95.374			
14	0.357	2.379	97.753			
15	0.337	2.247	100.000			

Table 2.8

Factor Matrix(a)

	Factor
	1
V1 Satisfied (-)	0.622
V2 Sleeping (-)	0.306
V3 Eating (-)	0.273
V4 Healthy (-)	0.544
V5 Tired	0.375
V6 Old	0.498
V7 Lonely	0.563
V8 Friends (-)	0.310

V9 Visitors (-)	0.478
V10 Gloomy	0.698
V11 Boredom	0.453
V12 Excited (-)	0.729
V13 Helpless	0.429
V14 Weak	0.457
V15 Future (-)	0.614

Extraction Method: Maximum Likelihood.

a 1 factors extracted. 4 iterations required.

Table 2.9

Goodness-of-fit Test

Chi-Square	df	Sig.
230.364	90	0.000

If we may believe these factor loadings – which we know only after the statistical tests of the next section – we can conclude this:

- The highest loading is the one of v12 (Excited). This item correlates .729 with the common factor, the latent variable *Depressivity*. The communality is the square of it, .532. This shows how much variance of Cheerfulness is explained by *Depressivity: 53.2%*.
- The lowest loading is obtained for v3 (Food). This item correlates .273 with the factor *Depressivity*. The explained variance is .074.

In Figure 2.1, we may write these loadings next to the arrows from *Depressivity* to the items. The loading then indicates the strength of the corresponding arrow. (You can do this yourself. For an example, look at Figure 1 of Church et al. (1999).)

From the table of eigenvalues we can conclude that the factor *Depressivity* explains 25.829% of the variance of all the variables together.

Note, however, that the word '*Depressivity*' does not appear in the output. The factor is simply described as '*Factor 1*'. The data only indicate that there is a common factor. That '*Depressivity*' is a good name for it, is something we infer from the contents of the items. That also implies a limitation to the answer that we may give to the research question. The question was whether the items measure depression. But from factor analysis we may at best conclude that the items measure *the same* construct. Whether we shall call this '*Depressivity*' or not, does not follow from the data.

Reading Question. What is a factor loading? In which table of the output do you find the factor loadings ?

Example 2

In this example, an exploratory factor analysis on the correlations of Diesfeldt was conducted with PCA extraction and Varimax rotation. Moreover, in SPSS the command was given to sort the variables on the basis of factor loadings. The output is displayed in table 2.10 up to and including 2.13.

Table 2.10

Communalities

	Initial	Extraction
V1 Satisfied (-)	1.000	0.488
V2 Sleeping (-)	1.000	0.713
V3 Eating (-)	1.000	0.563
V4 Healthy (-)	1.000	0.581
V5 Tired	1.000	0.518
V6 Old	1.000	0.411
V7 Lonely	1.000	0.547
V8 Friends (-)	1.000	0.467
V9 Visitors (-)	1.000	0.663
V10 Gloomy	1.000	0.620
V11 Boredom	1.000	0.418
V12 Excited (-)	1.000	0.658
V13 Helpless	1.000	0.396
V14 Weak	1.000	0.744
V15 Future (-)	1.000	0.627

Extraction Method: Principal Component Analysis.

Table 2.11

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4.555	30.369	30.369	4.555	30.369	30.369	2.785	18.566	18.566
2	1.603	10.684	41.053	1.603	10.684	41.053	2.168	14.451	33.017
3	1.237	8.245	49.299	1.237	8.245	49.299	2.078	13.855	46.871
4	1.018	6.790	56.088	1.018	6.790	56.088	1.383	9.217	56.088
5	0.914	6.093	62.181						
6	0.815	5.436	67.617						
7	0.795	5.299	72.916						
8	0.758	5.051	77.968						
9	0.684	4.557	82.525						
10	0.567	3.782	86.306						
11	0.484	3.228	89.535						
12	0.453	3.017	92.552						
13	0.423	2.822	95.374						
14	0.357	2.379	97.753						
15	0.337	2.247	100.000						

Extraction Method: Principal Component Analysis.

Table 2.12

Component Matrix(a)

	Component			
	1	2	3	4
V12 Excited (-)	0.741	0.048	-0.018	-0.326
V10 Gloomy	0.725	0.083	-0.288	-0.069
V1 Satisfied (-)	0.661	-0.007	0.209	-0.083
V15 Future (-)	0.644	-0.036	0.166	-0.430
V7 Lonely	0.615	0.376	-0.084	0.146
V4 Healthy (-)	0.612	-0.379	-0.139	0.208
V6 Old	0.551	-0.241	0.216	-0.053

V9 Visitors (-)	0.531	0.417	-0.328	0.314
V11 Boredom	0.500	0.207	0.170	-0.309
V13 Helpless	0.471	-0.109	-0.297	-0.271
V14 Weak	0.509	-0.527	-0.356	0.283
V3 Eating (-)	0.317	0.511	0.401	0.202
V8 Friends (-)	0.354	0.480	-0.133	0.306
V5 Tired	0.440	-0.471	0.172	0.271
V2 Sleeping (-)	0.374	-0.173	0.671	0.305

Extraction Method: Principal Component Analysis.

a 4 components extracted.

Table 2.13

Rotated Component Matrix(a)

	Component			
	1	2	3	4
V15 Future (-)	**0.771**	0.116	0.025	0.137
V12 Excited (-)	**0.750**	0.204	0.229	0.029
V11 Boredom	**0.598**	-0.081	0.185	0.140
V1 Satisfied (-)	**0.548**	0.224	0.205	0.308
V10 Gloomy	**0.531**	0.361	**0.440**	-0.122
V13 Helpless	**0.483**	0.299	0.095	-0.254
V6 Old	**0.439**	0.346	0.003	0.313
V14 Weak	0.096	**0.849**	0.097	-0.067
V4 Healthy (-)	0.253	**0.693**	0.154	0.115
V5 Tired	0.119	**0.596**	-0.045	0.383
V9 Visitors (-)	0.138	0.192	**0.775**	-0.077
V8 Friends (-)	0.044	-0.007	**0.680**	0.050
V7 Lonely	0.340	0.115	**0.639**	0.100
V3 Eating (-)	0.153	-0.264	**0.490**	**0.480**
V2 Sleeping (-)	0.120	0.201	0.013	**0.811**

Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalization.

a Rotation converged in 7 iterations.

The Rotated Component Matrix contains the rotated factor pattern. Loadings greater than 0.4 are marked bold by me. As you can see, the previously discussed pattern returns here. A simple structure emerges, with a few exceptions.

Reading Question. What are the exceptions?

2.11 Plot of factor loadings

In a factor analysis with **two orthogonal** factors it may be helpful to plot the loadings. In such a plot, each factor is shown as an axis, and each variable as a point. The factor loadings of the variable are the coordinates of the point. If the factor model is correct, the correlations between the variables can be inferred from the plot. Variables that are close to each other, away from the origin, correlate strongly positive. Variables that are opposite to each other, far away from the origin, correlate strongly negative. Variables that, viewed from the origin, are perpendicular with each other, have correlation zero with each other. Variables that are close to the origin, have correlations close to zero with all other variables.

Explanation

If there is only one factor, one can make a plot of the loadings, but this plot will have only one axis. SPSS clearly feels too good for this and refuses to make it. If there are more than two factors, the plot cannot be properly drawn on paper, since such drawings can only be two-dimensional. SPSS can plot a projection of the first three factors, which you can rotate interactively, but I've never seen a case in which this was useful.

In an oblique rotation the correlations cannot be inferred from the plot. The plot is less meaningful in this case.

The correlations that one infers from the plot are predicted by the model. These are called the **reproduced correlations** (see next section). They may differ from the observed correlations that are calculated directly from the data.

How correlations can be inferred from the plot is discussed in the section on visualisation at the end of this chapter.

Example

In an analysis that focuses on data reduction, I sometimes use the following method to get an overview of the correlations: A PCA with two factors and varimax rotation. In addition, I let the factor loadings be plotted. The thus rotated factor pattern from the example of Diesfeldt is shown in Table 2.14, and the corresponding plot is shown in Figure 2.2.

Table 2.14

Rotated Component Matrix(a)

	Component	
	1	2
V14 Weak	0.728	-0.081
V4 Healthy (-)	0.713	0.099
V5 Tired	0.639	-0.082
V6 Old	0.578	0.166
V12 Excited (-)	0.540	0.510
V15 Future (-)	0.519	0.383
V1 Satisfied (-)	0.514	0.416
V13 Helpless	0.432	0.217
V2 Sleeping (-)	0.398	0.106
V7 Lonely	0.233	0.681
V9 Visitors (-)	0.143	0.660
V8 Friends (-)	-0.034	0.596
V3 Eating (-)	-0.082	0.595
V10 Gloomy	0.506	0.526
V11 Boredom	0.253	0.479

Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalization.

a Rotation converged in 3 iterations.

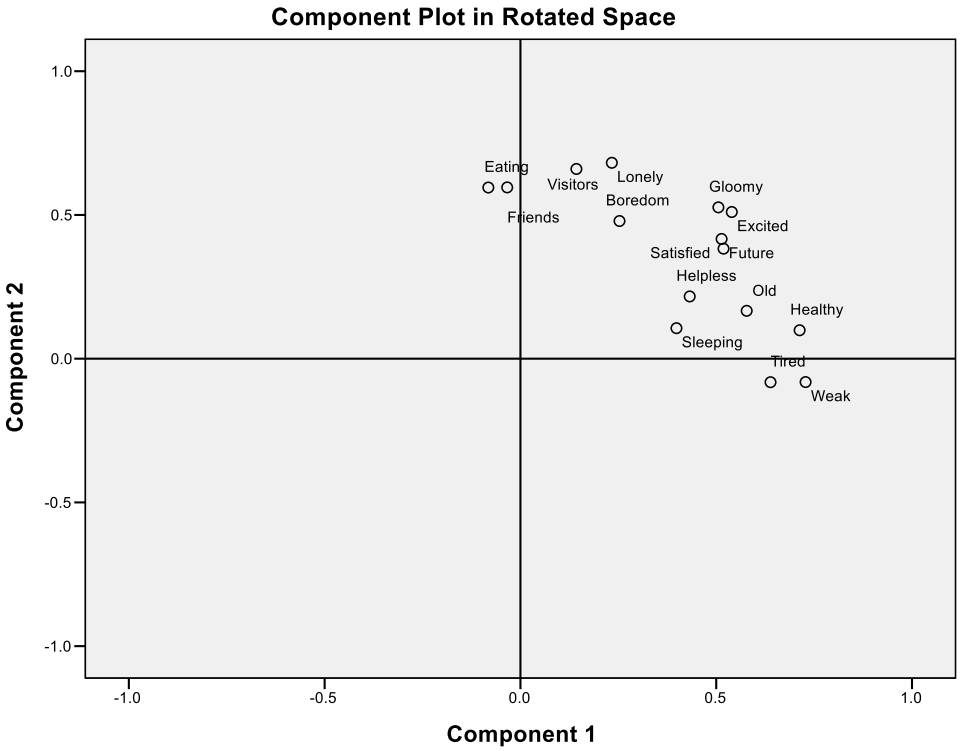

Figure 2.2

Such a plot can be very useful if the variable names aren't printed over each other. In the plot you can see, for example, that Gloomy and Excited have a relatively high correlation with each other, while Healthy and Eating hardly correlate with each other.

According to Diesfeldt there are four factors, but in this section two factors are drawn. Is that allowed? No. But the aim was now alone data reduction, not theory.

2.12 Test statistics

Based on the pattern factor and the correlations between the factors one can predict the values of the correlations between the manifest variables. These predicted correlations are called **reproduced correlations**. The difference between the observed and the reproduced correlation is called a **residual correlation.** Evaluation of the validity of a factor model is in principle based on the residual correlations (Mulaik, 1972; Browne, 1974, 1984; Jöreskog & Sörbom, 1996). This can be done in the following ways.

The *p*-value. The size of the residual correlations is usually summarised in a statistic that, under the null hypothesis, asymptotically has a **chi-square** (χ^2) distribution. The higher the residual correlations, the greater χ^2 (Tucker & Lewis, 1973; Browne, 1984; Jöreskog & Sörbom, 1996 Flora & Curran, 2004). On this basis, the *p*-value is calculated, and the null hypothesis is rejected if the *p*-value is too small (less than 0.05). In a purely confirmatory analysis, the chi-square contains at the same time a test for the number of factors and the specified pattern. In SPSS such an analysis is not possible with multiple factors, and the chi-square will test only the number of factors. In that case, it will be still be needed to check manually if the pattern factor corresponds to the hypothesis. The *p*-value is usually computed after ML extraction, and some of the other extraction methods (e.g., ULS) will not produce a *p*-value.

A goodness-of-fit index. Because a small violation of the model often can lead to a significant violation of the model if the sample is large (Tucker & Lewis, 1973; Hu, Bentler & Kano, 1992), it is recommended to use **goodness-of-fit** index which reflect the size of the violation (similar to Cohen's *d* in a *t*-test). There are many goodness-of-fit indices (for discussion see Finch & West, 1997; Jackson, Gillaspy & Purc-Stephenson, 2009). We will discuss only the Root Mean Square Error of Approximation (RMSEA) (Steiger, 1990), because its theoretical foundation best reflects that the null hypothesis is not exactly the hypothesis that we want to test (Raykov, 1998). The RMSEA is not calculated by SPSS; you have to do it manually. The formula is

$$RMSEA = \sqrt{\frac{\frac{\chi^2}{df} - 1}{N - 1}}$$

$RMSEA = 0$ if the value under the root sign is negative.

Here N is the number of subjects. RMSEA values less than 0.08 represent an acceptable fit and values of 0.05 or less represent a good fit (Browne & Cudeck, 1993). Note that it is actually a *badness*-of-fit index.

The eigenvalues. In ancient forms of factor analysis a PCA is done in the first stage. A classic criterion is to retain only factors with eigenvalue greater than 1 (Guttman, 1954; Kaiser, 1960, 1961). This is also known as the Guttman-Kaiser criterion, the Kaiser-Guttman criterion, the Kaiser criterion, the Guttman criterion or the minimum eigenvalue criterion (but the name ‘Kaiser criterion’ should be discouraged, because there is already a ‘Kaiser Standardization’ in rotation, which is something quite

different). The intuitive appeal of the criterion is that factors with a smaller eigenvalue than 1 have an explanatory variance that is smaller than the variance of one manifest variable. Such factors are not very influential. Although the Guttman-Kaiser criterion is one of the most commonly used criteria, there is hardly any statistical justification for. Therefore it is not appropriate for a confirmatory factor analysis, although it is perhaps useful in an exploratory factor analysis. However, the latter is questionable, and the popularity of the Guttman-Kaiser criterion is perhaps partly due to the fact that it is the default in SPSS (Yeomans & Golder, 1982 Reise, Waller & Comrey, 2000). Fabrigar et al (1999, p. 278) even write: "In fact, we know of no study of this rule that shows it to work well."

Another method which is based on the eigenvalues, is the *scree test* (Cattell, 1966). In this method, a plot is made of the successive eigenvalues. If there is a sharp bend in the plot, this indicates the number of factors. This works well when the number of factors is clear. Finch and West (. 1997, p 466) comment: "The primary problem with the scree test Is that it is an 'eyeball test'; the point of the break in the plot can difficult to determine or there may be more than one such break. In such ambiguous cases, analysts can easily reach different conclusions concerning the proper number of factors to extract (Kaiser, 1970)."

Perhaps the best method based on eigenvalues is a so-called *parallel analysis* (Horn, 1965). Here, the eigen values are compared with the eigenvalues of a large number of simulated samples. See Finch and West (1997), Fabrigar et al. (1999) and Reise, Waller and Comrey (2000) for further discussion. This method is not implemented in SPSS, but in the free statistical software package R.

Reading Question. What is the formula for RMSEA?

Explanation

1 When reporting an article

Jackson, Gillaspy and Purc-Stephenson (2009) provide an overview of the ways in which, in practice, confirmatory factor analyses are reported in APA journals. They recommend in a confirmatory factor analysis to report always the following measures:

1. chi-square, number of degrees of freedom, and *p*-value;
2. an incremental fit index such as the Tucker-Lewis index or the comparative fit index; and
3. a residue based fit index such as the RMSEA.

In this book we use only (1) and (3), because this is only an introduction. For the benefit of readers who want to follow the above recommendation, the Appendix to this chapter describes how to calculate the indices of (2).

2 Criticism of the use of fit indices

The use of fit indexes is debatable. First, the cutoff at 0.05 is not robust in the sense that there might be other cutoffs that should be used for non-normal distributions (Yuan, 2005). Second, a violation that is 'small' as measured by fit indices, can be important. Fit indices measure only the overall fit of the model, but are insensitive to local violations (Fan and Sivo, 2005, 2007, Saris, Satorra & Van der Veld., 2009; Heene et al, 2012) and say little about the importance of the causal relationship that created the violation (Barrett, 2007; Hayduk et al., 2007). Consider the imaginary situation in which two intelligence items correlate more with each other than the model predicts, and that the cause of this is that both items depend in part on skin color. This will show up as a large residual correlation. However, if the other items do satisfy the model, then that single residual correlation will change little to the value of a fit index, and the index will therefore indicate that the fit is good. But is this violation unimportant? Not when you are assessed on intelligence during a job application and these two items cause you to be dismissed. "Close but significant ill-fit in SEM-speak, translates as 'close to being sued' in legal-speak" (Hayduk et al., 2007, pp. 848-849). A small violation is not necessarily a trivial violation, and some authors therefore recommend to also investigate the residues (Hayduk et al., 2007; Jackson, Gillaspy & Purc-Stehpenson, 2009). But how exactly this should done and whether it is a solution is not clear, and there are other possible approaches (Saris, Satorra & Van der Veld, 2009). Mulaik (2007, p. 890) notes that the history of science is one of increasingly better approximations in each field. Wegener's theory of continental drift could occur, for example because the coastlines of Africa and South America approximately fit into each other. According Mulaik it would therefore be ridiculous to banish the idea of approximations from science. There must be room for approximations. Nevertheless, it is possible to have valid criticisms to the currently existing indexes.

The criticism to fit indices is primarily that they are sometimes too *optimistic*, in the sense that they can qualify a violation as 'small' even when it is important. To my knowledge, the criticism is never that established fit indices would sometimes be too *pessimistic* in the sense that they can qualify a deviation as 'big' when it is actually unimportant. If the chi-square is significant and the RMSEA is large, there is therefore no doubt that the model does not fit the data well. In other cases, discussion may arise. As will be argued in the next chapter, you should always compare the model with alternative models.

Example 1

In the confirmatory factor analysis with one factor for the Diesfeldt data we use the part of the output displayed in table 2.15. Based on the *p*-value, the model of a single factor for all items would be rejected. On the basis of this output, the RMSEA can be calculated as 0.089. This, too, indicates that the one-factor model is not suitable for these data.

Table 2.15

Goodness-of-fit Test

Chi-Square	df	Sig.
230.364	90	0.000

Example 2

In the exploratory factor analysis that Diesfeldt himself conducted, he deployed the minimum eigenvalue criterion. The relevant part of the output is displayed in table 2.16. In the Initial Eigenvalues/ Total column, there are four eigenvalues which are greater than 1. On this basis, there are four factors retained. (Earlier we discussed just three factors; we will later discuss the fourth factor).

Table 2.16

Component	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	4.555	30.369	30.369
2	1.603	10.684	41.053
3	1.237	8.245	49.299
4	1.018	6.790	56.088
5	0.914	6.093	62.181
6	0.815	5.436	67.617
7	0.795	5.299	72.916
8	0.758	5.051	77.968
9	0.684	4.557	82.525
10	0.567	3.782	86.306
11	0.484	3.228	89.535
12	0.453	3.017	92.552
13	0.423	2.822	95.374
14	0.357	2.379	97.753
15	0.337	2.247	100.000

2.13 Decision

In a confirmatory factor analysis, the decision contains an assessment of the validity of the hypothesis. If the hypothesis is found to be invalid, it should be described – if possible – whether the violation concerns the number of factors or the presumed factor pattern. In the latter case, also try to indicate the loadings that violate the model. An analysis in SPSS does not provide significance tests for the loadings, and rejection of the null hypothesis thus leads to the conclusion that more factors are needed.

The usual practice is to consider not only the *p*-value, but also the goodness-of-fit indices and even the interpretation (see the next section). In this book you should adhere to the rules listed in table 2.17. The boundaries of 'acceptable' and 'good' are based on Browne and Cudeck (1993). According to this table, the decision should be based largely on the *p*-value and the RMSEA, but in some cases the interpretability plays a role. However, there is criticism possible to this decision rules (see notes).

Table 2.17 *Decision rules about number of factors*

P-value	*Goodness-of-fit*	*Meaning*	*Decision*
$p > 0.05$		Retain the null hypothesis	The model has enough factors
$p < 0.05$	*RMSEA* < 0.05	The null hypothesis not exactly true, but it has a good fit	The model has enough factors
$p < 0.05$	*RMSEA* > 0.05 *RMSEA* < 0.08	The null hypothesis is not exactly true, but has an acceptable fit	Doubt; decision depends on competing theories, interpretation and other factor analyses
$p < 0.05$	*RMSEA* > 0.08	the null hypothesis is not true, and the fit is poor	The model has too few factors

In a confirmatory fixed-model analysis in test construction one has to specify which items do not fit the model. Those are

- Items that have a high loading on another factor, and
- *probably* items that have a low load on the intended factor.

In an exploratory factor analysis, there is no hypothesis and therefore one cannot decide about it. Nevertheless, it is useful to report in this section what you decided about the number of factors.

Explanation

1 The null hypothesis is the research hypothesis

Earlier you learned in statistics that the null hypothesis is usually the inverse of the research hypothesis, that you're 'happy' if the null hypothesis is rejected. But if you really have a good memory, you also remember that there is a rule that is even more important: the null hypothesis must always be testable. This latter rule is now important. The research hypothesis in factor analysis is that there are at most *k* common factors. This hypothesis is been testable, but the reverse hypothesis, that there are more than *k* factors is not testable. The correlation matrix can always be reproduced perfectly by assuming as many factors as there are manifest variables. Therefore, in factor analysis **the null hypothesis must be equal to the research hypothesis.** Therefore, you should be happy if the test **is not** significant.

If you remember our earlier messages of how some psychologists deal with statistics, you will surely understand that this reversal sometimes leads to hilarious moments. For example, one student happily visits her supervisor, as a factor analysis with the hypothesised three factors showed a large *p*-value, and the supervisor is disappointed because the result is "not significant." (Please note that in this example the student is right). Or conversely, the student is happy because after days of "data massage" she finally got the data so far that the result is “significant”. Oh well, with any luck that person also confuses whether one should retain or reject the null hypothesis if $p < 0.05$, so it might still be good.

2 Huh, retain the null hypothesis when $p < 0.05$?

In Table 2.17 decision rules are given that deviate from the rules that we used until now in *t*-tests and ANOVA. There you could never maintain the null hypothesis after a significant result. It’s different here, sometimes. Why?

The research hypothesis is the null hypothesis, but the problem is that the null hypothesis is so exact that there will always minor deviations be found in practice if the sample is large enough. These abnormalities may be unimportant. The question is not whether the null hypothesis is *exactly* true, but whether it is *approximately* true (Tucker & Lewis, 1973; Steiger, 1990; Hu, Bentler & Kano, 1992; Browne & Cudeck, 1993; Raykov, 1998). That's why you often base the decision not only on the *p*-value, but also the goodness-of-fit. This is compatible to what you previously learned about effect size in ANOVA.

Only if the deviation is non-significant ($p > .05$) the above argument does not apply, and then you look at the *p*-value. In that case there is, moreover, still the problem that the sample might be too small. Whether the standard bounds for the RMSEA are exceeded, is in itself a testable hypothesis (MacCallum, Browne & Sugawara, 1996). On this basis Steiger recommends only to retain the null hypothesis if the RMSEA is

significantly smaller than 0.07, which is possible only if the sample is large enough. How to test this, is too complicated to discuss here.

There is criticism possible on the decision rules given here. The previous section has already discussed criticism on fit indices such as RMSEA. Someone who thinks that you should not use fit indices, will not endorse the above decision rules. In addition, one may criticize the recommendation to look at the interpretation when deciding on the number of factors. In short, the decision in factor analysis is not always as straightforward as you might think. The journal *Personality and Individual Differences* devoted a special issue to the question in 2007 whether and how to make that decision. And all eleven articles obviously disagreed (Barrett, 2007; Bentler, 2007; Goffin, 2007, Hayduk et al., 2007; Markland, 2007; McIntosh, 2007; Miles & Shevlin, 2007, Millsap, 2007 b; Mulaik, 2007; Steiger, 2007). The editors (Vernon & Eysenck, 2007, p. 813) concluded thus: "The diversity of the opinions overexpressed in these papers suggests that the best way to evaluate model-fit is still an evolving issue which will continue to be a subject of debate." As this debate lasts for 70 year already, you can assume that you will not live long enought to see the solution. Nevertheless, it is desirable to provide some simple rules of thumb in this introduction, so that you at least take a defendable decision.

3 What to do with doubt

A third deviation in the decision rules is that there is room for doubt. I guess that some students do not love that. But the possibility of doubt reflects better the actual situation than a black or white decision. The fact is simply that we sometimes do not know the answer because the data do not provide clarity.

Let's see what you should do in that situation. In itself an RMSEA between 0.05 and 0.08 provides no hard evidence to refute the theory. In a purely confirmatory factor analysis, and if there is no competing theory with acceptable fit, it is logical to retain the theory in this case. If there is a competing theory with an acceptable fit, an assessment must be made taking into account the fit and the parsinomy of both theories. That's beyond the scope of this chapter.

In a more exploratory analysis, there is always the competing theory that one needs one factor more than in the current analysis. Or two factors. Or three. And so on. Or one factor less. Or two. And so on. In short, in this case there is a series of competing theories, some of which will also have an acceptable or even good fit. Since there was no a priori theory, the fact that the current analysis resulted in an acceptable fit, is no convincing reason to decide that this is the correct number of factors. In that case, it is necessary to do multiple factor analysis, each with a different number of factors. The decision should be taken in this case by an assessment (see next chapter) of interpretability, fit and parsinomy.

Example 1A
In the example of Diesfeldt we did a confirmatory factor analysis with one factor. Because $p < 0.05$ and RMSEA > 0.08 we conclude that one factor is not enough to explain the correlations. There are many items with a low loading (< 0.5). That may be caused by the presence of a second factor. These items are eligible for removal from the Depression List (see section 2.14 INTERPRETATION).

Example 1B
In Example 1A, we used a fixed-model analysis: after rejecting the hypothesis we remove items that do not fit. In a fixed-domain analysis, on the other hand, the solution is to assume more factors, without deleting items. That means that follow-up analyses should be done, with the number of factors made larger until the fit is good enough. Since we have no a priori hypothesis anymore (because it was rejected in Example 1A), this is now an exploratory analysis. Table 2.18 shows the results for the *p*-value and the RMSEA.

Table 2.18

number of factors	*Chi-square*	*df*	*Sig.*	*N*	*RMSEA*
1	230.364	90	0.000	197	0.089
2	151.346	76	0.000	197	0.071
3	91.694	63	0.011	197	0.048
4	57.626	51	0.244	197	0.026

If we use only the *p*-values, the conclusion is that there are at most four factors, in agreement with what Diesfeldt concluded on the basis of the minimum eigenvalue criterion. If we consider only the RMSEA-values, it is concluded that good fit is achieved with three factors, and an acceptable fit with two factors. There are two to four factors. The further choice between them should depend on the interpretability of the factor patterns, the efficiency of the resulting theory, and the fit. This will be further discussed in the next chapter.

Example 2
Four factors with eigenvalue greater than 1 were found in the exploratory factor analysis of Diesfeldt. If the Guttman-Kaiser criterion is used, therefore, the decision is that there are four factors. This criterion is not recommended (see section 2.12 Assessment).

> *Reading Question.* Which combination of *p*-value and RMSEA will certainly reject the null hypothesis? What do you need after rejection: more factors, fewer factors, more items, less items, more people, fewer people, or a combination?

2.14 Interpretation

In the interpretation section you describe the meaning of the factors. In a confirmatory factor analysis, this follows from the hypothesis. In an exploratory factor analysis you determine this on the basis of the factor pattern and the content of the items. The variables with the highest factor loadings (say, > 0.90) correlate highly with that factor and are the most decisive for the name of the factor. If the interpretation of a factor is dubious because there are few variables with a high loading on the factor, or because no characteristic feature in the content of these items is discovered, state these doubts. The name of a factor should – to avoid confusion – preferably not be the same as the name of a manifest variable, and one factor should not be defined by only one item (Henson & Roberts, 2006).

In the case of an item analysis you also describe what the consequences are for the way the items are used in the scoring of the scale. In addition, you give

- whether it is justifiable to use a single overall score per person, and if not
- into which sub-scales the item set should be divided, and / or
- which items have to be removed.

For this you rely on the factor pattern :

- if there is only one factor, this justifies the use of a single total score per person
- the subscales consist of the factors, and each subscale consists of the items highly loading on one factor and only on that factor.

Items that load low on each factor, or that load high on several factors, must be removed. That is, they are not used in a (sub)scale to determine a score and they can be omitted from the test.

The next question is at which values one has to consider a loading as 'high' or 'low'. Floyd and Widaman (1995, p. 294) write : "In exploratory analysis, factor loadings are generally considered to be meaningful when they exceed .30 or .40." Henson and Roberts (2006, p. 402) report on the basis of 37 articles that in practice a cutoff of 0.30 to 0.50 is being used, with 0.40 as median. This says that many people have the thought that the bound should be around 0:40; but any justification for that thought is missing. Lambert, Wildt and Durand (1991) argue that such rules of thumb erroneously ignore the effects of sampling fluctuations. A loading of 0.50 in the population might be 0.30 in the sample, or vice versa. Without confidence intervals of the loadings, little can be said about it. Therefore, it is not yet possible to give a well-founded rule of thumb. In a confirmatory analysis, the question is irrelevant because if SEM is used, it is tested whether the loadings which would have to be ‘low’, are equal to zero.

Explanation

The factors are latent variables and it is not immediately clear what their meaning is. Therefore they will be labeled in the output as *Factor 1, Factor 2,* and so on. For theory formation this is only useful if a substantive meaning can be attached to the factors.

Indeed, when reading this chapter, you might occasionally have felt some discomfort. We assume a latent variable that we vaguely describe as a common factor. But what is the nature of that variable? That is exactly what you need to think about in the interpretation phase.

The interpretation phase has a fundamentally different course than with analysis of variance and *t*-tests, in which only manifest variables were used. The question there was what the decision meant for causality, but it was never asked what the *variables* meant. In factor analysis, this is the one of the main questions.

At an abstract level, it is, actually still comparable. For causality is not observable, and therefore latent. That is why it is sometimes so hard to say anything about it. Especially, inventing plausible confounds is difficult for many students. That's because confounding variables are usually latent. You could say that a common factor is actually one big confounding variable. Since factor analysis has a different objective, we do not see that as negative but as positive, provided that we get a clear picture of that variable.

In an exploratory factor analysis, it is therefore important that we get a clear picture of the factors. Usually this amounts to inventing good names for the factors. But it is supposed to be more than a word game. What matters, is that you have a *simple rule* describing which items load highly on which factor. You must seek the characteristic feature in the content of the items that have loadings on the factor. That characteristic feature should preferably enable predictions about new items (i.e., all items that have this feature load high on the factor, all items that do not have this feature load low on the factor).

By the way, it frequently happens that it is not possible to find such a simple rule. One possibility is that there are very few items with high loading on the factor. Usually it is then possible to recognise more than one common feature, which means that the meaning of the factor is unclear. In the running example (examining Diesfeldt the Depression List) the fourth factor, for example, has only the item *Sleeping;* In addition, the loading of *Eating* is also high, but this item also loads on another factor. What is the characteristic feature here?

If the factor pattern is not interpretable, this can be a reason to opt for a larger or smaller number of factors, and therefore redo the analysis with the new number of factors. In doing so, exploratory factor analysis, at least as it is used in psychology, has a larger subjective component than analysis of variance. This is also a reason why many prefer confirmatory factor analysis.

In a confirmatory factor analysis the interpretation is, in principle, less problematic than in exploratory factor analysis:

- In a confirmatory analysis, if the hypothesis is retained, the hypothesis already describes the meaning of the factors, and then it is sufficient to refer to it, and to identify any abnormalities or additional interpretations.

- In a confirmatory factor analysis, if the null hypothesis is rejected, you will not actually believe in the existence of common factors, and it makes no sense to interprete them.

Example 1
In this example, we did a confirmatory factor analysis with one factor. We concluded that the model should be rejected. The interpretation is: the common factor *Depressivity* does not exist. There is no factor that is the only factor governing these items. In other words, these items do not measure all the same. And therefore, they do not measure only depression. Maybe some items measure depression and other items measure something else. Or maybe the construct 'depression' is not good and there are actually two or more different types of depression. Anyway, it's not justified with these items to use one test score per person.

Because the analysis was fixed-model, the next question is which items can be removed to come to an acceptable scale. These are mainly items that have a low loading in the analysis of a single factor; So especially Eating, Sleeping and Friends. As a rule of thumb, I often require that an item loads at least 0.50 on the intended factor. But as argued above, there is really no good cutoff possible.

If we apply the bound of 0.50 to the loadings, only the following items are retained. The items are sorted by decreasing load (see Table 2.19).

Table 2.19

V12 Excited (-)	0.729
V10 Gloomy	0.698
V1 Satisfied (-)	0.622
V15 Future (-)	0.614
V7 Lonely	0.563
V4 Healthy (-)	0.544

This rule has the appearance of objectivity, but the next item has a load of 0.498 and you probably understand that this is so close to 0.50 that you could include it as well. But the next item has a loading of 0.478, so would you include that one as well? Does it ever stop?

If the aim is to avoid including items that have a negative contribution to the reliability, then it is not the height of the loading itself, but its relationship with the other loadings that matter. For example, suppose that 10 items are essentially tau equivalent, and that the last item has loading 0.5. If the loadings of the other items are

0.7, then it can be calculated that the last item has a negative effect on the reliability of the total score. But if the other items have loading 0.5, the last item has a positive contribution to the reliability of the total score. Whether a loading of 0.5 is high enough, therefore, depends on loadings of other items. Only when an item has a small loading compared with the other items, it will have a negative impact on the reliability. *Thus, there is no absolute bound possible* and it is pointless to debate whether it should be 0.30, 0.50 or 0.70.

Anyway, we should decide. Because there is still a reliability analysis that should be done later and in which more items can be removed, I would remove a minimal number of items and suggest this scale:

> *Spirited*: Excited, Gloomy, Satisfied, Future, Lonely, Healthy, Old, Visitors, Weak, Boredom, Helpless, Tired
>
> Removed: Eating, Sleeping, Friends

After this, the one-factor model should be tested again for the remaining items.

Example 2
In this example, we did an exploratory factor analysis, which resulted in four factors. On the basis of the loadings, Diesfeldt called these factors *Spirited*, *Health*, *Social* and *Sleep*. The last factor only has the item Sleeping. Therefore, one can hardly consider it as a *common* factor. This factor cannot be interpreted. The factors *Health* and *Social* both have only three items with loadings that exceed 0:50. Due to this small number it can be difficult to determine whether the characteristic feature is indeed understood. Therefore, the interpretation of these factors is somewhat dubious.

For the scale construction we only use three factors: *Spirited*, *Health* and *Social* (given the doubts about the latter two factors, one could also decide to only use *Spirited*). The items Gloomy, Helpless, Eating should be removed from the list, because they load low on all factors or high on two factors. Sleeping should also be removed, because it belongs to a factor that is removed. The resulting scales are thus:

> *Spirited:* Future, Excited, Boredom, Satisfied
> *Health:* Weak, Healthy, Tired
> *Social:* Visitors, Friends, Lonely

Next, we have to conduct a reliability analysis to examine whether these scales contain enough items – and it looks bleak.

2.15 Summary basic report

In table 2.20 the differences between exploratory and confirmatory factor analysis are put together. This is just a recommendation, not a description of what is actually done by everyone.

Table 2.20

	Confirmatory	*Exploratory*	*Data reduction*
Goal	Test the hypotheses	Forming hypothesis	Reduce number of variables
Hypotheses	Number of factors and scheme of factor pattern	-	
Extraction	ML	ML	PCA
Number of factors	Number of scales	Vary in multiple analyses	Depending on purpose, often 1 or 2
Rotation	Promax *	Promax	None or varimax
Decision	*p*-value RMSEA**	Comparison of analyses on *p*-value, RMSEA, and interpretability	None or Eigenvalues
Interpretation of factors	Follows from hypothesis	Characteristic feature	None
Effect on scale	Remove items	Split into subscales	Not applicable

*) See note in the text; **) also see Table 2.17.

In the Confirmatory column, the rotation 'promax' is specified, but note the following. In a confirmatory factor analysis with one factor, no rotation is possible. A confirmatory factor analysis of multiple factors must be preferably be conducted in SEM, and then rotation is not an issue. However, when it is desired to approach a confirmatory factor analysis of multiple factors in SPSS, then it is best to use an oblique rotation such as promax.

The column Exploratory states that one must make a decision by comparing multiple analyses. This is discussed in the next chapter. The basic principle is: many confirmatory analyses together form an exploratory analysis. Indeed, if one randomly examine all hypotheses, one is looking for a good hypothesis, which is exploratory.

In practice, the number of factors in the factor analysis is often based on the eigenvalues, for example, the Guttman-Kaiser-criterion (number of eigenvalues > 1) or the scree test. These methods are not recommended here, but if someone uses them, then the resulting analysis should be viewed as exploratory, because the number of factors is not determined by a hypothesis.

If we agree that the decision in a confirmatory factor analysis should be based on the *p*-value and RMSEA and possibly other fit indices (see paragraph 2.18), then it seems illogical to base the decision in an exploratory analysis on other statistics, such as the eigenvalues. After all, the hypothesis which is formed in an exploratory analysis, should later be evaluated with a confirmatory study. If one uses different criteria, this process might not converge.

2.15.1 Example 1

Design
Manifest variables: the 15 items from the Depression List.

Degree of control
Passively observed.

Aggregated data
The correlation matrix of items (see Table 2.3).

Hypotheses
A confirmatory factor analysis with one factor. In case of violation, the analysis will be based on a fixed model, because it concerns the construction of a scale.

Analysis method
ML extraction; no rotation is possible.

Estimators
The factor loadings (from which the communalities can be derived by squaring) are shown in Table 2.21. (This is essentially the SPSS-output Table 2.8, but formatted differently. One may prefer to display loadings in two decimal places. This is not done here to avoid confusion. One should preferably also report the largest eigenvalues, even though they are not relevant in a confirmatory factor analysis.)

Table 2.21 Factor Loadings

Item			*Factor loading*
Satisfied	v1	(-)	.622
Sleeping	v2	(-)	.306
Eating	v3	(-)	.273
Healthy	v4	(-)	.544
Tired	v5		.375
Old	v6		.498
Lonely	v7		.563
Friends	v8	(-)	.310
Visitors	v9	(-)	.478
Gloomy	v10		.698
Boredom	v11		.453
Excited	v12	(-)	.729
Helpless	v13		.429
Weak	v14		.457
Future	v15	(-)	.614

Test statistics
(Table 2.9 :) $\chi^2(90) = 230.364$, $p < .001$, RMSEA = 0.089.

Decision
The items do not satisfy the one-factor model and the violation is not acceptable (RMSEA > 0.08). There are many items with a low loading (< 0.50), and perhaps these items depend on another factor.

Interpretation
There is not one common factor 'depressivity' that describes these items properly. The items do not measure all the same factor. There is no justification with these items to use one test score per person. The item list should be divided into multiple scales or some items should be removed. In the latter case, the items Eating, Sleeping, and Friends should probably be removed.

2.15.2 Example 2

(This is essentially the analysis reported Diesfeldt.)

Design
Manifest variables: the 15 items from the Depression List.

Degree of control
Passively observed.

Aggregated data
The correlation matrix of items (see Table 2.3).

Hypotheses
An exploratory factor analysis.

Analysis method
Principal component analysis with varimax rotation.

Estimators
The factor loadings (from which the communalities can be computed) are shown in Table 2.22. (This is essentially the SPSS-output of Table 2.13, but formatted differently. One may prefer to display loadings in two decimal places. This is not done here to avoid confusion.)

Table 2.22 Factor Pattern of four varimax-rotated principal components

Item			*Spirited*	*Health*	*Social*	*factor 4*
Future	v15	(-)	**.771**	.116	.025	.137
Excited	v12	(-)	**.750**	.204	.229	.029
Boredom	v11		**.598**	-.081	.185	.140
Satisfied	v1	(-)	**.548**	.224	.205	.308
Gloomy	v10		**.531**	.361	**.440**	-.122
Helpless	v13		**.483**	.299	.095	-.254
Old	v6		**.439**	.346	.003	.313
Weak	v14		.096	**.849**	.097	-.067
Healthy	v4	(-)	.253	**.693**	.154	.115
Tired	v5		.119	**.596**	-.045	.383
Visitors	v9	(-)	.138	.192	**.775**	-.077
Friends	v8	(-)	.044	-.007	**.680**	.050
Lonely	v7		.340	.115	**.639**	.100
Eating	v3	(-)	.153	-.264	**.490**	**.480**
Sleeping	v2	(-)	.120	.201	.013	**.811**

Test statistics
The number of factors is determined by the minimum eigenvalue criterion. There were four eigenvalues greater than 1 (see Table 2.11, column Initial Eigenvalues, Total).

Decision
There are four factors with an eigenvalue greater than 1. Therefore, four factors are retained.

Interpretation
On the basis of the loadings, the factors may be interpreted respectively as *Spirits*, *Health*, *Social* and *Sleep*. The fourth factor is not interpretable, because only one item loads on it. Interpretation of *Health* and *Social* dubious, since there are only three items that load on these factors. The items Gloomy, Helpless, Eating should be removed from the list, because they load low on all factors or high on two factors. In addition, in this case the item Sleeping has to be removed, because it belongs to a factor that is removed. The resulting scales which are eligible for further reliability analysis, are therefore:

Spirited: Future, Excited, Boredom, Satisfied
Health: Weak, Healthy, Tired
Social: Visitors, Friends, Lonely

2.16 Concise Report

A concise report (or short report) is a summary of a basic report in a continuous text that can serve as the basis for a report in an article. The concise report contains

- the design;
- the analysis (possibly with the hypothesis);
- the test statistics;
- decisions;
- the core of the interpretation.

For guidelines on reporting in an article, see Floyd and Widaman (1995), Fabrigar, Wegener, MacCallum and Strahan (1999), McDonald and Ho (2002), Costello and Osborne (2005), Henson and Roberts (2006), Jackson, Gillaspy and purc-Stephenson (2009) and Ten Holt, Van Duijn and Boomsma (2010).

2.16.1 Example 1

With the 15 items from the Depression List a confirmatory ML–factor analysis was conducted for the hypothesis that there is one common factor for these items. There was a significant violation of the model (χ^2 (90) = 230.364; $p < 0.001$). The fit was not acceptable (RMSEA = 0.089). On this basis the one-factor model was rejected for these items. The conclusion is that the items do not constitute one scale.

2.16.2 Example 2

The 15 items of the Depression List were examined with an exploratory principal component analysis with varimax-rotation. There were four factors with an eigenvalue

> 1 (the eigenvalues were 4.56, 1.60, 1.24 and 1.02). The fourth factor has no obvious interpretation. The other three factors can be described as *Spirited*, *Health* and *Social*. For the scale construction, five items must be removed because they do not clearly belong to one of these three factors. The remaining three scales contain four, three, and three items, respectively.

2.17 Visualization : reading a loading plot

In a factor analysis with **two orthogonal** factors, factor loadings are often displayed in a plot. In that case, each factor is shown as an axis, and each variable as a point. The factor loadings of the variable act as the coordinates of the point. If the factor model is correct, the correlations between the variables can be inferred from the plot: Indicate the line segment from the origin to one variable as *a*, and the line segment from the origin to the other variable as *b*, then the correlation between the variables is

$$r_{ab} = length(a) * length(b) * \cos(\text{angle } a, b)$$

Explanation

The above formula is only useful if you remember that $\cos 0° = 1$, $\cos 90° = 0$, and $\cos 180° = -1$. The formula can also be written differently. If one variable has loadings (a_1, a_2), and the other variable has loadings (b_1, b_2), then the correlation is equal to the so-called inner product of the loadings:

$$r_{ab} = a_1 b_1 + a_2 b_2$$

The last formula is convenient for computing the correlation, but the first formula is more helpful if you want to estimate the correlation from visual clues in the plot. If you apply the last formula to compute the reproduced correlation of a variable with itself, you get $r_{aa} = a_1^2 + a_2^2$, and that is the communality of the variable. In the plot this is the square of the length of the line segment *a*.

Example

Earlier the plot of Figure 2.3 was obtained with PCA of the Diesfeldt data.

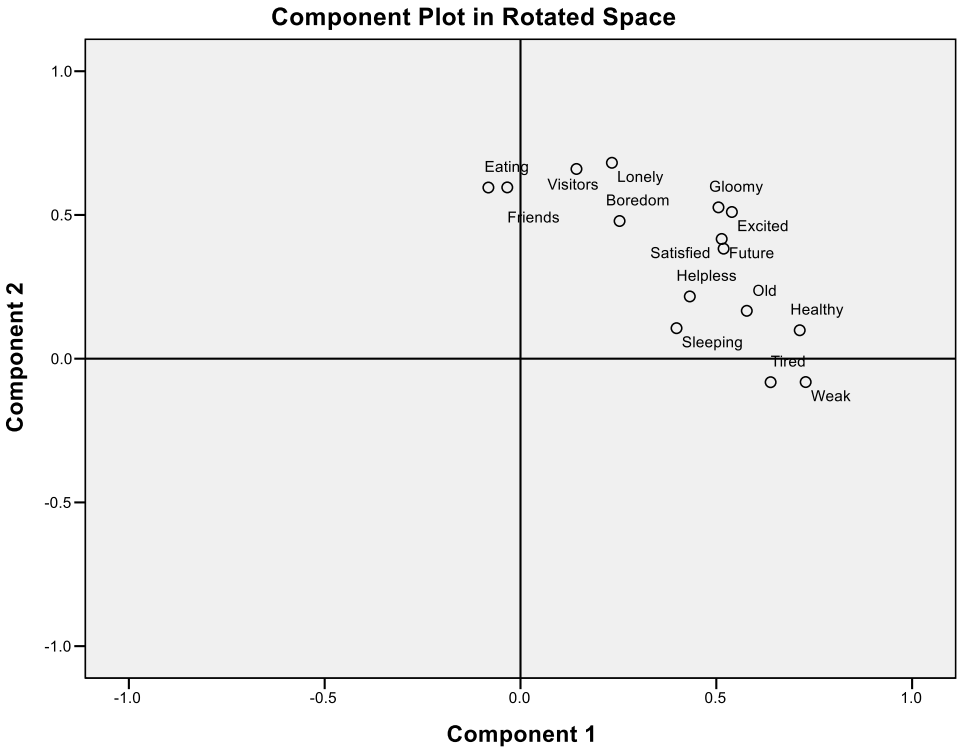

Figure 2.3

Table 2.23 shows some examples in which a conclusion is drawn about a correlation on the basis of the lengths and the angles in the plot. This is compared with the actual correlation in the right-hand column. In these examples, the conclusion from the plot always agrees with the observed correlation. In other cases, there can sometimes be discrepancies. But for a rough idea of the correlations, the plot can still be helpful.

Table 2.23

Conclusion of plot	*Observed correlation*
Tired and Weak (bottom right) correlate positively and strongly with each other, because they are close together, far from the origin.	r = .36
Gloomy and Excited (top right) correlate positively and strongly with each other, because they are close to each other, away from the origin.	r = .52

Eating and Friends (top left) correlate positively and strongly with each other, because they are close together, far from the origin.	r = .25
Weak and Lonely have a small positive correlation, because the angle (Weak, origin, Lonely) is large but still clearly smaller than 90 degrees, and they are far away from the origin.	r = .14
Eating (left) and Healthy (right) have a correlation of about 0 because relative to the origin, they are perpendicular to each other.	r = .04
Eating (top left), and Weak (bottom right) even have a slightly negative correlation with each other, because their angle is slightly larger than 90 degrees and they are far from the origin.	r = -.12
Helpless and Sleeping have a positive correlation, because their angle is small, but this correlation is close to zero because the two points are relatively close to the origin.	r = .03
Sleeping has a smaller correlation with Future than Old has with Future. For the first angle (Sleeping, origin, Future) is approximately equal to the second angle (Old, origin, Future), but Sleeping is closer to the origin than Old.	r = .17 and r = .37
The correlation between Gloomy and Excited is larger than the correlation between Satisfied and Future. The angle is in both cases approximately the same, but Gloomy and Excited are further away from the origin.	r = .52 and r = .42

For completeness' sake we now consider an example in which the correlation is calculated by the inner product formula. Suppose we want to know the correlation between Old and Healthy. The figure shows the following coordinates: Old ≈ (0.6, 0.2) and Healthy ≈ (0.7, 0.1). The reproduced correlation is 0.6 * 0.7 + 0.2 * 0.1 = 0.44. If you use the precise values shown in the factor pattern, then the calculation is 0.578 * 0.713 + 0.166 * 0.099 = 0.43. In reality, this correlation was 0.34. The fact that this is different from the prediction, means that this output of PCA does not provide a perfect description of the correlations. The plot only summarises the correlations, but not all details are accurate.

2.18 Appendix to Chapter 2

This appendix contains no content belonging to the learning objectives for this book. Nevertheless, the information can come in handy if you want to use more fit indexes in

a different context, or when you compare the results of SPSS and LISREL for the same model.

2.18.1 The Tucker-Lewis-index and the comparative fit-index

The Tucker-Lewis-index (TLI, Tucker & Lewis, 1973), also called the non-normed fit index (NNFI; Bentler & Bonett, 1980), and the comparative fit index (CFI; Bentler, 1990), together with the RMSEA are currently the most popular fit indexes (Jackson, Gillaspy & Purc-Stephenson, 2009). Both are called incremental fit indexes. Here, the model is compared to a **null model.** The null model is usually defined as the model in which all correlations are 0; that model explains no correlation at all.

The TLI is defined as

$$\text{TLI} = \frac{\chi^2_{\text{null}} / df_{\text{null}} - \chi^2_{\text{model}} / df_{\text{model}}}{\chi^2_{\text{null}} / df_{\text{null}} - 1}$$

Here, TLI > 0.95 can be regarded as a good fit (Hu & Bentler, 1999), 0.95 ≥ TLI ≥ 0.90 as acceptable fit, and TLI < 0.90 as bad fit. The TLI can be interpreted as the degree to which the model explains the correlations.

The CFI is defined as

$$\text{CFI} = \frac{(\chi^2_{\text{null}} - df_{\text{null}}) - (\chi^2_{\text{model}} - df_{\text{model}})}{\chi^2_{\text{null}} - df_{\text{null}}}$$

For the CFI the same boundaries are used as for the TLI.

If you want to compute the TLI and CFI, you may want to use a SEM program, because it will provide you with these indexes and many other fit indexes. But if you want, you can do it with SPSS. First, determine the chi-square of the null model. This can be done in the dialogue box of factor analysis by asking under the Descriptives button for 'KMO and Bartlett's test of sphericity', or in the syntax after the PRINT command – add the keyword SME. The chi-square and *df* which you get with Bartlett's test of sphericity, constitute a test for the null model, and they are identical to the chi-square and *df* which you would get from a ML factor analysis with zero factors (this can be derived from the formulas in the Algorithms section of SPSS Help given for Bartlett's chi-square and the ML chi-square). Then you can enter it in the formula for the TLI and CFI.

Explanation

The 'best' model is the model that is exactly correct in the population. For this model, the expected value of χ^2 equals df. The TLI makes use of the ratio χ^2 / df, while

the CFI makes use of the difference : $\chi^2 - df$. Applied to the best model we have $\chi^2 / df = 1$ and $\chi^2 - df = 0$. Both indexes have the form

$$\frac{\text{null model - my model}}{\text{null model - best model}}$$

Thus, your model is placed on a continuum between the null model and the best model (this explanation is based on the David Kenny webpage). Figure 2.4 illustrates this.

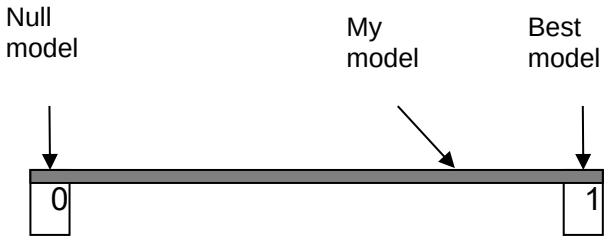

Figure 2.4

The above definition of the TLI is the way it is currently written. According to the original definition of Tucker and Lewis one would, if the chi-squares of SPSS are being used, first have to multiply χ^2_{null} with
$(N-1-(2k+5)/6-2m/3)/(N-1-(2k+5)/6)$.

Example
For the one-factor model with Diesfeldt's data we found earlier $\chi^2 = 230.364$ and $df = 90$. Bartlett's test of sphericity yields $\chi^2 = 765.483$ and $df = 105$. Then,

TLI = (765.483 / 105 – 230.364 / 90) / (765.483 / 105 – 1) = 0.752.

CFI = ((765.483 – 105) – (230.364 – 90)) / (765.483 – 105) = 0.787.

According to these indices the fit of the one-factor model is bad, too. With the original definition of Tucker and Lewis one gets here TLI = 0.751.

2.18.2 The relationship between the results of SPSS and LISREL

Sometimes, the same factor model can be tested with both SPSS and LISREL. That is true for the one-factor model. Testing a factor model with the ML method in SPSS produces essentially the same results as in LISREL. In case anyone wants to calculate

this, I will put it more precisely now. The algorithm used by SPSS is a variant of the method of Jöreskog (1967). In SPSS 19 and LISREL 8 (Jöreskog & Sörbom, 1996, 2006) the same loss function *F* is minimized. With a one-factor model the resulting loadings are therefore the same in both programs, provided that they both converge to the minimum. For a model with multiple factors the factor patterns are equal up to rotation. The value of chi-square is not quite the same, however. Both programs use a formula of the form $\chi^2 = n'F$, and both have the same value for *F*, but LISREL uses $n' = N - 1$ while SPSS uses $n' = N - 1 - (2k + 5)/6 - 2m/3$. Here, k is the number of manifest variables and *m* the number of factors. When calculating the fit indices, LISREL seems to use not the chi-square (the Minimum Function Fit Chi-Square), but the Normal Theory Weighted Least Squares Chi-Square. The original definition of the TLI is based on the Minimum Fit Function Chi Square (Tucker & Lewis, 1973).

Example
In Table 2.24, the results of SPSS are compared with those of LISREL for the one-factor model in the Diesfeldt data. The chi-square in the column *LISREL minimum fit* relate to the chi-square *SPSS Factor ML in* the manner explained above:

$$\chi^2_{\text{Lisrel minimum fit}} = \chi^2_{\text{SPSS}} * (N-1)/(N-1-(2k+5)/6-2m/3) .$$

Table 2.24

	SPSS Factor ML	*LISREL minimum fit*	*LISREL normal theory weighted least squares*
Chi-square one-factor (m = 1)	230.364	238.27	247.48
df	90	90	90
Chi-square null model (m = 0)	765.483	788.96	1493.66
Df null model	105	105	105
RMSEA	0.089 *	0.092 *	0.094
TLI	0.752 *	0.747 *	0.88
CFI	0.787 *	0.783 *	0.89

* Not in output, but calculated from the above results.

2.18. 3 The use of the term confirmatory factor analysis

In the literature, the term confirmatory factor analysis is usually used for analyses with a SEM program. In this book the choice was made to call some analyses with SPSS confirmatory too. In this section, I defend that choice (see Stewart et al., 2001).

A situation often discussed in this book, is that one wants to examine multiple scales on unidimensionality. The items are the manifest variables, and the items within

the same scale are believed to have one underlying factor. The hypothesis specifies both the number of factors and the pattern of zeros in the factor loadings. It would be natural to test that with SEM programs like LISREL. But it is unrealistic to expect this from a student of this book. Therefore we try to approximate this analysis with SPSS. The number of factors can be tested with the ML chi-square and the fit indices derived from it. Additionally we compare the promax-rotated factor pattern with the theoretical classification of items in scales. Only the part of the hypothesis that specifies the number of factors is tested in the sense that there a *p*-value is obtained. Nevertheless, the nature of the hypothesis is confirmatory, and also part of the analysis is confirmatory in the sense that a hypothesis is tested (namely, the hypothesis about the number of factors). Another part of the analysis (namely, the rotation) uses an exploratory technique, but the aim is still confirmatory. In my opinion it the analysis is altohether closer to confirmatory than exploratory factor analysis.

If the hypothesis is that there is one factor, the outcomes of SPSS and LISREL are essentially the same, at least if you used the ML used method in both programs (see Section 2.18.2). If the model has several factors then one can, in principle, specify for each factor model in SPSS a similar model in LISREL, with the same number of factors and essentially the same results for the test (chi-square, *df* and *p*-value; paragraph 2.18.2 describes what is meant by "essentially the same" here). It would be confusing to call the same analysis confirmatory if it is conducted with LISREL but exploratory if the same is done with SPSS.

3 Comparing multiple factor analyses

3.1 Background

In the previous chapter we concentrated mainly on simple situations where one factor analysis already leads to a conclusion. We have seen that in a factor analysis there is sometimes doubt about the number of factors. In these more difficult situations, it is advisable to base the conclusion on multiple factor analyses.

In addition, computational problems can sometimes occur in a factor analysis, so that the program can not complete the analysis. We will briefly discuss what the causes and solutions are.

3.2 Learning goals

After studying this chapter, you can compare the results of multiple factor analyses with each other in order to arrive at a decision about the number of factors. You can also suggest solutions for the most important computational problems of a factor analysis.

3.3 When to compare factor analyses?

Comparing various factor analyses is especially important if the number of factors is unclear. We can distinguish these situations:

a. A questionable outcome in the statistical evaluation of the model. We saw in the previous chapter that it sometimes happens that $p < 0.05$ while RMSEA is between 0.05 and 0.08, and it is doubtful in this case whether there are enough factors. In this case it is good to see if there should not be an extra factor.
b. Confirmatory factor analysis if there is a competing theory. For example, suppose the theory of Cattell says that there are four factors, while the theory of McGrae and Costa (1987) says that there are five factors in the same data. These are two competing theories. If we do a confirmatory factor analysis for the theory of Cattell, then it is conceivable that the fit is good. However, that is not sufficient to dismiss the theory of McGrae and Costa. After all, the fit of the last theory may be even better. Therefore, a confirmatory factor analysis must be done for both theories to compare the advantages and disadvantages of both theories. Jackson, Gillaspy and Purc-Stephenson (2009) recommend always to compare plausible alternative models with confirmatory factor analysis.
c. In an exploratory factor analysis, the number of factors is unknown to begin with. Often, the number of factors is determined initially on the basis of the

eigenvalues, but this is not a very defensible criterion. It may therefore happen that an analysis with a different number of factors yields a better description of the data.

If there is any doubt about the number of factors, it is wise to do multiple factor analyses, with a different number of factors. The decision on the number of factors is then taken by comparing the results of the analyses, and choosing the analysis that offers the best description of the data. The question is then how we can decide which analysis is the 'best'. We discuss this in the following sections.

This discussion is, by nature, theoretical and perhaps even philosophical. In short, it is difficult and vague. Do not ask me to give a simple rule of thumb. If there was a simple rule, I would have written it down in the previous chapter. But there isn't. And that is also logical, because factor analysis is used mainly in theory formation. Why would that be simple?

3.4 The problem

Perhaps you think: let's just do a lot of factor analyses and take the one with the best fit. You could do that, but it is not fruitful. The problem is caused by a conflict of two laws. The first law is *mathematical* in nature:

The more factors, the better the fit.

Let us first consider why that is, and then study what the consequences are.

In a factor analysis with k variables, there are $k * (k - 1) / 2$ correlations. In a model with f factors, the correlations are explained by $f * k$ factor loadings. The factor loadings are estimated by a system of equations to solve it as good as possible, in which the correlations act as known, and the factor loadings as unknowns quantities. If the number of factors f is large enough, there are more unknown than known quantities, and then a perfect solution always possible.

The second mechanism is more *philosophical* in nature:

The fewer factors, the more informative the model.

Above we stated that one can always achieve a perfect fit by assuming enough factors. That solution, however, will be meaningless precisely because it is always possible. It does not impose any restriction on the data. It is therefore no longer a testable theory, and therefore not informative.

For example, if there are 10 manifest variables, there are 45 correlations (if you do not count the diagonal). If you assume five factors, there are 50 factor loadings. In other words, you try to explain 45 correlations with 50 other correlations. As if you summarise a 45-page book in 50 pages. That does not help. That is no longer a theory, it just converts the data into a figment of imagination. You could compare it with the

'theory' of a homunculus: it doesn't explain anything, it only shifts the problem to another level.

Even in less extreme cases, you can say that a theory with fewer factors is much stronger. Suppose that in the example with the 45 correlations between 10 manifest variables, we explain the correlations with four factors, so 40 factor loadings. Then we have essentially 45 'observations' explained by 40 assumptions. That is testable, but it is still not very impressive. You could say that we have explained not 45 correlations, but only 5. A one-factor model, on the other hand, would need only 10 'assumptions' about the loadings to explain the correlations. The fewer factors, the greater the difference. In other words, the **explanatory power** of the model is increased when there are fewer factors. The model becomes more **parsimonious**. Therefore, it is conceivable that a model with fewer factors is still preferred, even though its fit is not as good as the bigger model.

3.5 The basic principles

When comparing multiple factor analyses to decide on the number of factors that you are going to use, a balance has to be found between the following criteria:

- Good fit
- Parsimony (high explanatory power)
- Interpretability

The first two principles are discussed in the previous section. It also became clear that the principle of parsimony is mainly philosophical in nature. It depends on what you expect from a good theory. But the answer is not only parsimony. Interpretability is important too. If a less parsimonious model leads to factors that are better understood in terms of content, then it still may be preferred.

3.6 Elaboration of the basic principles

Previously we saw that goodness-of-fit can be expressed by the RMSEA. Let us now first look at the parsimony of the theory.

3.6.1 Parsimony

In section 3.4 has been argued that the explanatory power of a factor model can be expressed by comparing two things together:

- the number of correlations that are explained, and
- the number of parameters (such as factor loadings) that are used in the explanation.

About the difference between them, I pointed out that this is actually the number of correlations that really is explained. For example, if there are 55 correlations explained with 50 other correlations (such as factor loadings), then there are only 5 correlations explained really. This difference is the **number of degrees of freedom.** That is the

number that we see in the SPSS output in the table *Goodness-of-fit test* under ***df*** when ML extraction is being used:

$$df = \begin{pmatrix}\text{number of correlations}\\ \text{that is being explained}\end{pmatrix} - \begin{pmatrix}\text{number of parameters}\\ \text{that is being used}\end{pmatrix}$$

(Incidentally, the number of parameters is somewhat more complicated than just the number of factor loadings.)

In previous analyses such as *t*-tests, we described the number of degrees of freedom as a number indicating at which row of the table one has to look. Now, in factor analysis, the number of degrees of freedom has more meaning. The number of degrees of freedom is a measure of the **explanatary power** of the model. The greater the number of degrees of freedom, the stronger the model is.

If you enter too many factors in a confirmatory factor analysis, the number of degrees of freedom will be **negative.** A negative number of degrees of freedom means that there are so many parameters that the theory is no longer falsifiable.

3.6.2 Goodness-of-fit

Earlier it was stated that the RMSEA is a measure of badness-of-fit. If you study the formula, you can see that it also includes the degrees of freedom, and the RMSEA decreases with *df*. That is why the earlier presentation is incomplete. The RMSEA not only takes account of the badness-of-fit (χ^2), but also the parsinomy of the theory. The fit (measured with χ^2) can only get better as the number of factors increases, but the RMSEA could decrease as the improvement in fit is not proportional to the decrease in parsinomy.

Nevertheless, you still have to weigh the fit and the parsinomy separately. The RMSEA only takes into account the statistical implications of the *df*. In addition, an parsimonious theory has advantages that are difficult to quantify. For example, the theory of Spearman that intelligence is one factor, is considerably simpler than the theory of Cattell, which assumes a large number of factors. If both theories had the same RMSEA, Spearman's theory would still be preferable.

The extent to which one theory is better than another can also be tested, provided that one theory is a subset of the other: that is to say, one theory (the strong theory) includes all assumptions of the other theory (the weak theory) plus additional restrictions. In that case, you may deduct the χ^2-value and *df* of one theory from those of the other theory:

$$\chi^2(\text{difference}) = \chi^2(\text{strong theory}) - \chi^2(\text{weak theory})$$

$$df(\text{difference}) = df(\text{strong theory}) - df(\text{weak theory})$$

Subsequently χ^2(difference) is compared with a χ^2-distribution with the number of degrees of freedom given by *df* (difference). On this basis, the *p*-value can be calculated.

In this context, an important model is the so-called **null model.** This model implies that all correlations are zero, which could be described as a model with zero factors. This forms the basis for the so-called Comparative Fit Index (CFI) (Bentler, 1990). With $d = \chi^2$ - *df* the formula for this index is: CFI = 1 - *d* (proposed model) / *d* (null model). Hu and Bentler (1999) argue that CFI ≥ 0.95 indicates a good fit. It should be noted here that the null model is generally not a plausible model. Indeed, Kenny (2015) describes it as "the worst possible model". You may wonder whether it is so convincing that a proposed model is better than a nonsense model.

The procedure for testing the difference in chi-squares can therefore be criticised. It really only makes sense if the weak theory is acceptable. Otherwise, it may cause you to accept a nonsense model because a still-larger-nonsensical model exists. Millsap (2007b, p. 878) writes: "The practice of ignoring the global chi-square tests while at the same time conducting and interpreting chi-square difference tests between nested models should be prohibited as nonsensical."

3.6.3 Interpretability

Because factor analysis is used in theory formation, it is important that the results are also fertile for the formulation of theories. Therefore, as the number of factors is unclear, the factor patterns of different analyses are compared with each other on interpretability. A factor pattern for which a clear substantive theory can be conceived is preferable to a factor pattern that is incomprehensible. What matters is that we understand the data as good as possible, and that cannot be captured with just statistics such as RMSEA and *df.*

The extent to which a factor pattern can be interpreted also depends on the context of the research, namely on the theories that exist at that moment. It is conceivable that a factor pattern that is not well understood at first will become understood after the development of a new theory. For example, in the first factor analysis of Spearman, there were signs that intelligence had more than one common factor, in the sense that some tests correlated higher with each other than was predicted by the theory of a single common factor *(g).* Initially, these deviations were explained away by stating that such tests were too similar. Only when Thurstone formulated a theory with several factors (the primary abilities), the deviations were understood and were therefore taken more seriously.

That deviations from a theory only get consequences if there is an alternative theory, is a pattern that one can also see in other sciences. The famous example is of course the constancy of the speed of light, which is a deviation from the classical mechanics of Newton and Galilei. When this deviation was established empirically, it did not immediately lead to the rejection of classical mechanics. That only happened

when Einstein formulated his alternative theory, in which this fact was no longer a deviation, but a fundamental law of nature, that moreover could be derived from other laws of nature (the Maxwell equations of electromagnetism).

The extent to which a factor pattern can be interpreted also depends on a another context of the research, namely the practical application. For example: I am regularly involved in investigations in nursing homes. Usually it concerns the supply of management information. Managers often want very global information. One reason is that they have many different types of information about customers, employees, finances, subgroups, function groups, functions, products and so on. For an overview, one does not want too much information per part. That may be a reason to opt for a small number of factors. On the other hand, this detailed information may be important for a therapist, which may be a reason for a larger number of factors. So, with the same items you might come to the conclusion that there is one scale *'Mood Problems'* if the information is intended for managers, while there are two scales '*Depressivity'* and *'Fear'* if the information is intended for practitioners.

Does that mean that factor analysis can bring you to any conclusion you want? No, but there is some flexibility. Not because we ingnore the reality, but because different degrees of detail are possible in sketches of that reality. In the case of mood problems, for example, it is true that depression and anxiety are not exactly the same, but that they often go together. So one can wonder whether they are really so different. Maybe they are in part the same. A scale purist would say that adding fear and depression is comparable with the adding apples and pears. A scale pragmatist would not deny that, but say that this is not a problem if one just wants to know how much fruit there is.

3.7 Examples

3.7.1 Comparison of multiple analyses in Diesfeldt's research

See the Diesfeldt study, which was discussed in the previous chapter. Table 3.1 summarises the results for the *p*-value and the RMSEA.

Table 3.1

Number of factors	*Chi-square*	*df*	*Sig.*	*N*	*RMSEA*
1	230.364	90	0.000	197	0.089
2	151.346	76	0.000	197	0.071
3	91.694	63	0.011	197	0.048
4	57.626	51	0.244	197	0.026

In view of the *p*-values, a conclusion could be that there are four factors, which is in accordance with what Diesfeldt concluded on the basis of the minimum eigenvalue criterion. In view of the RMSEA values, the conclusion is that good fit is achieved with

three factors, and acceptable fit with two factors. Thus one needs two to four factors. The further choice between them should depend on the interpretability of the factor patterns, the parsimony of the resulting theory, and the fit.

If necessary, we can first test whether the difference between these models is significant. The results are shown in table 3.2.

Table 3.2

Comparison	*Chi-square*	*df*	*Sig.*	*N*	*RMSEA*
2 versus 3 factors	59.652	13	0.000	197	0.135
3 versus 4 factors	34.068	12	0.001	197	0.097

The differences between these models are thus significant and RMSEA of the improvement is in both cases substantial.

To evaluate the interpretability we have to study the factor patterns of the analyses with good or acceptable fit. Tables 3.3 to 3.7 display the factor patterns with sorted loadings. Loadings less then 0.30 have been omitted. The correlation matrix of the factors is also shown for the last two analyses.

Table 3.3

Pattern Matrix[a]

	Factor	
	1	2
Lonely	.642	
Eating (-)	.578	
Excited (-)	.542	
Visitors (-)	.496	
Boredom	.493	
Gloomy	.465	.319
Future (-)	.453	
Friends (-)	.438	
Satisfied (-)	.433	
Weak	-.306	.910
Healthy (-)		.689
Tired		.510
Old		.371
Helpless		

Sleeping (-)		

Extraction Method: Maximum Likelihood.

Rotation Method: Promax with Kaiser Normalization.

a. Rotation converged in 3 iterations.

Table 3.4

Pattern Matrix[a]

	Factor		
	1	2	3
Future (-)	.706		
Excited (-)	.626		
Satisfied (-)	.622		
Old	.509		
Sleeping (-)	.387		
Boredom	.380		
Gloomy	.365	.326	
Tired	.350		
Helpless	.339		
Visitors (-)		.944	
Lonely		.571	
Friends (-)		.415	
Weak			.948
Healthy (-)	.313		.425
Eating (-)			

Extraction Method: Maximum Likelihood.

Rotation Method: Promax with Kaiser Normalization.

a. Rotation converged in 6 iterations.

Table 3.5

Factor Correlation Matrix

Factor	1	2	3
1	1.000	.564	.390
2	.564	1.000	.266
3	.390	.266	1.000

Extraction Method: Maximum Likelihood.

Rotation Method: Promax with Kaiser Normalization.

Table 3.6

Pattern Matrix[a]

	Factor			
	1	2	3	4
Future (-)	.746			
Excited (-)	.722			
Satisfied (-)	.581			
Gloomy	.499			
Old	.444			
Helpless	.432			
Boredom	.352			
Visitors (-)		.963		
Lonely		.505		
Friends (-)		.367		
Weak			.932	
Healthy (-)			.446	
Tired			.311	
Eating (-)				
Sleeping (-)				1.012

Extraction Method: Maximum Likelihood.

Rotation Method: Promax with Kaiser Normalization.

a. Rotation converged in 7 iterations.

Table 3.7

Factor Correlation Matrix

Factor	1	2	3	4
1	1.000	.530	.404	.350
2	.530	1.000	.233	.118
3	.404	.233	1.000	.130
4	.350	.118	.130	1.000

Extraction Method: Maximum Likelihood.

Rotation Method: Promax with Kaiser Normalization.

Note now, first, that the last pattern factor is the only one showing simple structure, in the sense that each item is loading on at most one factor. In the two other patterns there are items that load on two factors. The simple structure can be a reason for choosing the solution with four factors.

It is also important to consider the content of the factors. In the following discussion, for convenience, each factor will be indicated with the item that loads the highest on it. This is to prevent me from inventing names for factors that will never be used again later (normally it is undesirable to identify factors with the name of an item). The factors for the three analyses are then shown in table 3.8.

Table 3.8

Number of factors	*Factor names*			
2	Lonely	Weak		
3	Future	Visitors	Weak	
4	Future	Visitors	Weak	Sleeping

The factors of the three analyses have a relationship with each other:

- First, you see that some factors occur in multiple factor patterns. For example, the factor *Weak.*
- Secondly, you sometimes see that a group of items that first load on one factor disintegrates into two factors in the next analysis. For example, the factor *Lonely*, which disintegrates into *Future* and *Visitors.*
- Third, you can see that, if more factors are assumed, some of the later factors have only one or two items.

I do not claim that these relationships are logically necessary (at least not that I know of), but they do occur often.

The consequences of this for the interpretation are as follows.

- As a factor recurs in multiple analyses, this reinforces confidence in the existence of that factor. After all, this factor is less the result of a fairly random choice in the analysis.
- A factor that disintegrates into several factors, can indicate two factors that correlate relatively high with each other. If two factors correlate extremely high (for example > 0.90), it is probably better to view them as one factor. As the correlation is lower, there is more reason to differentiate the factors.
- A factor that has just one or two items loading on it, is difficult to interpret, and this may be a sign that there are too many factors being assumed, and / or that the items should be used separately (i.e., not in a scale with other items).

Putting everything together, my reasoning in this example would be as follows:

- That we have a factor *Weak* is pretty clear, since it emerges in in all three analyses.
- The factors *Future* and *Visitors* emerge in two of the three analyses. This suggests that they are indeed two different factors.
- The fact that the two-factor combines these two factors into one factor can be explained by the relatively high correlation between the two factors (0.564).
- This correlation is not so high that the factors must be considered almost identical.
- The items of the factor *Visitors* have a clear characteristic feature.
- For the items of the *Future* factor this is less clear.
- All in all I would assume the existence of two different factors *Visit* and *Future.*

That means that the two factor solution falls and the only choice is between *three* factors or *four* factors. These two factor patterns are much the same in that they both factors *Future, Visitors* and *Weak.* The only question is whether Sleeping should be distinguished as separate factor (as in the four-factor pattern) or that it should be counted into *Future* (as in the three-factor pattern). My reasoning would be:

- The requirement of simple structure indicates the model with four factors.
- The relatively small correlation between the factors *Sleeping* and *Future* (0.350) points to the belief that they are different factors.
- Sleeping can also be seen as subtantively different from the other items. Together with Eating this is the only item which is about an activity (or at least a verb). In addition, sleep is often influenced by medicines, which would explain the existence of an extra factor.
- I prefer a clean scale, with items that I'm pretty sure of, over a longer scale with items that I don’t trust.

All in all I would opt for a four-factor solution, and accept that the factor *Sleeping* is difficult to interpret because there is only one item loading on it. For the scale construction, this means that the item Sleeping is not included in the *Future* scale, but

maybe it can be reported as a standalone item (which incidentally has the disadvantage that its internal consistency reliability cannot be determined).

Reflecting on my conclusions, you should notice that my conclusions are eventually the same as those of Diesfeldt. This, despite the fact that I have based them on other analyses. And my reasoning is very long, while that of Diesfeldt is very short. But the above reasoning cannot be written in an article as it is way too long. Such a long story would never be accepted. Certainly not if the scale construction is only a side issue in the article. An option is to use a simple criterion with the same conclusion. In this case one could write for example that the number of factors is determined on the basis of the significance test, and the model with four factors had no significant violation. Or, if the public is less sophisticated, you could write that the number of factors is based on the minimum eigenvalue criterion. In both cases you get to the same conclusion without all the talking. But it is that an honest description of the decision process? If the subjective criterion of interpretability played a role in the decision, that has to be reported, otherwise this constitutes a violation of academic integrity. What you often see is that it is reported that the number of factors is partly based on the interpretability, without explaining exactly what has been done. That is not entirely satisfactory, because it is not reproducible.

3.7.2 The Big Five, Six, Seven, Eight, Nine, Ten

De Raad and Barelds (2008) used 2365 adjectives on which 1466 subjects had to judge themselves and others. They write:

> With respect to the number or factors to extract, we were generally guided by two types or considerations. One was psychometrical, mainly determined by the eigenvalues, the scree test, and by studying the stability of the factors across subsets or the data set; the other was determined by both the interpretability or factors and expectations or factors that could appear and by factor structures advocated in the research literature. (De Raad & Barelds, 2008, p.354)

Thus, you see that the theoretical interpretation plays an important role here. If factor analysis is applied in such a way, it is not some kind of blind machine where you enter data on side and that spits out answers at the other side. It is rather a means to support theory formation.

In the end, De Raad and Barelds conclude that there are eight factors. They conclude this by comparing 10 factor analyses. The factors that resulted from these 10 analyses are shown in Figure 3.1. You can see in this figure that, regardless of how many factors are being assumed, *some* interpretation of the factors is possible. Why then eight factors, and not just the Big Five? The figure shows that the analysis with five factors does not result in the factors of the Big Five. That outcome would therefore agree only with the *number* of factors in the Big Five theory, but not with the *content*

of these factors. To get the five factors of the Big Five, eight factors have to be drawn. And then one gets also three factors which do not belong to the Big Five.

Why should the solution agree with the Big Five theory? It is not an irrefutible law that a factor analysis must fit in with previous theories. But in this case the already existing theory is pretty strong, according to the literature. You can't just push this aside and ignore it. In particular, as the literature – on the basis of similar analyses – distinguishes between Agreeableness and Conventionality, you cannot just say that this distinction is unnecessary. That would be possible if the data leave no other choice. But in this case the data allow a choice that agrees with the literature, namely by assuming eight factors.

If you look at this example, you see that the outcome of the analysis is partly motivated by the already existing theory of the Big Five. But it's not like that theory is blindly believed. We also look at the data, and the combination of theory and data analysis eventually leads to a theory that deviates from the earlier theory. The outcome of the analysis *incorporates* the existing theory, but it is not necessarily a confirmation of that theory.

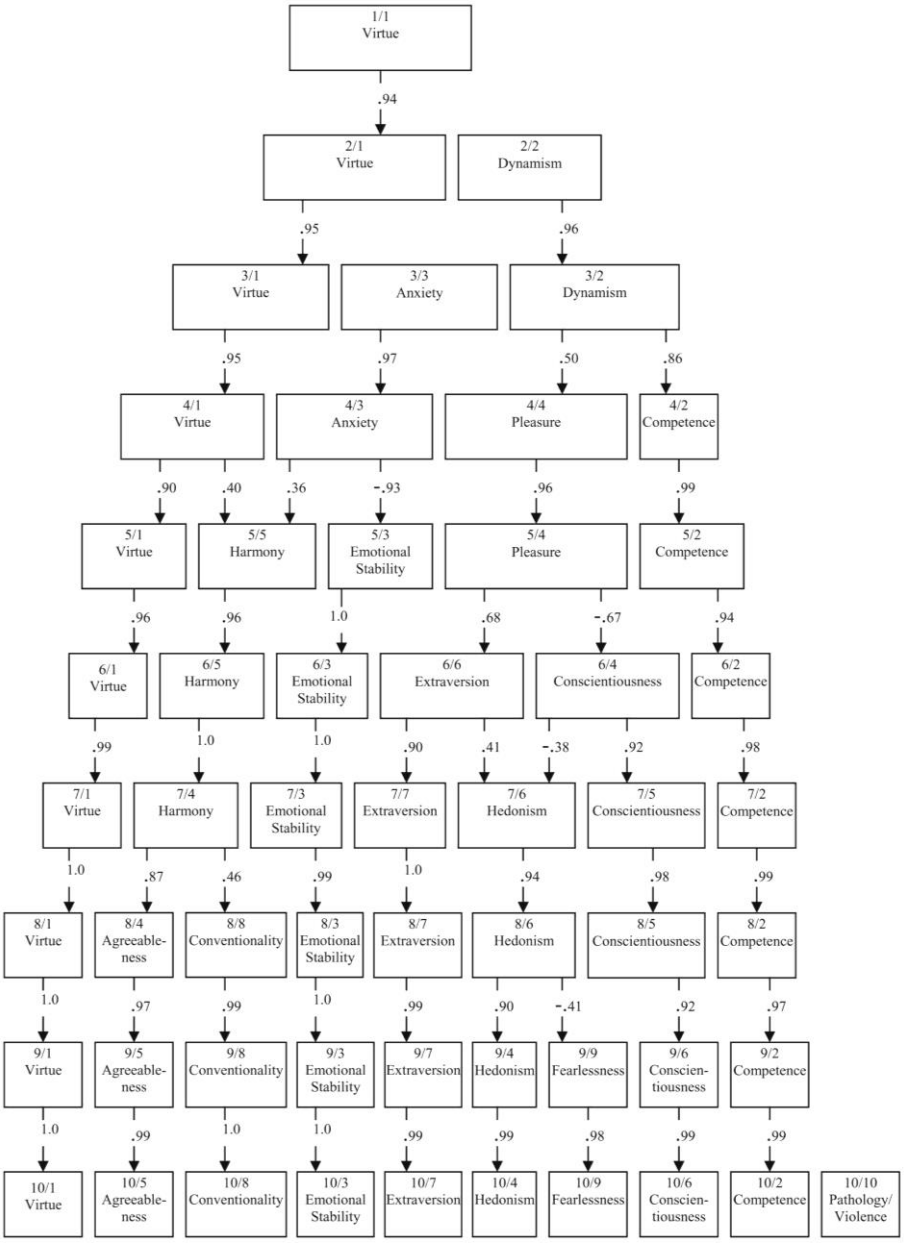

Figure 3.1 (Copyright © 2008 by the American Psychological Association. Reproduced with permission. The official citation that should be used in referencing this material is De Raad, B., & Barelds, D. P. H., 2008, 'A new taxonomy of Dutch personality traits based on a comprehensive and unrestricted list of descriptors', Journal of Personality and Social Psychology, 94, 347-364. The use of APA information does not imply endorsement by APA)

3.8 Computational problems with factor analysis

The calculations by a factor analysis are not as straightforward as in an ANOVA. The estimates are iteratively adjusted. As a result, various problems can arise.

3.8.1 Some variances 0

If there is a manifest variable with variance equal to 0, then factor analysis is impossible. For such a variable, factor analysis would be completely useless. Variables with a positive but small variance can also cause a problem. In SPSS this problem is flagged with this warning:

```
There are fewer than two cases, at least one of the
variables has zero variance, there is only one variable
in the analysis, or correlation coefficients could not
be computed for all pairs of variables. No further
statistics will be computed.
```

After this, as the warning promises, no more output follows.

To solve this problem, you must first carefully read the content of the warning. Each of the mentioned possibilities can be the cause. The first possibility is that there are fewer than two subjects. Here you must remember that subjects with missing observations are being removed by default (this is the option *exclude cases listwise* under the Options button of the factor analysis dialog box). So if all men miss question 1, and all women miss question 2, then all subjects are removed from the analysis. And then it is logical that no output appears.

The second possibility is that there is a variable with variance 0. That means that everyone has the same score. The correlations with that variable are then undefined, so that no factor analysis is possible. Whether one of these problems occurs can be verified by using the button Descriptives in the factor analysis, and asking for the univariate statistics. And then you have to solve the problem by removing the variable that causes the problem. (The options *exclude cases pairwise* and *replace with mean* should be discouraged, because they usually lead to other problems).

3.8.2 No convergence

This means that no final solution has been found. The outcomes shown are meaningless and can not be used. In SPSS, this problem is flagged by the following footnote:

```
a. Attempted to extract 3 factors. More than 25
iterations required. (Convergence = .010). Extraction
was terminated.
```

In addition, the goodness-of-fit tests are not displayed in these cases. Note that this situation is misleading: output is being produced, but the output cannot be trusted. The previous issue (variances 0) was much more pronounced, because you know right away that there is a problem from the fact that no output appears. But here, you have to read the footnote and understand it.

Sometimes it is helpful to increase the number of allowed iterations, for example, from 25 to 250. If convergence is achieved with that, the ensuing solution may be used. If that does not help, then the lack of convergence indicates that the number of factors is incorrectly specified. In addition, this can happen if the number of subjects is very small (< 100), as a result of which the correlation matrix is too irregular.

The issue at stake here concerns only convergence in the *extraction* phase. Non-convergence can also occur during the *rotation* phase, but that is much less of a problem. The rotation is then not optimal, and if the factor pattern is not interpretable, this can be caused by that non-convergence of the rotation. But the unrotated factor pattern and the chi-square test can still be trusted. If the rotated factor pattern is interpretable, then there is no problem.

3.8.3 Communality greater than 1

Communality is equal to the square of a multiple correlation and therefore it cannot be greater than 1. But sometimes a factor analysis leads to estimates where the communality is greater than 1. That cannot be a good estimate. Such events are known as **Heywood cases.** Modern factor analysis programs ensure that the communality usually stays smaller than one. SPSS issues in such cases a footnote – below the table Communalities – with the following warning:

```
a. One or more communalitiy estimates greater than 1
were encountered during iterations. The resulting
solution should be interpreted with caution.
```

If the final solution has a communality close to 1 (e.g., 0.999), this indicates still a problem. In psychology, that outcome is not realistic because it would mean that the relevant manifest variable would have a reliability of at least 0.999. Furthermore, the intervention (the communality being set at 0.999) does not eliminate the fact that there is something strange about the relationship between model and data. The chi-square statistic no longer has a chi-squared distribution with the specified degrees of freedom, but its distribution is a mix of chi-square distributions with different degrees of freedom (a so-called chi-bar-square distribution, see Dykstra, 1991). Consequently, the *p*-value in the output is no longer correct. Therefore, these cases should still be called Heywood cases – even if the reported communality is not greater than 1. In summary, you have a Heywood case if the above warning is given and the final communality is

0.999 or 1. In a Heywood case, the *p*-value reported in SPSS cannot be trusted, and when it comes to psychology, the factor pattern can not be trusted either.

Sometimes it helps to specify a different number of factors. Both a too large and a too small number of factors can lead to a Heywood case. Also, a small *N* can lead to a Heywood case, and there are several other possible causes (Kolenikov & Bulbs, 2012).

3.8.4 Correlation Matrix is not positive definite

A correlation matrix will always have certain properties such as symmetry, 1's on the diagonal, and so on. These properties can be summarised under the heading 'positive semi-definite'. If the correlations are, moreover, such that there are no multiple correlations of 1 (i.e., no single variable is 100% predictable from other variables by multiple linear regression), then the correlation matrixis 'positive definite'. So

- each correlation matrix is positive semi-definite;
- any natural correlation matrix is positive definite.

The exact meaning of these terms is not important here. What is important is that factor analysis is possible only if the correlation matrix is positive definite. But that would naturally almost always happen. How can it still go wrong?

This may be the case if there are missing observations and the correlations are computed with the pairwise option. For example, suppose a person has answers to questions 1 and 2, but not to question 3. Then that person can in theory still be used to calculate the correlation between questions 1 and 2, but not to compute the correlation between questions 1 and 3. Consequently, the correlation between questions 1 and 2 is computed in a different group than the correlation between questions 1 and 3. In such cases it is possible that the correlation matrix is not positive semi-definite, and then no factor analysis is possible. To avoid this, use list wise deletion of missing values under the Options button. This means that once a subject has a missing observation, that subject is not used throughout the analysis. It is important that each participant answers the questionnaire completely if this option is being used. If someone omits one answer, all answers of this person are useless.

- If there are fewer subjects than variables, the correlation matrix is positive semi-definite, but not positive definite.
- If the data consist of a correlation matrix, and you analyse it as if it's the raw data matrix (so first calculate the correlations of the correlations), then you should not expect anything meaningful in the output.

3.8.5 Hessian is not positive definite

The Hessian is a matrix which is used in order to optimize estimates, similar to a second derivative. If it is not positive definite, the program is unable to optimize the estimates. This is what happens with the correlation matrix in Table 3.9.

Table 3.9

	v1	*v2*	*v3*	*v4*
v1	1	0	0	0
v2	0	1	0	0
v3	0	0	1	.10
v4	0	0	.10	1

This correlation matrix is about the simplest one that one can think of, but if you conduct a ML factor analysis with one factor, you get the warning that the Hessian is not positive definite. It won't help to take more factors, because if you do an analysis with two factors, then the number of degrees of freedom is negative and the Hessian is still not positive definite. Sometimes it helps to choose a different analysis, such as PCA instead of ML.

4 Conducting and reporting a reliability analysis

4.1 Background

In scale construction one will initially conduct a factor analysis the items, and some items may be deleted. Furthermore, the set of remaining items will be partitioned into subsets, called scales, each corresponding to a factor. After this you have to analyse each scale separately for the question whether its reliability is large enough.

In a reliability analysis you compute primarily an internal consistency reliability, and you judge whether it is high enough. An internal consistency reliability indicates the reliability of the *total score*. More specifically, it provides a lower bound for that reliability. This increases with the number of items and the correlations between items. In addition you have to study for each item whether it has a positive impact on the reliability. Items with a negative contribution to the internal consistency reliability are removed from the scale.

4.2 Learning goals

After studying this chapter, you can use a reliability analysis to

- examine which items should be removed from the scale,
- determine the internal consistency reliability of the resulting scale, and
- judge whether the reliability is large enough.

Furthermore you can report this in a concise report.

4.3 Basic report a reliability analysis

The usual set-up of a basic report is not convenient here because many steps would be rather trivial. Actually, the analysis just too simple ☺ . Nevertheless, this chapter follows the pattern of a basic report, to be consistent. But we include only the relevant parts:

- Design
- Analysis
- Estimates
- Decisions
- Interpretation

4.4 Running example

As an example we use the research of Diesfeldt (1997) to the Depression List, described in section 2.4.

4.5 Design

In the design you specifiy which items belong to the scale for which the internal consistency reliability has to be calculated. *Within a process of scale construction this scale is to be selected as a group of items that is unidimensional or unifactorial.* If a preceding factor analysis concluded that there were several factors, one should conduct a reliability analysis separately for each factor. In each of these reliability analyses only items belonging to the same factor must be used. Those are the items that load highly on this factor and not on any other factor. Similarly, if IRT analysis concluded that there are several dimensions, one should conduct the reliability analysis for subsets of items that are unidimensional.

If the factor analysis had a factor that was not interpretable or that had only a small number of items, then one might decide to exclude this factor and its items from the measuring instrument all together. For such factors a reliability analysis is not needed.

Explanation

In a reliability analysis you calculate the reliability of a sum score on multiple items. For each item sum score that you want to calculate later, you must do a reliability analysis. If you have multiple factors, then these consitute different scales, from which you will calculate different sum scores, and for which different reliability analyses are needed.

The remark that you do not have to do a reliability analysis for items that are not proposed as scale, means that we are talking about items for which you will never calculate a sum score. It does not mean that you only need to avoid the word ‘scale’.

You have to do reliability analysis on a unidimensional or unifactorial group of items, for two reasons. Firstly, because of construct validity, only scores of unidimensional scales (i.e., items with the same factor) should be used. Then reliability analysis should cover these sum scores. However, if the aim was solely criterion validity and content validity, that will invalidate this argument – and then you did not have to do a factor analysis to begin with. Second, it is assumed implicitly or explicitly in a reliability analysis that the items are unidimensional. Cronbach’s alpha is a good estimate for the reliability only if the items are essentially tau-equivalent, and this condition implies that the items satisfy a one-factor model. In a multidimensional set of items, reliability analysis can mean that you remove items one by one, until the result is about unidimensional. Actually this resembles a primitive kind of factor analysis, and then you better do it properly right away.

Example 1
In the Depression List data of Diesfeldt, one of the factor analyses was followed by the conclusion that this list might be unifactorial if the items Eating, Sleeping and Friends are removed. If we analyse this further, then the design is:

> Items *Depression List* : Please t / m Future, excluding Eating, Sleeping and Friends

Another option was to split the list into subscales. Then design is

> Items of *Spirited* : Future, Excited, ..., Old
> Items of *Health* : Strong, Healthy, Tired
> Items of *Social:* Visitors, Friends, Lonely, Eating

This leads to three different reliability analyses: a reliability analysis for *Spirits,* a reliability analysis for *Health,* and reliability analysis for *Social.*

In this chapter I will continue with a single analysis of the entire list. That's actually a bad choice, I just do this because this analyis includes effects that I want to discuss. The design will be:

> Items of *Depression List* : Please t / m Future

Example 2
See the example of the BPS, which consists of three subscales, *Cognition, Mood* and *Contacts.* Three different sum scores are calculated per person. Thus, three reliability analyses have to be conducted. Each subscale has an internal consistency reliability that should be estimated. In the design you specify which items belong to which subscale.

4.6 Degree of control

Does not apply. Or passive-observing, you might say.

4.7 Aggregated data

The aggregated data is comprised of the correlations between the items, and the standard deviation of each item, and *N*. This is sufficient to complete the analysis.

Example
In the running example, Diesfeldt provided only correlations. This is not sufficient to enable reproduction of the reliabibility analysis because the standard deviations are

lacking. In the following it will be assumed that the standard deviations are all equal to 1.

4.8 Analysis

In this section you describe which kind of reliability coefficient is being used. We limit ourselves in this book to the **internal consistency reliability.** That is, by definition, the reliability of the total score, based on the statistical relationships between different items in a single test administration in a population of subjects. The most common estimate for internal consistency reliability is **coefficient alpha**. Alpha is used so frequently that it is often regarded as synonymous with internal consistency reliability. There are other, similar and sometimes better estimates, for example, based on the factor loadings. Especially Guttman's **coefficient lambda 2** is a better estimate for internal consistency reliability than coefficient alpha, and it is equally easy to calculate in SPSS. The term internal-consistency reliability therefore has a broad and a narrow meaning. In the broad sense, it is the collective name for such coefficients as alpha and lambda 2, which estimate the reliability of the total score based on relationships between items. In the narrow sense, it is another name for coefficient alpha. Therefore, as soon as a specific coefficient is chosen, it should preferably be referred to by the specific name of that coefficient, and not only with the somewhat ambiguous term 'internal consistency reliability'.

In the following, for the sake of readability, the internal consistency reliability will often simply be designated by the term **reliability**. However, please be aware that there are other forms of reliability, such as test-retest reliability and inter-rater reliability. These are not treated in the present chapter.

If the analysis is based on SPSS output, then lambda 2, and not alpha, may be the best estimate of reliability (at least as far as its population value is concerned). Nevertheless, in an article it is wise to also report alpha; not to estimate the reliability, but to prevent the manuscript being rejected by ignorant reviewers. Furthermore, it is definitely permitted to use other software than SPSS to obtain better estimates than lambda 2.

Explanation

Coefficient alpha was initially only calculated for dichotomous items, with a formula that became known as ‘KR-20’ (Formula 20 Kuder & Richardson, 1937). The general formula, which is also suitable for non-dichotomous items, was first formulated by Guttman (1945), who named this coefficient lambda 3 (λ_3). Guttman also proved that this is a lower bound for reliability. The coefficient was studied in an article by Cronbach (1951), who called it coefficient alpha (α). That name has since been customary. Cronbach showed that alpha equals the average of all split-half reliability coefficients of the test.

Although alpha is often called the internal consistency reliability, Cronbach (1951) already pointed out that ‘internal consistency’ is not a good interpretation of alpha because alpha increases with the number of items. That is, you may call it ‘internal consistency reliability’ but not ‘internal consistency’, just like Amsterdam is synonymous with ‘Dutch capital’ but not with ‘Dutch’. Apparently, it is necessary to repeat this every decade (Novick & Lewis, 1967; Green, Lissitz & Mulaik, 1977 Hattie, 1985; Cortina, 1993; Schmitt, 1996; Drenth & Sijtsma, 2006; Sijtsma, 2009). Nevertheless, the name internal consistency reliability will probably stay, so that each generation psychologists is confused again.

In order to clarify the preceding paragraph: ‘internal consistency’ means that the items fit to each other, and measure the same. This means they are unidimensional (or unifactorial). The item’s dimensionality affects alpha, but alpha also depends heavily on the number of items. The latter has nothing to do with the dimensionality. Check the following examples, in which it is assumed for convenience that the items all have equal variances.

- Suppose the items depend on five uncorrelated factors and that each item loads 0.9 on one factor and 0 on the other factors. The item set thus can be partitioned into five uncorrelated subscales. If each subscale contains the same number of items, then with 100 items the alpha for the total scale equals 0.95. For someone who erroneously believes that alpha measures the internal consistency, this high value of alpha would suggest that the overall scale is internally consistent, whereas it actually falls apart into five uncorrelated factors.
- Suppose that the correlations of all items is 0.1. The items then satisfy a one-factor model. With five of such items will be alpha be equal to 0.36. For someone who erroneously believes that alpha measures the internal consistency, this low value suggests that the items are not internally consistent, whereas they actually are perfectly consistent.

Given that the term 'internal consistency' is still prevalent, one wonders how many researchers are aware that this term suggests a misinterpretation. Of course it could be that they only use the words without attaching meaning to them, but then it would be better to call the coefficient ‘beautiful sunset’, as to avoid misinterpretations.

That alpha equals the average of all possible split-half-reliabilities, can be interpreted in this way: If you randomly split the test into two equally long subtests, then the *expected value* of the split-half reliability equals alpha. This argument can be generalized and applies also when the test is split into more than two parts. Related to this is the following interpretation of alpha: if two equally long tests each are composed by taking a random sample of items from a large universe of allowed items, then their correlation is expected to be about equal to their alpha. This is an interpretation in terms of the generalizability theory (Cronbach, Rajaratnam & Gleser,

1963 Rajaratnam, Cronbach & Gleser, 1965; Cronbach, Gleser, Nanda and Rajaratnam, 1972).

Generalizability theory uses analysis of variance to attribute the variance of scores to multiple sources. In the simplest case, the sources are Item and Subject. While in classical test theory (Lord & Novick, 1968) the items are considered as a fixed factor, generalizability theory assumes they are a random factor. Despite this difference in interpretation, the same coefficient alpha will be used if there are no other factors besides the factors Subject and Item. Generalizability theory also provides formulas for designs with more factors, for example if there are different observers in addition to various items and subjects (see for example Veldhuijzen, Goldenbeld & Sanders, 1993).

Alpha is not necessarily equal to the reliability, but is only a lower bound for the reliability. Often, the reliability will be somewhat greater than alpha. Guttman (1945) derived six coefficients, named λ_1 through λ_6, all of which are a lower bound for reliability. All of them can be calculated with SPSS. Of these, λ_3 is equal to alpha. Coefficient λ_2 is always at least as large as alpha (Guttman, 1945; Jackson & Agunwamba, 1977 Revelle & Zinbarg, 2009). In other words, if the reliability of the test is denoted by ρ_X, then

$$\alpha = \lambda_3 \leq \lambda_2 \leq \rho_X$$

It follows that one should use lambda 2 rather than alpha as an estimator for reliability. After all, in the worst case, lambda 2 equals alpha, and in other cases lambda 2 is closer to the reliability. *If only SPSS is being used, my advice is to use λ_2 for estimating the reliability.*

Revelle and Zinbarg (2009) analysed nine data sets where λ_4 is greater than λ_2, and often considerably larger. This would have been a reason to advise λ_4, except that this is a split-half reliability, so the outcome depends on how they divide the test into two halves. Revelle and Zinbarg used the maximum over all possible splits, but that value is not easy to obtain with SPSS.

One may wonder why one would not simply use the maximum of all six λ coefficients. However, this is not recommended by any modern author, possibly because they are afraid of capitalization on chance (Ten Berge & Zegers, 1978, p. 579).

Ten Berge and Zegers (1978) showed that there is an infinite series of increasingly better lower limits, of which alpha and lambda 2 are the first two (see also Ten Berge, Cutters & Zegers, 1981, Osburn, 2000). The other lambda coefficients of Guttman are, incidentally, not a part of that series. Jackson and Agunwamba (1977) and Woodhouse and Jackson (1977) developed a coefficient called the 'greatest lower bound' (*glb*) of reliability. This is always at least as large as any of the other coefficients discussed

above, and often larger. However, the sample estimator of the *glb* can have positive bias (Ten Berge & Sočan, 2004).

Drenth and Sijtsma (2006; Sijtsma, 2009) are amazed that these improved lower limits are rarely used. They are usually higher than alpha, and you would think that psychologists are happy with that. Perhaps you think this is because these coefficients are not in SPSS is. That may be the reason why the *glb* is not used, but lambda 2 is actually in SPSS ! But it's not the default ... Sigh. Lambda 2 is better than alpha. That's known for more than 60 years. It has been in SPSS for decades. And yet the masses continue to report alpha. I sometimes dream of a program Shocking Statistical Program against Secondhanders which punishes users with an electric jolt if they use the default.

While alpha can be studied with analysis of variance (where the items are a fixed or a random factor), alpha can also be studied with factor analysis (see, for example, Cronbach, 1951; Mulaik, 1965). Within a factor analysis, there are also other estimates for reliability possible (Cronbach, 1988; Ten Berge & Hofstee, 1999; Zinbarg, Revelle, Yovel & Li, 2005). Revelle and Zinbarg (2009) advocate the use of McDonald's coefficient ω (omega). This is based on a factor analysis of the items. If there are *k* are items, and *V* is the sum of all elements in the correlation matrix (including diagonal), and *H* is the sum of all communalities, then the value of ω for standardized items can be calculated as follows:

$$\omega = 1 - \frac{k - H}{V}$$

According Revelle and Zinbarg this is a better estimate for the reliability than the *glb*. Coefficient ω can be calculated with the program **psych** in the free statistical package R (Revelle, 2008). A disadvantage of ω is that it depends on the number of factors that one assumes, and we have seen in previous chapters that this is not always something we quickly agree about. An advantage of ω is that, once one has the correct number of factors, it is a better estimate of the reliability than *glb* and the λ coefficients.

To summarise: As an estimate of the internal consistency reliability one can best use ω, provided that the number of factors is clear. Next best is the *glb*. Confined to SPSS, the best choice is λ_2. Those who want to publish shall also report α, but preferably interpret it as 'estimated generalizability coefficient' rather than 'estimated reliability'. It is formally correct to call α the 'internal consistency reliability', but many readers might interpret it incorrectly.

Running SPSS

Analyze > Scale > Reliability Analysis
Move the variables with item scores into Items
Select at Model: Guttman
Click the Statistics button

Check under *Descriptives for*: Scale if item deleted
Check under *Inter-item*: Correlations
Check under *Intraclass correlation coefficient*
Model: Two-way mixed (default)
Type: Consistency (default)
Continue
OK

If the data contains the correlation matrix instead of the raw data, the analysis must be done from syntax. Click Paste instead of OK. After the / VARIABLES subcommand insert:

```
/ MATRIX = IN (*)
```

Example

Diesfeldt used alpha as a measure of internal consistency reliability. If you run SPSS in the way described above, you get alpha and lambda 2 and a confidence interval for alpha.

```
RELIABILITY
    / VARIABLES = v1 to v15
    / MATRIX = IN (*)
    / FORMAT = NOLABELS
    / SCALE (Guttman) = ALL / MODEL = Guttman
    / STATISTICS = CORR
    / SUMMARY = TOTAL
    / ICC = MODEL (MIXED) TYPE (Consistency), CIN = 95 TESTVAL = 0 .
```

4.9 Estimators

If an estimated lower bound of the reliability, report lambda 2 when the analysis is performed with SPSS. In addition, alpha (lambda 3) should be reported. For each item, determine whether the reliability increases when the item is removed. For this purpose, alpha-if-item-deleted can be used. The values of alpha-if-item-deleted do not have to be reported, but it must be reported whether there are items whose removal would result in a larger alpha.

Explanation

The previous section advised to use lambda 2 for estimating the reliability and alpha was criticized severely. The main reason to report alpha too was the opportunistic consideration that your article is otherwise unlikely to be accepted by a top journal (Sijtsma, 2009), since these jounals are quite conservative in statistical matters.

The values of alpha-if-item-deleted can be compared to alpha in order to improve the scale (see section 4.11 Decision). In theory, one might be able to do better with lambda-if-item-deleted, but these values are not provided by SPSS. That is another reason why one could report alpha despite all the criticism. Alternatively, one could of course take the trouble to do a separate analysis for each item, and still get the values of lambda-if-item-deleted. But that take may take up to fifteen minutes, and psychology is not important enough for that, right?

Example
The SPSS output is presented in Table 4.1 to 4.3. The correlation matrix is omitted by me because it is already shown in table 2.3.

Table 4.1

Reliability Statistics

Lambda	1	.771
	2	.832
	3	.826
	4	.796
	5	.811
	6	.848
N of Items		15

Table 4.2

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Satisfied (-)	.00	56.00	.563	.415	.808
Sleeping (-)	.00	59.62	.311	.248	.824
Eating (-)	.00	60.46	.255	.229	.828
Healthy (-)	.00	56.68	.514	.406	.811
Tired	.00	58.94	.357	.244	.821
Old	.00	57.52	.455	.298	.815
Lonely	.00	56.58	.521	.412	.810

Friends (-)	.00	60.00	.285	.185	.826
Visitors (-)	.00	57.76	.438	.397	.816
Gloomy	.00	55.12	.626	.464	.803
Boredom	.00	58.22	.406	.265	.818
Excited (-)	.00	54.98	.637	.498	.803
Helpless	.00	58.78	.368	.207	.820
Weak	.00	58.36	.397	.417	.819
Future (-)	.00	56.44	.531	.383	.810

Table 3.4

Intraclass Correlation Coefficient

	Intraclass Correlation[b]	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.240[a]	.199	.290	5.739	196	2744	.000
Average Measures	.826[c]	.788	.859	5.739	196	2744	.000

Two-way mixed effects model where people effects are random and measures effects are fixed.

a. The estimator is the same, whether the interaction effect is present or not.

b. Type C intraclass correlation coefficients using a consistency definition. The between-measure variance is excluded from the denominator variance.

c. This estimate is computed assuming the interaction effect is absent, because it is not estimable otherwise.

Coefficient alpha is given under Lambda 3, and it is equal to 0.826 here. A better estimate for the reliability is Lambda 2, which has the value 0.832 here. The reliability is therefore estimated to be at least 0.832. The values of alpha, if one item is removed, are in the column Alpha if Item Deleted. The Eating item has alpha-if-item-deleted 0.828, which is slightly larger than the value of alpha. Removing this item would therefore lead to a larger value of alpha. In section 4.11 Decision, this is further discussed.

For completeness, we also consider the value of ω. For this we first have do a factor analysis. If one factor is extracted, you get from the communalities $H = 3.874$. From the correlation matrix we obtain $V = 65.42$. The number of items is $k = 15$. If you enter

this into the formula, you get: ω = 1 - (15-3.874) / 65.42 = 0.830. With four factors and ML extraction, you get ω = 0.875. With nine factors you get ω = 0.918, but that is a bit silly, because no one seriously claimed that there are nine factors in these data. Nevertheless this makes clear that you must first know the number of factors before you can calculate ω.

4.10 Testing

It is not customary to conduct a statistical significance test on the reliability. The reliability is usually so high that it is significantly greater than 0. Moreover, this is not an interesting null hypothesis, since the reliability has to be much higher than 0.

However, it may be relevant to report a confidence interval for alpha. SPSS does not provide confidence intervals if you ask only for alpha. But the Reliability procedure can also calculate intraclass correlations, and if you take the two-way mixed model and consistency type, then the outcome reported for the average rater is equal to alpha. And for that intraclass correlation, a confidence interval will be calculated. Incidentally, this is based on the outdated method of Kristof (1963) and Feldt (1965) (see McGraw & Wong, 1996), and there are now better methods (Van Zyl, Neudecker & Nel, 2000; Kistner & Muller, 2004; Maydeu-Olivares, Coffman & Hartmann, 2007).

Example

The value of alpha is 0.826, and the confidence interval ranges from 0.788 to 0.859. Note that alpha itself has again an unreliability. The latter unreliability mainly depends on the number of people, while the size of alpha depends mainly on the number of items and their correlations. The fact that one can easily calculate a confidence interval for alpha, might be a third reason to report alpha, and not only lambda 2.

4.11 Decision

In this section of the basic report you describe which items need to be removed from the scale (or vice-versa, which are retained in the scale). Those are the items that reduce the reliability, so that removal actually leads to an increase of reliability. Items with negative correlations with other items are also eligible to be removed, but usually these items will also have a negative effect on reliability.

After removal of the item, the reliability analysis must be done again with the remaining items, because it is possible that subsequently another item shows a negative contribution to the reliability, even if that item did not have negative contribution initially. The reverse is also possible. Therefore it is advisable to remove one item at a time, and redo the analysis after each item removal.

If no items with negative contribution are left, then a reliability estimate for the resulting scale has to be reported.

Explanation
If the items are one-factorial and have equal loadings, they can not reduce the reliability. In other cases, it is possible that an item has a negative effect on the reliability and that removing the item increases reliability. Take the extreme example of a situation where you have an intelligence test with a reliability of .95. And now you add this item to it: the number of eyes that the tested person throws with a dice. This item has factor loading 0, but it would not violate the one-factor model. Nevertheless, it is intuively clear that adding this item will reduce the reliability of the total score.

The instruction to remove items with a negative contribution to alpha, seems quite logical, yet it is controversial. After all, alpha is generally not equal to the reliability and maximising alpha can therefore lead to a decrease in reliability (Raykov, 2007) and criterion validity (Raykov, 2008). Furthermore, if one wants to maximise alpha it would be more apropriate to weigh the items differentially in the sum score, and then items with a negative contribution to alpha may still positively contribute to the reliability of the weighted sum score.

Example
We had already found that alpha is 0.826. If you look in the output in the column *Alpha if item deleted*, you see that after removal of item v3 (Eating), the alpha would increase from the current 0.826 to 0.828. This item should therefore be removed. You might wonder whether such a small increase is significant. Probably not, but on the other hand keeping the item is certainly not a significant improvement, so why would one keep the item? In addition, this item also has a negative correlation (-0.12), with another item.

Many people seem to have some emotional resistance to the removal of items. They seem to think "the more items the better”. If it is not quite certain that the item should be removed, they want to keep it. But then one confuses ends and means. The reason that having more items is better in general, is that it increases reliability provided that the items are good. But that last assumption is under discussion here. Better create a short, pure scale than a long scale that can be disputed.

If v3 is removed, and we re-do the analysis without this item, then it turns out that v8 (Friends) has a negative contribution. And if we remove that one too, then v2 (Sleeping) contributes negatively. After removal of that one, there are no more items with a negative contribution. Then alpha is 0.830, and lambda 2 is 0.835 (you cannot infer that from the output given here).

The conclusion should thus be that (if the scale is not split into sub-scales) the items Eating, Friends and Sleeping should be removed. These are also precisely the items with the lowest loadings in the factor analysis with one factor. That's no coincidence. A low factor loading means that the item correlates lowly with the factor, so it has a low reliability. If the item is too unreliable, then it adds more to the measurement errors than to true scores, and thus reduces the reliability of the total. Or, even worse, the item

measures a different factor and thus provides a systematic bias when you use it anyway. To illustrate this effect in this chapter I made the wrong choice of using all items, although the factor analysis indicated that we had different scales. Figure 4.1 shows a plot of alpha-if-item-deleted on the factor loadings. As you can see: potayto potahto.

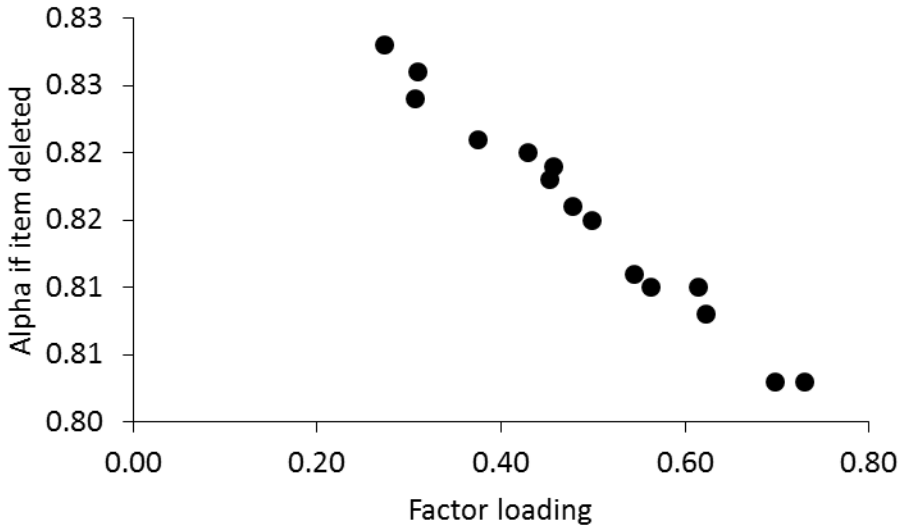

Figure 4.1

4.12 Interpretation

in this section of the basic report you assess whether the reliability of the resulting scale is high enough. There are no objective limits. It depends on the application (Cortina, 1993; Schmitt, 1996). If the test scores are used for decisions about individuals a higher reliability required than if the test is only used to establish in group research whether the test correlates positively with another variable. Because a low reliability means that each score contains a lot of random noise, and such noise can change the decisions about individuals but it cannot change the sign of a correlation.

Moreover, the boundary in decisions on individuals depends on how important the decision is. The higher the cost of a wrong decision, the higher the reliability should be. In a national final mathematics exam for thousands of high school pupils we demand more of the reliabiblity than in an exam for 50 students at a university. The demands of the university exam to the students hopefully exceed the requirements of the school exam, but the standards we set for the reliability of the university exam are

lower. Because firstly, the number of students is smaller and secondly for each student the cost of a wrong decision is smaller. This is so because if the student drops wrongly, the student will not have to retake the whole year (unlike the school pupil), and if the student erroneously succeeds, there are many other examinations to come where things can be put right.

In sum, it depends on how important the decisions are and how the decisions are influenced by unreliability. (That is, I suppose, the reason why during the World Cup nobody ever analyses the reliability of soccer games: it is completely unimportant). Most authors in the field of reliability refrain wisely from identifying any numeric boundary.

Explanation

Psychometricians have gone to great lengths to define and estimate the reliability coefficient, and professional organisations like the APA require that reliability coefficients be reported for every psychological test, but now we find out that nobody really knows how big it should be. Different views are discussed below.

1 The classic standard

While most authors avoid to specify any numeric border, one person (Nunnally, 1978; Nunnally & Bernstein, 1994) wrote down which value of alpha one should require. That bound was promptly embraced by a subpopulation psychologists as salvation from their suffering, because now they can at least refer to someone when they claim that their tests alpha is high enough. According to some, this benchmark is 0.70, while according to others, this benchmark is 0.80 (Clark & Watson, 1995; Duhachek & Iacobucci, 2003). And they refer to the same author. Sigh. The original text of Nunnally (1978) is less certain:

> In the early stages of research on predictor tests or hypothesized measures of a construct, one saves time and energy by working with instruments that have only modest reliability, for which purpose reliabilities of .70 or higher will suffice. (…) For basic research, it can be argued that increasing reliabilities much beyond .80 is often wasteful of time and funds. (…) In those applied settings where important decisions are made with respect to specific test scores, a reliability of .90 is the minimum that should be tolerated, and a reliability of .95 should be considered the desirable standard. (Nunnally, 1978, pp. 245-246)

Thus:

> 0.70 is acceptable in preliminary research
> 0.80 is good enough for group research
> 0.90 is the minimum required for individual decisions.

Group research has lower reliability requirements than individual decisions because in group research one is only interested in the correlation or the effect size. This will now

be explained in more detail. Two formulas are important: the attenuation formula and the Spearman-Brown formula.

The *attenuation formula* (Spearman, 1904b; see also Lord & Novick, 1968; Borsboom & Mellenbergh, 2002, Charles, 2005) shows how correlations are influenced by reliabilities. Suppose the manifest variables are X and Y, and their true scores are $T(X)$ and $T(Y)$, and their reliabilities are ρ_X and ρ_Y. The correlation of the true scores can be denoted with $\rho_{T(X)T(Y)}$ and the correlation of the observed scores with ρ_{XY}. One should be aware of the following: we would like to know the correlation between the true scores, but it is unknown. If we calculate the correlation based on a sample of data, we get an estimate for the correlation between the observed scores. The relationship between these two correlations is given by the so-called attenuation formula:

$$\rho_{XY} = \rho_{T(X)T(Y)}\sqrt{\rho_X \rho_Y}$$

From this formula it appears that the correlation between the observed scores is pushed down (attenuated) by the presence of the test's unreliability. For example : if the correlation between the true scores is .90, and the two tests have a reliability of .50, then the correlation between the observed scores will be only .45; much smaller than the true score correlation. Therefore, it is undesirable to use very unreliable tests in group research on correlations. The same formula, however, also shows that when the observed variable Y has a reliability of .80, to further increase the reliability will have a relatively small effect on the correlation.

An analogous formula is valid for the effect size in a study in two groups in which a *t*-test is being used. If the observed value of Cohen's *d* in the population is denoted by δ_Y, and the value of Cohen's *d* for the true scores of the dependent variable is denoted by $\delta_{T(Y)}$, then the relationship between them is (Cohen, 1988, p. 536):

$$\delta_Y = \delta_{T(Y)}\sqrt{\rho_Y}$$

As a result, the power of the *t*-test will be limited if the dependent variable has low reliability. Here too, increasing the reliability above .80 has a relatively small effect on the effect size. A similar formula is valid for one-factor ANOVA with a fixed factor (Feldt, 2011).

The *Spearman-Brown formula* shows how the reliability of a test increases if the test is lengthened (i.e., if the number of items increases) (Brown, 1910; Spearman, 1910; see also Lord & Novick, 1968). If the items are parallel, and if the current reliability of the test is ρ, then after lengthening the test by a factor k (for example, a test of 10 items is extended to a test of *k* * 10 items), the reliability is equal to :

$$\rho_k = \frac{k\rho}{1+(k-1)\rho}$$

A plot of this function shows that lengthening the test increases reliability, but the effect diminishes as the reliability gets closer to 1.

The attenuation formula and the Spearman-Brown formula together lead to the classic standard: lengthening the test has a positive but diminishing effect on the reliability (Spearman-Brown), while raising the reliability above .80 has a relatively small effect on the correlation and effect size (attenuation), so increasing reliability above .80 would be a waste of time and money (Nunnally, 1978).

2 Typical values of reliabilities

Frisbie (1988) argues that most published and standardized tests have a reliability of 0.85 to 0.95, and that those tests are usually evaluated as very acceptable, and that the lower limit for use in individual decisions should therefore be 0.85 when the decision is based on only that test score. Well, that's like the argument "$p < 0.05$ is a good bound to reject the null hypothesis, because almost everyone uses that bound" – a fallacy. Anyway, Frisbie says that for the group decisions the "generally accepted" lower bound is 0.65, and that paper and pencil tests of teachers have an average reliability of 0.50.

3 Optimizing the power given budget research group

The classic standard that increasing the reliability above .80 is a waste of time and money is a premature conclusion, because these costs vary by domain. Some experiments take several minutes, others take several decades. The classic standard ignores those differences.

Suppose we do an experiment with two groups, and that each subject spends b minutes undergoing the experimental conditions, and that it takes a further c minutes to take the test that measures the dependent variable. Suppose the test will be lengthened by a factor k to make it more reliable. In one experiment with n subjects, the total time you have pay the subjects will be $n(b+kc)$. Ellis (2013b) shows that under this condition, based on the attenuation formula and the Spearman-Brown formula and a result of Allison et al. (1997), the power of the *t*-test is maximized at a given budget if the reliability of the test is equal to

$$\rho_{efficient} = \frac{b}{b+c}$$

For example : if the experimental manipulation lasts 50 minutes, and the test takes 10 minutes, then the efficient value of the reliability is 50 / (50 + 10) = .83. This is the best

value of the reliability, the value that you have to achieve if you want to maximise the power of the *t*-test with a given budget. A higher reliability is inefficient: one would spend too much in the measurement while the power would increase more if the same money is spent to a larger number of participants. A lower reliability is inefficient too: it uses too many subjects, while it would be better to spend more time on measurements.

In this example, the efficient reliability is .83. However, if the experiment takes 180 minutes, with on top that a test session of 10 minutes, then the efficient value of the reliability is not .83, but 180 / (180 + 10) = .95. According to this analysis it is not possible to give a single value like 0.70 or 0.80 as standard for the reliability. It depends on how time consuming the experiment is, and how much time the test takes. In a long and expensive experiment it is worth pursuing a high reliability, perhaps .95. Conversely, if the test administration takes a relatively long time, it is better to accept a somewhat lower reliability, perhaps .60.

4 The probability of a wrong decision about an individual

The idea that the reliability should be higher if the test is being used for individual decisions is generally accepted, but it is illogical because the concept of reliability has no meaning for a single person. Reliability is defined as the proportion of true score variance, but for that one person the true score variance is equal to 0 and thus the reliability is 0 too. Another argument is that one person is a member of many different populations (e.g., I'm a Dutch man, but also a Gelderlander, and a restaurant customer, and coffee drinker) and in each population the reliability of the test is different, so which of those reliabilities would one have to use for that person? In deciding on individuals it is more logical to calculate a confidence interval for the true score, or to calculate the probability of a wrong decision (Swaminathan, Hambleton & Algina, 1974, Charter & Feldt, 2001).

A possible decision that one could study is whether a subject is above or below average (Veldhuijzen, Goldenbeld & Sanders, 1993). If the true score of the subject is above average and the observed score is above average, then the subject is correctly classified by the observed score. If the true scores and error scores are bivariate normally distributed, it is possible to derive the probability of correct classification from the reliability. These probabilities are shown in Table 4. 4.

Table 4.4

Reliability	*Probability of correct classification*
0.50	0.75
0.60	0.78
0.70	0.81
0.80	0.85

0.90	0.90
0.95	0.93
0.97	0.94
0.98	0.96

A similar question is : when two subjects are randomly drawn from the population, what is the probability that their observed scores have the same rank order as their true scores ? If true scores and error scores have a bivariate normal distribution, then the probabilities are the same as those in the right column of Table 4.4. The table shows :

- with a reliability of 0.50, the probability of correct classification is still 75%, where one should note that this probability starts at 50% if the test has reliability 0;
- if the desired error probability is 5%, then the reliability should be between 0.97 and 0.98.

Since it is commonly required that error rates may be at most 5%, one could argue that the reliability of a test should be at least 0.97 if one uses the observed score to compare an individual with the average or with another subject. Of course the 5% limit is just as well arbitrary ...

5 The standard error of measurement

The standard error or standard error of measurement is calculated as follows:

$$\text{SEoM} = \sigma\sqrt{1-\rho}$$

(The usual abbreviation is SEM, but this abbreviation is already used in this book for structural equations models). Here, σ is the standard deviation of the test scores in the population, and ρ is the reliability of the test scores in the population. The SEoM is equal to the standard deviation of the errors in the test score. When the standard deviation for each person is equal and the errors are normally distributed, then, a 95%-*confidence interval for the true score of a person* can be calculated:

$$X \pm 1.96 * \text{SEoM}$$

Here, *X* is the test score of the person. For example, suppose that an examination takes place in which participants receive a grade between 1 and 10, and that these grades have a reliability of 0.682 and a standard deviation of 1.346. With this it can be calculated that the standard error is equal to 0.759. If anyone on this exam has grade 5.5, the confidence interval for his or her true score ranges from 4.0 to 7.0. In other words, the grade that the person deserves (the true score), might be 1.5 marks higher or

lower than the grade that the person gets. If you find this error too big, then the exam has to be made more reliable.

The advantage of this calculation is that it can be applied to individuals. One limitation is that the assumptions made – the errors are normally distributed with equal variance for each person – are often not true, as we will see in the chapter on IRT.

A second limitation is that the problem is thus shifted to the question as to what is an acceptable margin of error. With exam results you have perhaps an intuition about this, but can you imagine what would be a good margin of error for the scores on the Depression List in Diesfeldt's research? And even with exam results, you might wonder whether the intuition you have is so reasonable. The foregoing example, wherein the confidence interval for one's grade ranges from 4 to 7, is based on real data. Did you expect such a wide range? Now, it is of course easy to scream that tests should be much more reliable. But in part 2B of this book there was a problem on a multiple choice exam of four possible answers per question, in which they wanted to distinguish reliably between 5 and 5.5. This showed the exam would need 2562 questions to achieve this. Assuming you answer an exam question every minute, the examination would take 42.7 hours...

Example
After removing the three items Eating, Friends and Sleeping the internal consistency reliability of the Depression List is 0.830. According to the classic standard this is high enough to use the scale in group research, but not high enough to allow individual decisions based solely on the scale score. If one compares any two individuals on the total score, then there is a probability of 10% to 15% that the wrong conclusion is drawn on who has the highest true score. If test sessions last 10 minutes, the test would be efficient for experiments of about an hour (including test administration). For longer experiments, a higher reliability would be more efficient.

4.13 Summary

Design
Items of *Depression List* : Please through Future.

Aggregated data
In the running example Diesfeldt provided only correlations.

Analysis
Diesfeldt used alpha as a measure of internal consistency reliability. This is normal, but it would have been better to use lambda 2.

Estimators
After removal of the items with a negative contribution to alpha, alpha is 0.830, and lambda 2 is 0.835. The estimated internal consistency reliability was 0.835.

Decision
The items Eating, Friends and Sleeping should be removed from the scale.

Interpretation
The reliability of the resulting scale is high enough for group research in accordance with the rules of thumb of Nunnally (1978).

4.14 Concise Report

An analysis of the internal consistency reliability of the Depression List showed that the items Eating, Friends and Sleeping had a negative contribution to alpha. These items were therefore removed from the scale. The estimated reliability of the resultant scale, consisting of the rest of the items, was lambda 2 = 0.835, while alpha = 0.830. According to the rules of thumb of Nunnally (1978) that is high enough to use the scale in group research.

5 Conducting and reporting a Rasch analysis

5.1 Background

Using factor analysis in the construction of tests has several drawbacks. The problem is that items usually have a small number of response categories, which is not compatible with the assumptions of factor analysis. There are alternative forms of factor analysis that are suitable for such data (Christoffersson, 1975, Muthén, 1978, 1984, Bartholomew, 1980). An alternative analysis method is based on the **Item Response Theory** (IRT). This is a collection of models for items with a limited number of answer alternatives (see Sijtsma & Junker, 2006, for an overview).

The Rasch model (Rasch, 1960) is the simplest IRT model, and will be introduced in this chapter. Because of its simplicity, the Rasch model has a number of attractive features that make the estimation and testing of the model relatively easy. However, that simplicity also means that the model is very restrictive. In practice, the model rarely holds for psychological tests. Nevertheless, the model is important, because it describes how the ideal test works.

The following first describes how an Rasch analysis is done. A later chapter discusses the important theoretical properties of the Rasch model.

5.2 Learning goals

After studying this chapter you can

- indicate the drawbacks of applying factor analysis to items;
- explain and motivate the assumptions of the Rasch model;
- describe the characteristics of the IRFs of the Rasch model;
- on the basis of given item parameters, sketch the IRFs and TRF in a Rasch model;
- assess which estimation method (CML , MML , WML , EAP) can best be used in a given case;
- schematically indicate how the variance of the person parameter depends on the sum score;
- on the basis of the results of the R_0, R_1 and R_2 tests, draw the correct conclusion about the unidimensionality of a set of items;

- write a concise report based on RSP output;
- indicate the similarities and differences between the Rasch model and the 3PL model, in particular as regards their applicability to multiple choice tests.

5.3 The problem of factor analysis

The use of factor analysis for analysing items in a test has a number of drawbacks:

- It is assumed that there is a linear relationship between the factor and the manifest variables. In practice, item scores are almost always bounded from above and below. The assumption of linearity is then implausible.
- Because of these bottom and ceiling effects, items with different averages usually also have different skewnesses. This limits their correlations, which leads to additional factors (McDonald & Ahlawat, 1974).
- Statistical tests usually assume that the manifest variables are normally distributed, but items usually have a limited number of categories, so this assumption is violated.
- By assuming a normal distribution for the manifest variables you essentially assume that you can already measure something, but that can be doubted (Fischer, 1974). The elementary observations in test data are not numbers, but responses from persons.

There are nonlinear factor analyses (McDonald, 1967; Yalcin & Amemiya, 2001), but they are not widely used in psychology. There are also tests for linear factor analysis that assume no normal distribution (Browne, 1984), and these are not used much in psychology either. Some methods of nonlinear, non-normal factor analysis can be seen as a form of IRT analysis (Muthén, 1978, 1984; Knol & Berger, 1991).

Explanation

If we assume a one-factor model for the items, and if we would know the factor scores, we could consider a scatter plot of, for example, four items. We could then create a plot that indicates how the item score depends on the score on the common factor. That should be a linear relationship for each item. The following figure contains an example in which the scores of 500 people were simulated according to a one- factor model. The syntax used is

```
compute g = rv.normal (0,1).
compute x1 = 0 + g + rv.normal (0,1).
compute x2 = 2 + g + rv.normal (0,1).
compute x3 = 3 + g + rv.normal (0,1).
compute x4 = 5 + g + rv.normal (0,1).
```

```
execute.
```

Figure 5.1 shows a plot with the factor scores *g* on the horizontal axis and the item scores *x*1 and *x*4 on the vertical axis. The scores of *x*2 and *x*3 have been omitted for greater clarity.

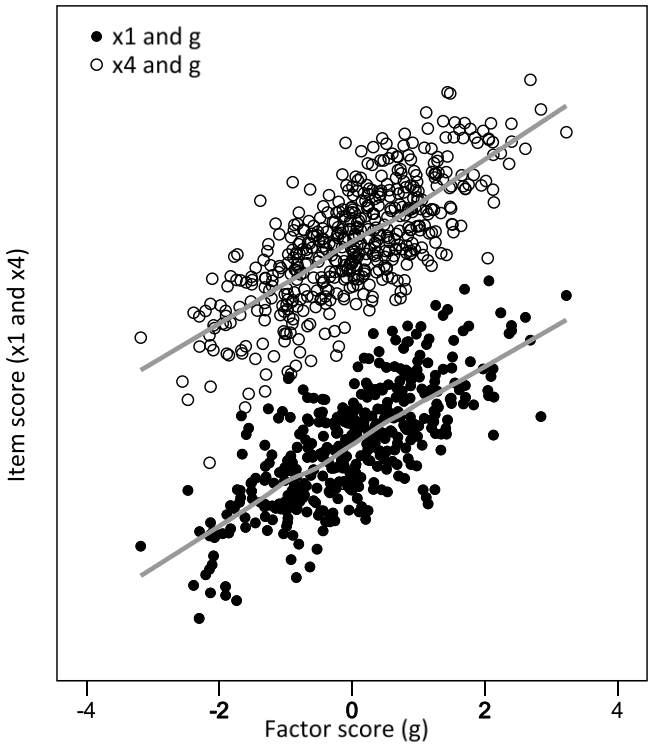

Figure 5.1

If we test a one-factor model with ML extraction on these data, then we get $\chi^2(2) = 1.555$, $p = .460$. In this analysis , therefore, we arrive at what we have put into it: one factor.

But suppose now that the researcher submits the questionnaire to a sample of subjects, where it is not allowed to give answers lower than 0 or higher than 5. The answers will then be truncated (assuming that the test subjects do not change their answers more than necessary). The plot then becomes as in figure 5.2.

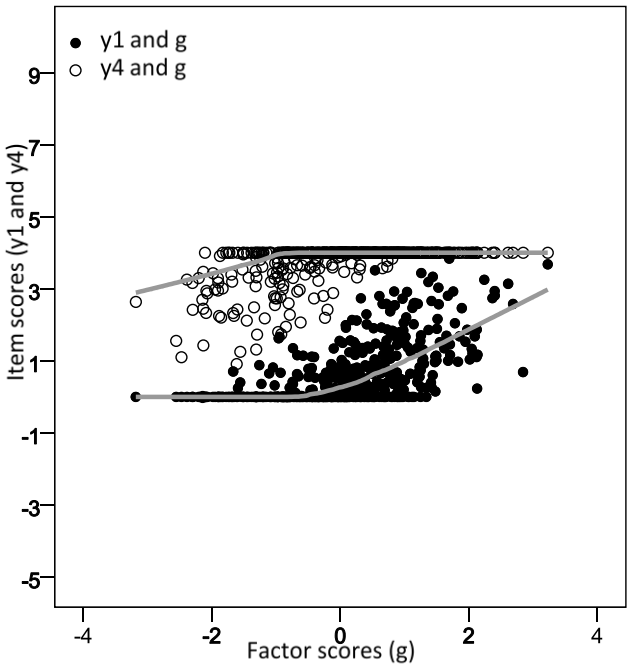

Figure 5.2

The truncation mainly concerns items 1 and 2 at the bottom and items 3 and 4 at the top. As a result, items 1 and 2 have become right-skewed, while items 3 and 4 have become left-skewed. But then it is impossible that they still have a positive linear relationship with each other. And thus they can no longer have a linear relationship with the same underlying factor. The trimming changes the pattern of correlations, and if we now do the same factor analysis on these data, we get $\chi^2(2) = 8.775$, $p = .012$. In this case we do not get what we put into it. We started with a one-factor model, but after truncation of the scores this is no longer correct. The researcher's conclusion would be that there is a second factor. In terms of content, however, nothing has changed, and the scores still depend only on the latent variable g . The researcher will therefore never be able to interpret his second factor properly. The second factor is *artificial*.

If the researcher tries to interpret the second factor, then it is possible to say what will happen. In this example, let's do a PCA because there are not enough items for ML extraction. The loadings are given in table 5.1 . The averages of the items are also shown there.

Table 5.1

Item	*Loading component 1*	*Loading component 2*	*Average score on item*
y1	.707	-.533	.5996
y2	.802	-.129	2.0439
y3	.784	.045	2,8298
y4	.641	.694	3.7869

You can see that the loadings on the second factor go hand in hand with the average. Item 1 has the lowest average and also the lowest loading on the second factor. Item 4 has the largest average and also the highest loading. In the case of an intelligence test, the averages would reflect the difficulty of the items. The second factor is therefore called a **difficulty** factor (McDonald & Ahlawat, 1974). For empirical examples see, among others, Miecskowski et al. (1993) and Van der Ven and Ellis (2000).

> *Summary:* The truncation of data that satisfy a one-factor model, can lead to an artificial second factor, the difficulty factor.

Data of items usually have an upper and lower bound, and can therefore be seen as truncated data. The conclusion is that factor analysis may not be such a good way to analyse such data. (That you still had to learn it, is because it is nonetheless used more frequently than IRT (Ten Holt, Van Duijn & Boomsma, 2010)).

5.4 Basic concepts of IRT

Item response theory assumes that the items have only a small number of response categories that do not necessarily indicate a quantity. The simplest IRT-models assume that each item has only two response categories. In this book only these simplest models will be discussed.

Items with only two response categories are called **dichotomous** or **binary**. Examples are:

- items where the answers are classified as ‘good’ or ‘wrong’;
- items where only ‘yes’ or ‘no’ can be answered;
- items where only ‘positive’ or ‘negative’ can be answered;
- items that establish whether a certain behavior is ‘present’ or ‘absent’.

That what is called the ‘common factor’ in factor analysis is called the **latent trait** in IRT. In both cases it is a latent (non-observable) variable. The only difference is that other assumptions are made about the relationship with the manifest variables.

IRT models are often developed with cognitive tests in mind. The latent trait is therefore often called the mental **ability** of a person, or the **difficulty** of an item. This

will also be done in this text. This may give the impression that IRT models only apply to cognitive tests, but that is not true.

5.5 Basic report of a Rasch analysis

We stipulate that a basic report of a Rasch analysis consists of these sections: design, degree of control, hypothesis, aggregated data, analysis, estimating, testing, decision and interpretation. These parts will be discussed below.

5.6 Running example

As an example, we take items B01 to B12 of Raven's Standard Progressive Matrices, administered in the Netherlands to a group of children aged 12 to 15 years. With these data a Rasch analysis was done by Van der Ven and Ellis (2000). The data can be found in the dataset *Raven.sav*.

5.7 Design

The design describes the items on which the analysis is done. A Rasch-analysis can be done on dichotomous (binary) items. Therefore, it should also be mentioned that the items are **dichotomous**. Sometimes a Rasch analysis is done on items with multiple response categories, which are then first recoded into two categories by combining some of the original categories. For example :

1 and 2 $\rightarrow$ 0;
3 and 4 $\rightarrow$ 1.

In that case, should be mentioned that the items are **dichotomized,** and which transformation has been used.

Explanation
Usually all items of the test are administered to all persons in the sample. You could call that a within-subject design, but it is not customary to mention that. In some programs, other, more complex designs are possible, for example group A items 1 through 15, and group B items 11 through 25. The groups then have only 5 out of 25 items in common. If such an unusual design is used, it must be described.

Example
The items are the twelve items from subtest B of Raven's Standard Progressive Matrices test. The items will be named B01 through B12 (dichotomous).

5.8 Degree of control

The degree of control is usually passive-observing.

5.9 Hypothesis

The hypothesis is that the items satisfy the Rasch model. The Rasch model has the following assumptions:

> *Unidimensionality.* Each person and each item is characterized by one number, which is called the person's ability and the item difficulty, respectively.
>
> *Double monotonicity.* The probability that the person will give a correct answer to the item, increases with the ability of the person, and decreases with the difficulty of the item.
>
> *Local independence.* The probability that the person has an item correctly depends only on the person's ability and, given this ability, not on the person's other answers.
>
> *Sufficiency of the sum scores.* Each person's sum score on the test contains all information about that person's ability that is contained in the person's response pattern. For each item the sum score on that item in the sample contains all information about the difficulty of the item that is present in the sample.

From these assumptions, supplemented with some technical assumptions such as differentiability, it can be deduced that the probability to answer the item correctly depends on the ability and the difficulty in the following way (Fischer, 1974, 1995):

$$P[X = 1 \mid \theta, \beta] = \frac{e^{\theta - \beta}}{1 + e^{\theta - \beta}}$$

The Greek letters θ (theta) and β (beta) represent this:

θ = the (mental) **ability** of the person
β = the **difficulty** of the item.

Implicitly, the function $\Psi(x) = e^x / (1 + e^x)$ is used, with $\theta - \beta$ being entered for x . The function Ψ is called the **logistic function**. Because each item is characterised by 1 parameter β, the Rasch model is often also called the **1-parameter logistic model** (1PL model).

Next, one can plot how the probability depends on the person's ability and the item difficulty. If this is done for one item, this is called the **Item Characteristic Curve** (ICC); the associated function is called the **Item Response Function** (IRF). That last name is nowadays the most common .

Explanation

Unidimensionality
The assumption that each person can be characterized by one number is comparable with the hypothesis of a one-factor model. This is a sensible hypothesis for scale analysis, because we want to investigate whether it is possible to characterize the person with one score (the test score).

The assumption that each item is characterized by one number is comparable in factor analysis to the hypothesis that the items have the same factor loadings while they can have different averages.

Because both people and items have only one parameter, we can represent them on a common scale, for example, like in figure 5.3.

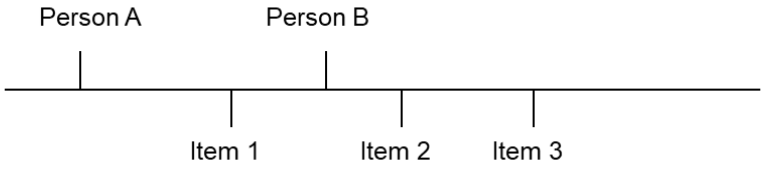

Figure 5.3

Here, the ability of person B is greater than that of person A. The difficulty of item 2 is greater than that of item 1. Person B has a greater ability than the difficulty of item 1, but smaller than the difficulty of item 2.

Double monotonicity
The assumption (or rather requirement) that the probability of a correct answer increases the *ability* of the person is a reasonable assumption. We do not want a test where the smart people get lower scores than the stupid ones.

The assumption that the probability of a correct answer decreases with the difficulty of the item, also seems obvious, but ... that is a bit more complicated. In many other IRT models this probability also depends on a second item characteristic, the discrimination.

Local independence
For example, if you need the correct answer to question 1 in order to answer question 2 correctly, then the questions 1 and 2 are **not** locally independent. By contrast, if you

have an equal chance on question 2, whatever your answer to question 1 was, then questions 1 and 2 are locally independent.

Sufficiency of the sum score

Some psychologists say that they not only want to look at the total test score, but also to the answer pattern: which items did the person answer correctly? Or, in the case of a questionnaire, to which items did the person respond positively? I heard psychologists saying this with pride, because in their view it means that they use the test intelligently. In a test that satisfies the Rasch model, however, that strategy is just as intelligent as talking to the rabbit that you see in the clouds. Sufficiency of the sum score means that all relevant information is contained in the sum score. Considering the specific response pattern can only produce noise.

Sufficiency of the sum score is a reasonable requirement, given the purpose of scale analysis. The sum score will be used to say something about the person's ability, and then it is desirable for that sum score to be a good summary of all the information contained in the person's answers. Sufficiency means that the sum score is a *perfect* summary of the empirical information about the person's ability.

The IRF

Figure 5.4 shows how in the Rasch model the probability of a correct answer depends on the person's ability. The ability is shown on the horizontal axis. Each item has its own curve, and that curve is called an ICC or IRF. Each of these curves is S-shaped. In this example, items 1 through 5 have respective difficulties β = -2, -1.5, 0, 1.5, and 2. The item difficulty is the ability at which the IRF of the item has a height of 0.5. The items have increasing difficulties in this example. In the figure this implies that the curve of an item is always shifted to the right in relation to the previous item.

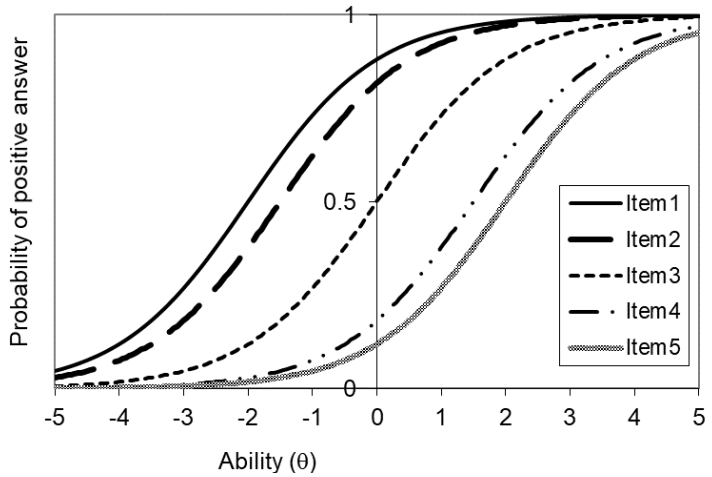

Figure 5.4

From this figure you can see the following:

The IRFs are **increasing.**
The IRFs are **S-shaped.**
The IRFs go to **0** (minimum) at the left and to **1** (maximum) at the right.
The IRFs do not intersect each other.
The IRFs are **'parallel'** (the horizontal distance between two IRFs is constant).

The difficulty of an item is the point above which the IRF assumes the value 0.50. So the harder the item, the further the IRF of that item is to the right. A person whose ability is the same as the difficulty of the item will have a 50% probability of making the item correctly. *(The next two sections were not translated)*

5.10 Aggregate data

5.11 Analysis

5.12 Estimators

The estimators consist of two tables:

- for each item the estimated item parameter (the difficulty of the item), and
- for each possible sum score the estimated person parameter (the ability of the person).

If MML is used, the estimated mean and standard deviation of the normally distributed latent trait is also reported.

Explanation
The estimated person parameter does not have to be reported separately for each person, because it depends only on the sum score of the person. In practice, the estimated person parameters usually correlate about .99 with the sum score.

Example
The following tables contain the estimated item parameters and the estimated person parameters, respectively. In the case of person parameters, the standard deviation and frequency are also mentioned in this example, although this is not necessary according to the above rules.

Table 5.4 **CML estimates or item parameters**

Item	*Difficulty (B)*
B03	-1.609
B04	-1.459

B05	-.951
B06	.380
B07	.610
B08	.892
B09	.231
B10	-.335
B11	.579
B12	1.662

Table 5.5 **Weighted ML estimation of ability**

Score	*Parameter*	*St.Dev*	*Freq*
0	-3.579	1.606	0
1	-2.271	.997	1
2	-1.533	.831	0
3	-.953	.755	3
4	-.445	.717	5
5	.027	.703	9
6	.490	.709	16
7	.973	.741	41
8	1.520	.814	91
9	2.228	.980	216
10	3.516	1.589	328

Based on the estimated item parameters we can now draw the IRFs, assuming that the Rasch model holds (which we do not know yet). These are shown in Figure 5.7.

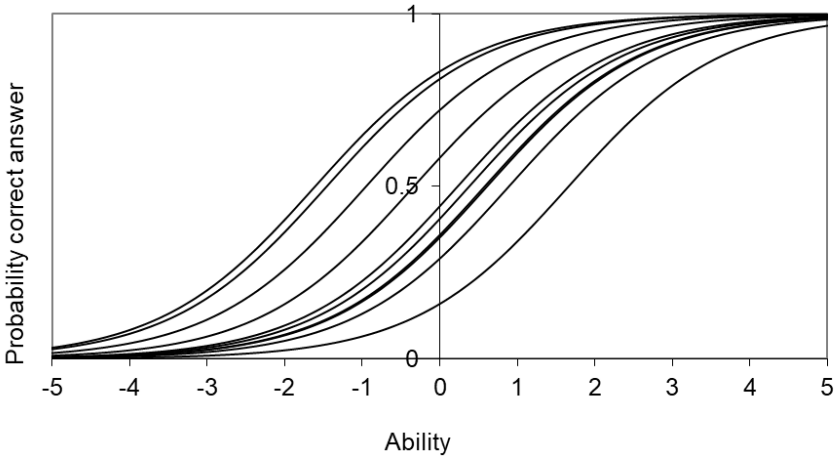

Figure 5.7

By adding the curves you get the so-called Test Characteristic Curve or Test Response Function (TRF), which indicates how the expected test score depends on the ability of the person. This is shown in Figure 5.8. This curve is often S-shaped , but not always. That the curve is steeper in the middle, near 0, means that the test discriminates best among individuals with ability around 0.

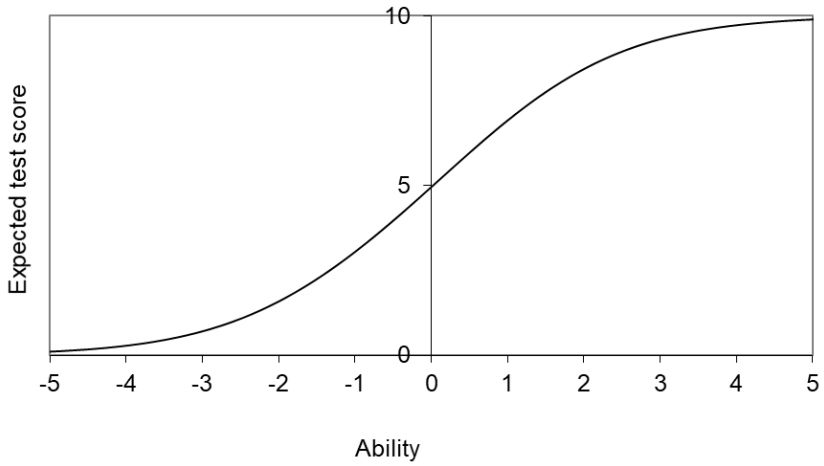

Figure 5.8

The curve is not linear. Thus, if another variable is linearly related to the ability, then it *does not* correlate linearly with the sum scores. GLM can apparently be trashed if you want to use the sum scores as a dependent variable or covariate in a follow-up study. But in the middle the curve is almost linear. If almost all persons are in that area in terms of their ability, there may not be a major problem with GLM. However, that situation may not occur.

Remember now that within a group with the same sum score there are still people with different abilities. The sum score is not a perfect measurement of the ability. According to Table 5.5, the distribution of abilities is particularly great with the extreme sum scores 0 and 10. This is plotted in figure 5.9. This pattern is typical for IRT models: extreme sum scores have large measurement errors. Such persons fall probably outside the range of the scale.

The fact that these standard deviations are so different implies that the assumption of homogeneous error variances in GLM is not plausible. Nor is it adequate to express the reliability of the test in a single number, as is done in classical test theory. According to the Rasch model measurements are much more precise for some persons than for others.

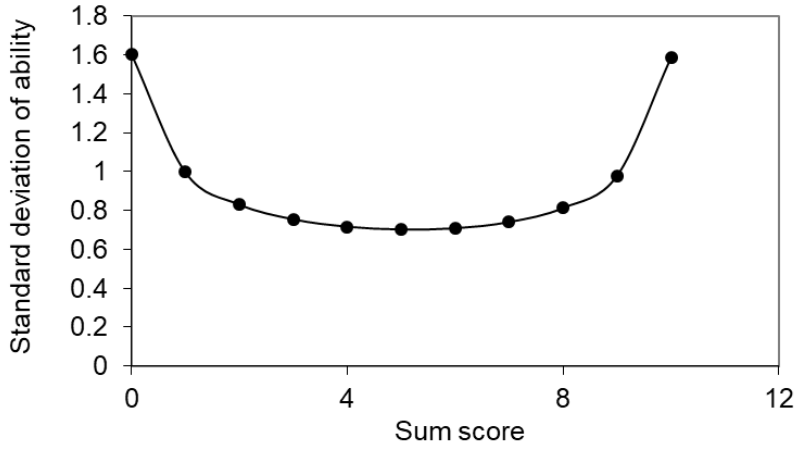

Figure 5.9

(Sections 5.13 – 5.17 were not translated)

5.18 The 3PL model

The **three-parameter logistic model** (3PL model) (Lord, 1980) states that the IRFs have the following form:

$$P[X=1|\theta,\alpha,\beta]=c+(1-c)\frac{e^{\alpha(\theta-\beta)}}{1+e^{\alpha(\theta-\beta)}}$$

Here every item has three parameters:

β = the difficulty of the item
α = the discrimination parameter of the item
c = the guessing parameter of the item

Each person has one parameter, θ, the person's ability. The IRFs of this model look like in figure 5.11. Unlike the Rasch model, the IRFs can cross each other.

The guessing parameter indicates the minimum of the IRF. That is about the height at which the IRF starts in the figure. This stands for the probability that someone with an ability of minus infinity still gives the right answer by guessing. Imagine that, ability minus infinity.

All curves are S-shaped in this model. The discrimination parameter indicates how steeply the curve increases at its steepest point. The difficulty indicates at which ability value the curve is the steepest. The greater the difficulty, the farther to the right is the curve. The discrimination parameter indicates how strongly the probability to answer

the item correctly depends on a person's ability. The discrimination parameter is in many ways comparable to a factor loading. When determining a total score, an item should be weighted more heavily as it has a larger discrimination parameter.

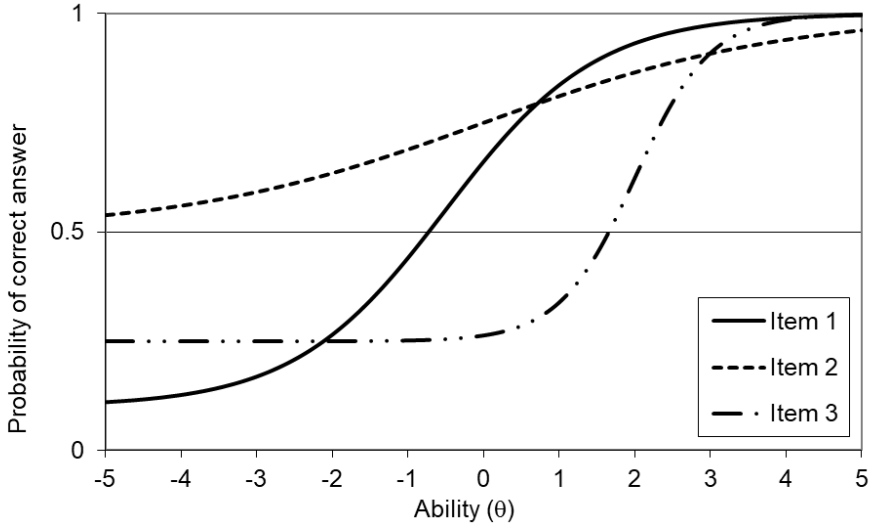

Figure 5.11

In Figure 5.11, item 1 has a high probability of guessing (0.5) and a low discrimination parameter (0.5). If these items were exam questions, it would mean that someone who knows nothing of the examined topic still has a probability 0.50 to answer the item correctly, while this probability increases only slowly as someone has learned more. In contrast, item 3 has a smaller guessing probability (0.25) with a large discrimination parameter (2), and a high difficulty (2). For persons with ability around 2 (the difficulty) the probability to answer the item correctly increases rapidly (high discrimination parameter) with their ability.

To give you an idea of the distinction between discrimination and difficulty: if an exam question is unclear or unexpected, this may result in a low discrimination parameter. If the question is clear and to be expected, the discrimination parameter may be high, but the item can still be difficult or easy.

The 3PL model is more flexible than the Rasch model, because it has more parameters. That is why it is also used more often. In particular, the 3PL model is suitable for multiple choice exams, because it has a guessing parameter. The Rasch model does not take into account the possibility that guessing can give a correct answer, and is therefore not plausible for multiple choice questions.

Another advantage of the 3PL model is that the probability of guessing the correct answer can be estimated *from the data*. As a result, a substantiated correction for guessing can be given when computing the grades. That is much better than the classic

assumption that all alternatives are equally likely if someone guesses (Holzinger, 1924), which is almost always implausible.

Between the Rasch model and the 3PL model lies the 2PL model (Birnbaum, 1968). That is a 3PL model 3 whereby it is assumed that there is no guessing: $c = 0$ for each item. However, the items may still have different discrimination parameters. Because of estimation problems with the 3PL model, the 2PL model is used frequently.

6 Conducting and reporting a Mokken analysis

6.1 Background

The Rasch model is one of the strictest IRT models, and as a result there are few tests for which it holds. Since the 1970s there has been a group of psychometricians who feel that such a strict model is not necessary for many applications in the social sciences. The Rasch model was developed with cognitive tests in mind, which often contain large numbers of items and are widely used to make important decisions about individuals. Many scales in the social sciences, however, involve attitude or personality questionnaires, which contain a small number of items, which are examined in relatively small samples, and which are mainly used in group research. With such items, the Rasch model is less plausible and less necessary in the eyes of some (Junker & Sijtsma, 2001b).

For example: the SAT and the ACT are two tests in the United States and are being used for admission to universities and colleges. The SAT was administered to more than 1.4 million students in 2006 (College Board, 2007). The ACT consists of 215 items and was administered in 2007 to 1.3 million students (ACT, 2008). Such tests have to meet different standards than a scale of 11 items that is used only once in an experiment with 50 students to investigate whether an information spot of the Ministry of Social Affairs has changed their attitude.

Partly for this reason, various psychometricians developed *nonparametric* IRT models, in which only ordinal restrictions are imposed. In particular, it is not assumed that the IRF is logistic. In the Netherlands, the model of ‘monotonous homogeneity’ and ‘double monotonicity’ has been developed by Mokken (1971, 1997, Mokken & Lewis, 1982) and extended to items with more than two answer categories by Molenaar (1991, 1997). The basic elements of this analysis method are discussed here. In the United States and Canada similar models have been developed by Stout (1990) and Ramsay (1991), but with different methods of analysis. See Junker and Sijtsma (2001a, 2001b) and Sijtsma and Meijer (2007) for an overview. *(6.2-6.6 not translated)*

6.2 Learning goals

6.3 Basic report of a Mokken analysis

6.4 Running example

6.5 Design

6.6 Degree of control

6.7 Hypothesis

Mokken (1971) distinguishes two different models. The first is the least restrictive and is called **monotone homogeneity**. It contains the following assumptions:

> *Unidimensionality.* Each person is characterized by one number, which is called the ability of the person.
>
> *Monotonicity.* The probability that the person will give a correct answer to the item increases with the ability of the person.
>
> *Local independence.* The probability that the person answers an item correctly depends only on the person's ability and, given that ability, not on the person's other answers.

With these assumptions it is (in theory) possible to order the persons according to their ability, but it is not always possible to order the items consistently in difficulty. The latter is possible in the second, slightly more restrictive model, with the following additional assumption:

> *Double monotonicity.* Each item is characterized by one number, which is called the difficulty of the item. The probability that the person will give a correct answer to the item decreases with the difficulty of the item.

The hypothesis which will be discussed here, is that the items satisfy monotone homogeneity. The investigation of double monotonicity will not be discussed here. The statistical test procedure also assumes that the items have sufficient discrimination in the population studied. That is why this will also be mentioned in the hypothesis.

Explanation

The '**ability'** of the persons is viewed as a **latent variable**, just like a factor in factor analysis and a true score in test theory. Thus the ability is not equal to the total score of the test. The total score is only an estimate of the ability, just as a sample average is an estimate for the population mean.

The term ability suggests that it is a cognitive test, but this is only a matter of jargon. The analysis can equally well be used for questionnaires and the interpretation must then be adjusted according to the content of the questions. Analogously, the term 'difficulty' is only a matter of jargon, and the interpretation must be adapted to the content. Suppose the items ask how many times you have been to the pub in the past week on each of the days in that week, then 'ability' must be interpreted as 'tendency to go to the pub' rather than a kind of intelligence. That Monday is 'more difficult' than Friday in that context means that people tend to go to the pub less often on Mondays than on Friday.

The **Item Response Function** (IRF) shows how the probability of a positive answer ('correct', 'yes', 'agree') increases with the ability. Figure 6.1 shows an example of four items that meet monotone homogeneity. The IRF can have any form, as long as it nowhere goes down. The IRFs illustrated here are purely theoretical. In practice, one cannot simply draw the IRFs because the latent abilities of the individuals are unknown and can only be estimated (however, see Ramsay, 1991 for a method to estimate the IRFs).

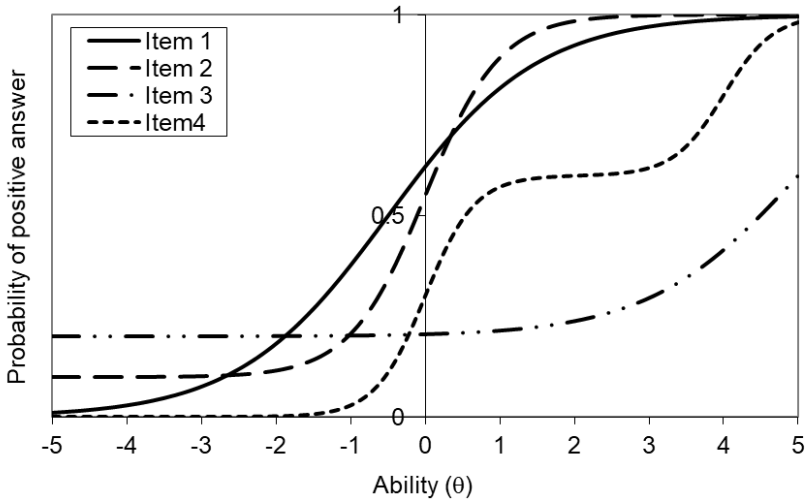

Figure 6.1

As you can see in the figure, it is possible that an IRF remains constant over a large interval. In the figure, for example, this is the case for Item4 between the values 1 and 3. This is not strictly in conflict with monotone homogeneity. However, if many abilities in the population are concentrated in this interval, the item have low discrimination, as these persons have the same probability to answer the item positively. The item is therefore useless in that group. The additional requirement of sufficient discrimination now says that this situation is not acceptable. For example, in Figure 6.1, items 1, 2 and 3 would be sufficiently discriminating for a population with

abilities between -1 and +1, but item 4 would not. In a population with abilities between 3 and 5, items 3 and 4, on the other hand, would be sufficiently discriminating, but items 1 and 2 would not. In a population with abilities from -1 to 5, all four items would be sufficiently discriminating. Whether the items are accepted, will therefore also depend on the location of the population relative to the IRFs. It is possible that in one population item 4 must be removed, while in another population items 1 and 2 must be removed, while no item has to be removed if both populations are joined.

We will assume a confirmatory approach, in which the hypothesis exists that the items form a scale. A Mokken analysis can also be done in an explorative way, in which case a set of items is searched for a subset that forms a scale.

In a Mokken analysis, double monotonicity is usually also investigated. But in most applications the purpose is primarily to measure persons, and not to measure the items. The most important hypothesis is therefore monotonic homogeneity, while double monotonicity is secondary. Since this is just an introduction, the discussion will be limited to monotone homogeneity .

Example

The hypothesis is that the items B01 to B12 satisfy monotone homogeneity and each have sufficient discrimination in the population.

(The rest of the chapter is not translated)

7 Exercises

- Make sure you always use syntax when you analyse a correlation matrix.
- Subsequent tasks can be more difficult and sometimes there are multiple solutions or no solution.

Answer-format

To streamline any discussion with others you need to report some analyses according to the format of the file *Answer Sheet.xls.* These questions are marked with → Answer Sheet. This file is described below. If you do not have the file, you can use the *Answer Sheet factor analysis* which is after the last question. *In addition, you also have to answer the other questions of the assignment!*

The Answer Sheet.xls file
This file has three worksheets:

- *Answers:* here you write up the design of the analysis and the conclusions.
- *TablesWithReport:* paste *only* output tables on which your conclusions are based.
- *All Output:* paste *all* SPSS output associated with the task, so you can still find stuff, in case you forgot this in *TablesWithReport.*

Parts A to E of the sheet *Answers* should be completed before you analyse the data. Make the other parts after analysing the data. Sometimes certain parts do not apply. You have to assess this yourself.

Exercise 0
Study carefully the text to be learned before you start with the exercises. It seems that some people start with the exercises before reading the text. That is just silly. It saves no time at all. It only creates confusion because you do not take the information in a logical order. Thus, if you have not yet thoroughly read the text, now is the time to do it.

Exercise 1
This exercise is about the theoretical concepts. The premise of this book is that you not only should you be able to 'do' a factor analysis, but you have to understand some theory as well, at least to the extent that can be understood without much mathematics.

It is advised to alternate between theory and practise. The theory becomes easier by applying it and applying it becomes easier if you understand the theory. Make this task before you make the other assignments, and make it a few times again later. The ultimate goal should be that you can answer all questions without reverting to the text.

1. What is the purpose of factor analysis?
2. Which aggregated data serve as input for factor analysis?
3. What is a confirmatory factor analysis?
4. What is an exploratory factor analysis?
5. What is the maximum likelihood criterion?
6. What is principal component analysis?
7. What is a factor?
8. What is a factor loading?
9. What is a factor pattern?
10. What is a communality?
11. What is an eigenvalue?
12. What are orthogonal factors?
13. What are skewed factors?
14. What is rotation?
15. What varimax rotation?
16. What is oblique rotation?
17. What is the null hypothesis in factor analysis?
18. Which criteria can be used to determine the number of factors?
19. What is the function of the chi-square statistic?
20. What does the chi-square depend on?
21. What is a goodness-of-fit index?
22. What is the RMSEA?
23. At what values of RMSEA is there a good / acceptable / bad fit?
24. What is the minimum eigenvalue criterion?
25. How do you assess whether a factor is interpretable?
26. Which items should you remove in scale analysis after a factor analysis?

Exercise 2

See exercise 1a. Although it is often instructive to answer open questions, checking multiple choice questions is easier. They are also more convenient to provide direct, automated feedback. If you completed exercise 1, open the *MC Questions.xls* program. It contains about the same questions, but in multiple choice format. Start the macro / add-in Answer Questionnaire. Choose the subject *Theory Factor Analysis.* When done, the questions and answers are displayed in the sheet *Answers.*

Exercise 3

An important element in factor analysis is the plot of the factor loadings. By understanding what it says, you can get a pretty good idea of what a factor analysis does, without going into details. In this assignment we will look at that for the personality theory of Eysenck. In the theory of Eysenck, based on factor analysis of personality tests, the factors *Neuroticism*, *Extraversion* and *Psychoticism* are distinguished. Initially, Eysenck distinguished only the dimensions of *Neuroticism* and *Extraversion*; *Psychoticism* was only added later. From that time dates the following well-known figure, figure 7.1, which is a fair description of a factor analytic theory (Eysenck & Eysenck, 1985).

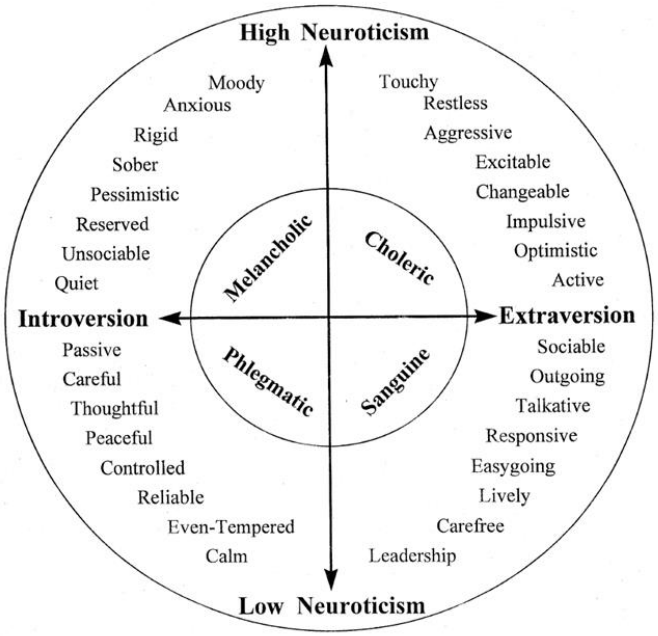

Fig. 1. The relationship between the four temperament types and the introversion-extraversion and neuroticism dimensions of personality (after Eysenck & Eysenck, 1985, p. 50).

Figure 7.1 (copied with permission from Robinson, 2001, page 1236)

This Figure is an idealized plot of the factor loadings. Each point in the outer circle represents a manifest variable. The variables in this case are, for example, Moody, Passive, Sociable, and so on. The factors are *Extraversion* and *Neuroticism*. These are latent variables that act as axes in the plot. The four classic temperaments in the inner circle are shown to clarify the relationship with this classical theory, but do not follow from the factor analysis.

According to this plot, every manifest personality trait can be seen as a mix of *Extraversion* and *Neuroticism* plus something that is specific to the trait. For example: Excitable (right) corresponds to a mix of high *Extraversion* and high *Neuroticism* plus something specific Excitable.

a. Which two variables have the highest correlation with Excitable?
b. With which two variables does Excitable have the lowest (the strongest negative) correlation?
c. With which four variables does Excitable have a correlation of approximately zero?
d. Make a schematic representation of the correlation matrix of the variables just discussed. Indicate the height of the correlations as 'high', 'zero' and '-high', the latter representing a strongly negative correlation.

Exercise 4

Let's begin to see if you can reproduce with SPSS the analysis of the main text. The correlation matrix is the *Diesfeldt.sav* file.

a. Do the confirmatory factor analysis described in the text (paragraph 2.10, Example 1) and make sure your results are the same:
    1. Open the data file.
    2. In the menu, specify which analysis should be done, but do not click OK yet. Ask for ML extraction and one factor.
    3. Instead of clicking OK, click on Paste. A new syntax window pops up with in it a syntax command.
    4. Remove the sub commands VARIABLES, MISSING and ANALYSIS. Put the subcommand / MATRIX = IN (COR = *) at that location.
    5. Put the cursor in the syntax command.
    6. In the menu bar click on the *Play* button.
    7. Should it fail, look in Exercise 4a_demo.zip.
    8. Check whether your outcomes match the text.
b. Do the exploratory factor analysis described in the text (paragraph 2.9), and make sure your results are the same (Section 2.10, Example 2).
c. Complete the Answer Sheet for (b). From now on, this instruction will be declared as → Answer Sheet.

Exercise 5

Now it is time to apply factor analysis in a simple situation of which you probably know the context, but of which you do not yet know the results. The file *Statistics 1.sav* contain the data on Statistics 1 of psychology students at Radboud University some years ago. The exam consisted of four partial examinations, A, B, C and D. The scores for the partial exams are in the variables a1, b1, c1 and d1 respectively. Exam Part B consisted of six questions, and marks for them are given in the variables quest1 to

quest6. The contents of these tasks are: (1) basic report of correlation and regression, (2) predicted scores and residues, (3) visualisation, (4), conclusions from the correlation, (5), contingency tables and the paradox of Simpson, and (6) the intuitive scientist. This exercise can also be made without syntax.

a. How many factors do you expect in quest1 to quest6?
b. Perform a factor analysis on the six questions of Exam Part B → Answer Sheet. If you do not succeed with SPSS, look in exercise 5b_demo.zip.
c. What is your psychometric evaluation of partial exam B?

Exercise 6

See exercise 5. It would be unsatisfactory if a slightly different choice in the analysis would yield a completely different result. That is a real danger in factor analysis. To what extent is that the case here?

a. If you have done a confirmatory analysis, do an exploratory analysis now. If you've done an exploratory analysis, now do a confirmatory analysis → Answer Sheet.
b. Check the extent to which the conclusions match the previous conclusions.

Exercise 7

Now we apply factor analysis in a simple situation that has been particularly important in theory. In some areas of psychology people are very attached to references to recent literature, which would make one believe that humanity is younger than 10 years. It can be enlightening to study the classics. The inventor of factor analysis is Spearman (1904a). In his research into intelligence he found the following correlations (table 7.1).

Table 7.1

	Classics	French	English	Mathem.	Discrim.	Music
Classics		0.83	0.78	0.70	0.66	0.63
French	0.83		0.67	0.67	0.65	0.57
English	0.78	0.67		0.64	0.54	0.51
Mathem.	0.70	0.67	0.64		0.45	0.51
Discrim.	0.66	0.65	0.54	0.45		0.40
Music	0.63	0.57	0.51	0.51	0.40	

The correlations are also in the *Spearman.sav* file (although these are actually average correlations, but let's just ignore that).

a. Study these correlations and formulate some simple laws for them. Do not think too complicated. Good laws are simple.

b. Based on the content of the variables and the correlations, formulate a theory about the number of factors in intelligence. Indicate which variables should load on which factor.
c. Test this theory by a factor analysis → Answer Sheet.

Exercise 8
See Exercise 7.

a. If you have done a confirmatory analysis there, do an explorative analysis now. If you've done an exploratory analysis, now do a confirmatory analysis → Answer Sheet.
b. Check the extent to which the conclusions match the previous conclusions.

Exercise 9
See Exercise 7.

a. Look up Spearman's article and find out how many factors there are according to his theory.
b. Test this theory, provided you have not already done so → Answer Sheet.

Exercise 10
Now we apply factor analysis in a slightly more difficult situation, but the results are simple because they correspond nicely with the theory. The file *Christel.sav* contains part of the data of a student. These are data from students too.

a. View the variable labels of the variables pers1 through pers25. Try to formulate a theory on the number of factors based on the contents of the variables. Indicate which items would load on which factors, i.e. which factor pattern you expect.
b. Test this theory by a factor analysis → Answer Sheet.

Exercise 11
See exercise 10. You may have invented a totally new theory. But a suitable theory already exists, so you had to use it, even though the exercise didn't tell you this. If you have used an existing theory, you can skip this assignment. Otherwise, your punishment is that you have to make the previous assignment again, but now correctly:

a. Formulate a theory that agrees with the psychological literature.
b. Use factor analysis in order to construct appropriate scales → Answer sheet.

Exercise 12
Now we apply factor analysis in a situation where the conclusions are less clear. The file *BPS Jan-Feb 2004 VZ.sav* contains data on clients of nursing homes. The caring staff completed a questionnaire for each client. One part of the questionnaire consists of the BPS (Van Loveren-Huyben et al., 1988). This instrument aims to screen the

clients in a house, in order to understand the extent to which various types of care needs exist. The items of the BPS are the variables zbps01 to zbps41. Initially, the BPS consisted of 33 items. The scales were *Cognition*, *Mood* and *Contacts*. Over the years, however, items were added for various substantive reasons. Investigate whether, if you take all 41 items into account, there is reason to distinguish more than three scales. Determine which scales you would eventually distinguish, and which items they contain. Motivate your choice by comparing the results of multiple factor analyses. Write this up as a clear comprehensible argument. Provide a clear overview of the relevant elements of the output. By the way, have you read in this exercise that the data in the file are clean? If you do not know what clean data are, look that up. Investigate with procedure Frequencies whether there are strange scores that require your action.

Exercise 13
The file *Judith.sav* contains part of the data of a student. Consider the variables work1 through work18. Use factor analysis to construct suitable scales. If you do multiple analyses, report the ones on which you base the scales. → Answer Sheet (is difficult to fill in some places).

Exercise 14
Lifestyle.sav The file contains part of the data in an investigation of elderly who are living independently (not in a nursing home). Module 3 of the questionnaire is about lifestyle. These are the variables v29_0 to v29_65. Use factor analysis to construct suitable scales.

Exercise 15
The *Simms.sav* file contains the correlation matrix reported by Simms (2007) as to a number of scales for psychological well-being. Consider to what extent it is possible to explain the correlations between these scales from one factor 'psychological well-being'.

Exercise 16
The *Nivel.sav* file contains the correlation matrix reported by the Nivel institute for 14 scales that were used to evaluate home care by clients. Consider to what extent it is possible to reduce these 14 scales to a small number of factors.

Exercise 17
The *Raven.sav* file contains data from several hundreds of children at Raven's Standard Progressive Matrices (Monks et al, 1986;. Also see Van der Ven and Ellis, 2000). First, find on the Internet or in a book what kind of test this is, so you get an idea of what you're dealing with. Next, investigate with these data whether the test is unidimensional.

Exercise 18
The *Antoine.sav* file includes MMPI subtest scores. Investigate the factorial structure of the main clinical scales.

Exercise 19
The file *Gerlinde.sav* contains data including the SCL-90 and POMS. What these two tests aim to measure is something you can look up yourself, right? Cause I often heard people saying that they don't want to learn something because they can look it up. Go ahead. While you're at it, think a moment about why you would want to look for something if you can also learn it.

a. Investigate the factorial structure of the subscales of these two tests.
b. Describe which problems you encountered and how you solved them.
c. Describe what is peculiar about the relationship between SCL-90 and POMS.

Exercise 20
The *Tessa.sav* file includes data on the WMS-R (Wechsler Memory Scale Revised) in epilepsy patients. Investigate the factorial structure of the subtests. In particular, examine the extent to which there is evidence for the hypothesis that there is a 'general memory', analogous to the concept of 'general intelligence'.

Exercise 21
The *Annicka.sav* file contains data about an investigation into quitting among smokers.

a. Investigate the factorial structure of items v33 through v52 with the aim of constructing (sub) scales.
b. Describe which problems you encountered and how you solved them.

Exercise 22
Watson and Tellegen (1985, p. 221) show Figure 7-2 on the relationships between various self-ratings of moods. The figure is based on factor analysis.

a. Watson and Tellegen explain the figure with the following text. Fill in the missing words. You must of course fill them in on the basis of your knowledge of factor analysis. To check if it is correct, you can look up the article.

> The figure can be interpreted in the following manner. Terms of the same octant are highly ... (1) ... correlated, whereas those in adjacent octants are ... (2) ... correlated. Words 90^0 apart are essentially ... (3) ... to one another, whereas those 180^0 apart are ... (4) ... in meaning and highly ... (5) ... correlated.

b. The factors (axes) of two possible rotations are also shown in the figure. In the so-called unrotated solution of a factor analysis the first principal factor or principal component is the one with the greatest eigenvalue. Which factor is the first principal factor in this figure?

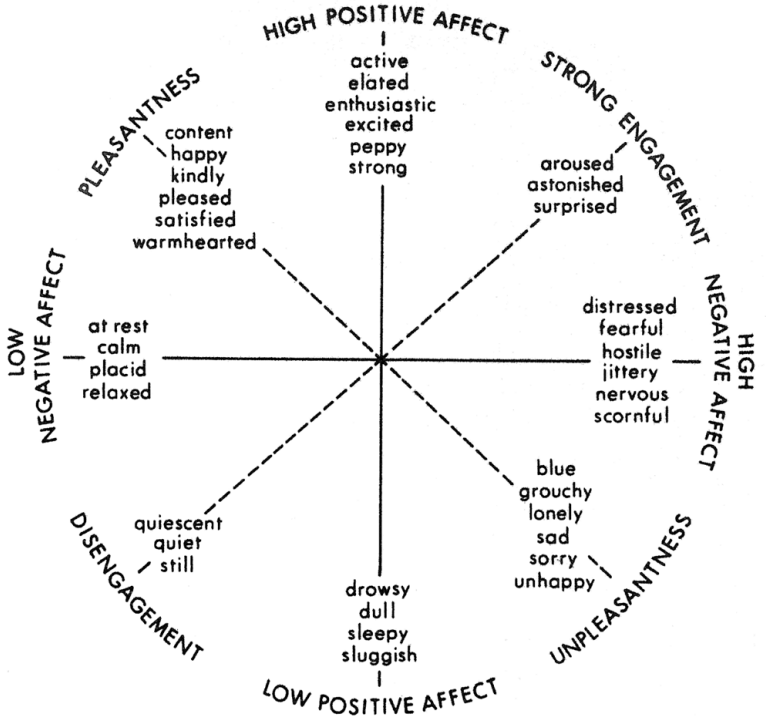

Figure 7.2 (Copyright © 1985 by the American Psychological Association. (Reproduced with permission. The official citation that should be used in referencing this material is Watson, D., & Tellegen, A., 1985, 'Toward a consensual structure of mood', Psychological Bulletin, 98, 219-235. The use of APA information does not imply endorsement by APA)

Exercise 23

Izard et al. (1993) used the Differential Emotions Scale (DES-IV) in 88 women for a period of three years after the birth of their child. This instrument consists of 12 subscales of about three items each. The correlation matrix of the subscale scores is given in the *DES.sav* file. Investigate to what extent the correlations of the subscales can be explained by a small number of factors.

Exercise 24

Church et al. (1999) had about two hundred students rating their mood of that day on a large number of adjectives. A second group of students had to do the same for their mood of the past week. A factor analysis was done in both groups. Since the factor patterns of the two groups were similar to each other, they were given in the same table; see Figures 7.3 and 7.4. Oops, that's where my cup of coffee spills over the factor names. Can you please tell me again what the names were?

Table 3
Rotated Factor Matrix for NA Words: Seven-Factor Solution in Combined Mood-Today and Mood–Past Week Sample

Filipino	English Translation							
Malungkot	Sad	**.67**	.24	.25	.07	.18	.07	.14
Nalulumbay	Sad	**.62**	.16	.26	.16	–.01	.08	.25
Mapanglaw	Lonely	**.62**	.23	.08	.19	.13	.12	.03
Malumbay	Lonely	**.59**	.21	.27	.02	.13	.08	.03
Namamanglaw	Sad	**.56**	.14	.23	.21	.05	.11	.10
Matamlay	Lifeless	**.56**	.21	.19	.08	.16	.16	.21
Luhaan	Weeping	**.53**	.19	.10	**.30**	.18	.03	.01
Masaklap	Bitter	**.52**	.20	.11	**.39**	.12	.13	–.06
Nag-iisa	Alone	**.49**	–.04	.13	.18	.10	.16	.13
Luksa	Mourning	**.49**	.15	.01	**.35**	.14	.04	–.07
Masakit	Painful	**.42**	.20	.17	**.31**	.12	.10	.03
Miserable	Miserable	**.41**	.17	.13	.15	**.36**	.09	.02
Di-maligaya	Unhappy	**.39**	.10	.04	.08	**.39**	.22	.07
Mabigat ang katawan	Sluggish	**.36**	.19	.17	.05	.16	.08	.16
Bigo	Disappointed	**.35**	.23	.12	.19	.27	.14	.07
Ligalig	Troubled	**.34**	.27	.18	.22	**.31**	.08	–.08
Tuliro	Stupefied	**.33**	.08	.25	.26	.24	.06	.12
Galit	Angry	.29	**.60**	.10	.27	.18	.05	–.01
Bugnot	Irritated	.16	**.60**	.13	.11	.18	.05	.10
Inis	Irritated	.18	**.60**	.18	.03	.27	.05	.27
Mainit ang ulo	Hot-headed	.21	**.59**	.21	.17	.12	.07	.14
Buwisit	Annoyed	.15	**.56**	.10	.20	.27	.04	.18
Yamot	Irritated	.27	**.55**	.22	.08	.06	.10	.22
Suya	Fed-up	.21	**.46**	.11	.26	–.01	.11	.29
Nasusuya	Disgusted	.21	**.46**	.11	.26	–.05	.10	**.36**
Kunsumido	Exasperated	**.30**	**.45**	.12	.14	.28	.09	–.05
Gigil	Furious	.04	**.42**	.14	.27	.26	.09	.02
Muhi	Disgusted	**.39**	**.41**	.01	.20	.01	.01	.08
Bagot	Weary	.15	**.40**	–.00	.02	.29	.06	.15
Desmayado	Dismayed	**.30**	**.38**	.14	.03	**.36**	.05	.08
Aburido	Disgusted	.11	**.38**	.18	–.03	**.32**	.09	–.00
Alboroto	Complaining	.16	**.35**	.04	.26	.29	.10	–.16
Asiwa	Awkward	.09	**.35**	.02	.20	.25	.06	–.00
Awang-awa	Commiserating	.04	**.30**	.25	.24	.20	.06	–.14
Kabado	Nervous	.17	.14	**.73**	.02	.27	.03	–.07
May-nerbiyos	Nervous	.18	.09	**.72**	.03	.22	.06	.02
Kinakabahan	Worried	.18	.19	**.68**	.05	.24	.08	–.04
Kakaba-kaba	Anxious	.22	.19	**.68**	.04	.27	.00	–.03
May-kaba	Apprehensive	.24	.10	**.67**	.07	.26	.00	.02
Takot	Afraid	.28	.09	**.52**	.27	.07	.08	.13
Nag-aalaala	Worried	.16	.16	**.49**	.04	.24	.03	.19
Nahihiya	Ashamed	.02	.03	**.47**	.11	.07	.10	.24
Taranta	Confused	.21	.08	**.42**	**.37**	.26	.08	.20
Nakukunsiyensiya	Guilty	.13	.08	**.35**	**.30**	.09	.06	.22
Naninibago	Unaccustomed	.14	.01	**.34**	.19	.15	.09	.28
Nagtataka	Puzzled	.09	.18	**.34**	**.32**	.12	.07	.20
Nanghihinayang	Regretful	.13	.16	**.34**	.16	.16	.09	**.31**
Nagsisisi	Repenting	.22	.27	**.32**	.27	.09	.08	.23
Namamahinga	Resting	.02	.19	.27	.08	–.11	.12	.09

Figure 7.3 (Copyright © 1999 by Blackwell Publishers. Reproduced with permission from Church et al. (1999))

Table 3 (cont.)

Nanglalait	Contemptuous	.09	.05	–.10	**.63**	.16	.06	.18
Nanghahamak	Contemptuous	.14	.07	–.12	**.58**	.14	.10	.07
Nagulat	Surprised	.11	.24	**.34**	**.52**	–.03	.15	.08
Nangingilabot	Goose flesh	.22	.20	.24	**.50**	.03	.09	–.01
Nagulantang	Startled	.12	.27	.17	**.49**	.09	.11	.02
Nauulol	Feeling Crazy	.21	–.05	–.04	**.45**	.21	.04	.23
Nag-aalab	Ardent	.19	.12	.18	**.45**	.13	.03	.00
Mangha	Amazed	.08	.21	.12	**.44**	.03	.09	–.06
Gimbal	Stunned	.17	.22	.19	**.41**	.14	.12	–.28
Nahahabag	Pitiful	.24	.26	**.32**	**.40**	.05	.10	–.02
Manhid	Numb	**.36**	.11	.08	**.40**	.09	.15	.09
Poot	Indignant	**.36**	**.31**	.11	**.36**	.04	.12	–.08
Nangingiba	Unfamiliar	.09	.05	.25	**.32**	.12	.16	.19
Nalulula	Dizzy	.24	.19	.25	**.31**	.09	.06	.14
Di-mapalagay	Uneasy	.22	.15	.20	.11	**.60**	.11	.12
Di-mapakali	Restless	.20	.11	.17	.10	**.59**	.11	.11
Alinlangan	Uncertain	.02	.10	.16	.11	**.50**	–.01	.18
Deskontentado	Discontented	.18	.23	.09	.01	**.47**	.06	.15
Alanganin	Reluctant	.07	.01	.13	.07	**.46**	–.05	.16
Balisa	Disturbed	.25	.24	.12	.11	**.43**	.11	.02
Duda	Doubtful	.08	**.31**	.17	.20	**.40**	.10	.12
Bagabag	Worried	.18	.20	.22	.15	**.39**	.10	.03
Diskompiyado	Distrustful	.14	.29	.13	.13	**.38**	.08	.06
Hibang	Delirious	.20	–.04	.06	**.36**	**.36**	.04	.05
Atubili	Hesitant	–.10	.14	.12	.12	**.34**	.05	.03
Lito	Confused	.29	.28	.28	.10	**.34**	.06	.23
Hapis	Sorrowful	**.31**	.16	.20	.25	**.32**	.11	–.19
Desperado	Desperate	.15	.22	.14	.19	.27	.13	–.07
Alarmado	Alarmed	.01	.13	.14	.03	.26	.02	–.21
Walang-pakiramdam	Insensitive	.12	.06	.10	.14	.06	**.72**	–.05
Walang-pag-ibig	Loveless	.10	.09	.02	–.02	.07	**.68**	.03
Walang-pag-asa	Hopeless	.12	.08	.04	.10	.13	**.65**	.03
Walang-hangad	Undesirous	–.03	.00	.08	.04	.06	**.63**	.03
Walang-panatag	Restless	.12	.11	.17	.16	.01	**.63**	–.07
Walang-inspirasyon	Uninspired	.12	.04	–.04	.00	.09	**.62**	.02
Walang-gusto	Uninterested	–.06	.04	.10	.09	–.07	**.62**	.04
Walang-tatag	Unstable	.11	.05	.04	.08	.01	**.62**	.05
Walang-damdamin	Indifferent	.01	.05	–.00	.19	.14	**.58**	.01
Walang-sigla	Lifeless	.25	.07	.09	.03	.05	**.57**	.16
Walang-kibo	Still	.06	.03	.01	.03	.03	**.48**	.10
Pagod	Tired	.06	.18	.22	–.03	.15	.13	**.52**
Nag-aantok	Sleepy	.05	.24	.14	.01	.16	.02	**.48**
Nagsasawa	Satiated	.20	.20	.02	.18	.14	.09	**.48**
Inip	Bored	.06	**.39**	.13	–.01	.24	.04	**.41**
Nalalabuan	Perplexed	.17	.14	**.31**	.22	.16	.11	**.35**
Naiinggit	Envious	.24	–.01	.20	**.32**	.14	.10	**.33**

Note. Factor loadings ≥ |.30| are in bold.

Figure 7.4 (Copyright © 1999 by Blackwell Publishers. Reproduced with permission from Church et al. (1999))

Exercise 25
Make the questions on the subject Visualise Factor Analysis of MC Questions.xls program (For an explanation, see Section 2.17).

Exercise 26
Repeat the reliability analysis of Diesfeldt's data, as stated in the text.

Exercise 27
See Exercise 5. Perform a reliability analysis for the marks in Statistics 1 exam Part B.

Exercise 28
See exercise 12 on the BPS. Perform a reliability analysis for each of the subscales that you constructed with the factor analysis. Assume five factors (not because that is unambiguously better, but to make sure that everyone is doing about the same analysis). Do not forget to clean the data first! You can do that with the syntax below. Both commands must be executed.

```
recode zbps01 to zbps41 (9 = sysmis).
execute.
```

To check, the first two scales are:

- zbps33 zbps40 zbps17 zbps41 zbps15 zbps20 zbps39 zbps10 zbps13 zbps25 zbps08 zbps22 zbps21
- and zbps23 zbps32 zbps07 zbps30 zbps27 zbps05 zbps35 zbps16

Exercise 29
See exercise 17 on the Raven test. Perform a reliability analysis for this test.

Exercise 30
See exercise 20 on the WMS-R. Investigate with reliability analysis to what extent it is reasonable to combine the subtests scores into one total score.

Exercise 31
See exercise 28 about the BPS. Calculate the subscale scores, and then their correlations. Also calculate the disattenuated correlations between the subscales (i.e., the true-score correlations).

Exercise 32
See exercise 28 on the BPS. Calculate the standard errors of measurement for each of the subscales.

Exercise 33
Open the program with multiple choice questions and complete the questions about reliability.

Exercise 34

a. Explain why factor analysis is not strictly suitable for item analysis.
b. Explain the assumptions of the Rasch model.
c. Describe the characteristics of the IRFs in the Rasch model
d. Name the estimation methods you know for the Rasch model. Describe the main differences between these estimation methods (what do they estimate and under which assumption).
e. Describe which model violations the three R-tests are most sensitive to.

Exercise 35

Table 7.2 shows a number of possible outcomes of the R-tests. Indicate in which cases the scale is unidimensional.

Table 7.2

Test	R_0	R_1	R_2	*Decision*
a	p = 0.01	p = 0.01	p = 0.01	
B	p = 0.01	p = 0.2	p = 0.2	
C	p = 0.1	p = 0.01	p = 0.2	
D	p = 0.1	p = 0.1	p = 0.02	
E	p = 0.2	p = 0.3	p = 0.1	

Exercise 36

In Figure 7-5 are three IRFs of the Rasch model. What are the difficulties of these items?

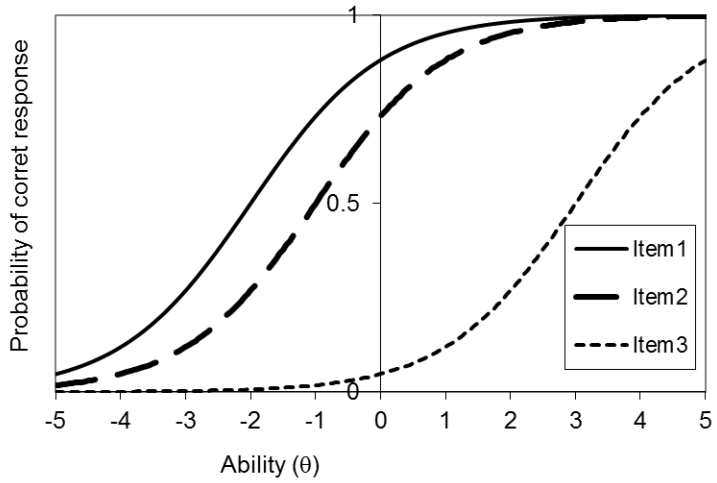

Figure 7.5

Exercise 37
In Figure 7.6, five IRFs of the Rasch model are given. Draw the test response function as accurately as possible without calculating. Where does the function run almost horizontally? Draw schematically how the standard error of measurement depends on the abilities of the subjects (this might not be described exactly in the text). Verify your outcomes with the Excel program *IRT plots.xlsm* by using the buttons Easier and Harder.

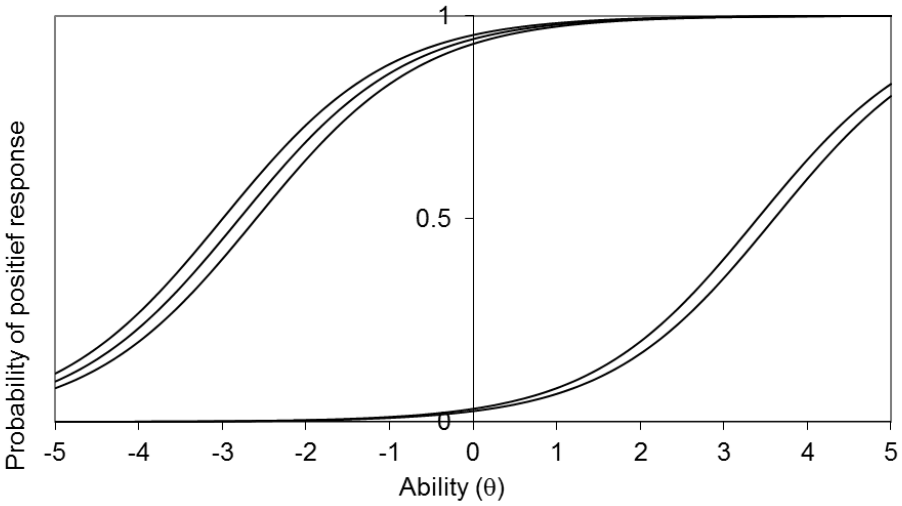

Figure 7.6

Exercise 38
The Excel program *IRT plots.xlsm* allows you to try the effect of the item parameters in the Test Response Function and the standard error of measurement. Play and observe.

Reset all items. Consider what happens to the standard error of measurement when you give item 3 a very large discrimination parameter. The same if you also give item 5 a large discrimination parameter.

Exercise 39
Explain the differences between the 1PL, 2PL, and 3PL model. Which model is preferable for multiple choice questions?

Exercise 40
Open the program with multiple choice questions and complete the questions about IRT.

Exercise 41

a. State the assumptions of monotone homogeneity and explain them.
b. Draw an example of three unequal IRFs that comply with monotone homogeneity.
c. Which of kind of correlations are the target of a prediction by the Mokken model, and what is the content of this prediction?
d. What kind of correlations are being used in an analysis with MSP?
e. Explain what manifest monotonicity is.
f. Describe the three types of H coefficients in an analysis of monotone homogeneity.
g. Describe the decision rule in an analysis of monotonous homogeneity.

Answer sheet factor analysis

The following questions correspond to the answer format of the file *Answersheet.xls.* A round ❍ indicates that a choice must be made. A square ❑ indicates that a yes or no answer has to be given, wherein yes may be chosen more than once. Square brackets with gray markings [...] indicate an open question; something needs to be written there.

Given the research question, what kind of factor analysis would you do here:

- ❍ explorative ❍ confirmatory

Given the research question and chosen kind of analysis, which choice do you make during the analysis:

- Extraction Method: ❍ PCA ❍ ML
- Factors: ❍ minimum eigenvalue [bound] ❍ number of factors [how many]
- Rotation: ❍ Varimax ❍ Promax

Given the research question and chosen kind of analysis and output, on what criteria you base the decision

- Eigenvalues ❑
- *p*-value and RMSEA ❑
- Interpretability of the factor pattern ❑

Given the research question and output and the decision criterion, what is your decision:

- ❍ More factors are required.
- ❑ Maybe more factors are needed.
- ❍ The number of factors is all right.
- ❑ Maybe fewer factors are needed.
- ❍ Fewer factors are required.

Given the research question and output and decision:

- Give an interpretation of the factors [name each factor]
- Indicate which factors are not interpretable [numbers of the factors]
- Specify which items should be removed [numbers of the items]
- Indicate which scales are to be made [per scale the item numbers]

Given the research question, output, decision and interpretation:

- Is a new factor analysis needed? ❑

Given the research question and output:

- Write a concise report [text]

References

(References are of the entire book, not just of the translated chapters)

Abdi, H. (2003). Factor rotations in factor analysis. In M. Lewis-Beck, A. Bryman & T. Futing (Eds.), *Encyclopedia of social sciences research methods*. Thousand Oaks, CA: Sage.

ACT (2008). *2007 ACT national profile report.* Retrieved from http://www.act.org/news/data/07/data.html

Allison, D. B., Allison, R. L., Faith, M. S., Paultre, F., & Pi-Sunyer, F. X. (1997). Power and money: Designing statistically powerful studies while minimizing financial costs. *Psychological Methods, 2*, 20-33.

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika, 38,* 123-140.

Andersen, E. B. (1980). *Discrete statistical models with social science applications*. Amsterdam: North-Holland.

Barrett, P. (2007). Structural equation modelling: Adjudging model fit. *Personality and Individual Differences, 42*, 815-824.

Bartholomew, D. J. (1980). Factor analysis for categorical data. *Journal of the Royal Statistical Society, Series B, 42,* 293-321.

Bartolucci, F., & Forcina, A. (2005). Likelihood inference on the underlying structure of IRT models. *Psychometrika, 70*, 31-43.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*, 238-246.

Bentler, P. M. (2007). On tests and indices for evaluating structural models. *Personality and Individual Differences, 42,* 825-829.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*, 588-606.

Bernstein, I. H., & Teng, G. (1989). Factoring items and factoring scales are different: Spurious evidence for multidimensionality due to item categorization. *Psychological Bulletin, 105*, 467-477.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores* (pp. 395-479). Reading: Addison-Wesley.

Bolt, D. M. (2001). Conditional covariance-based representation of multidimensional test structure. *Applied Psychological Measurement, 25,* 244-257.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika, 71*, 425-440.

Borsboom, D., & Mellenbergh, G. J. (2002). True scores, latent variables, and constructs: a comment on Schmidt and Hunter. *Intelligence, 30*, 505-514.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 3*, 296-322.

Browne, M. W. (1974). Generalized least squares estimators in the analysis of covariance structures. *South African Statistical Journal, 8,* 1-24.

Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology, 37,* 62-83.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Beverly Hills, CA: Sage.

Carroll, J. B. (1961). The nature of the data, or how to choose a correlation coefficient. *Psychometrika, 26,* 347-372.

Cattell, R. B. (1950). *Personality: A systematic theoretical and factual study*. New York: McGraw-Hill.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1***,** 245-276.

Cattell, R.B. (1971). *Abilities: their structure, growth, and action*. Boston: Houghton Mifflin.

Cattell, R. B., Eber, H. W., & Tatsuoka, M. M. (1970). *Handbook for the Sixteen Personality Factor questionnaire.* Champaign, IL: Institute for Personality and Ability Testing.

Charles, E. P. (2005). The correction for attenuation due to measurement error: clarifying concepts and creating confidence sets. *Psychological Methods, 10,* 206-226.

Charter, R. A., & Feldt, L. S. (2001). Meaning of reliability in terms of correct and incorrect clinical decisions: The art of decision making is still alive, *Journal of Clinical and Experimental Neuropsychology, 23*, 530-537.

Christoffersson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika, 40,* 5-32.

Church, A. T., Katigbak, M. S., Reyes, J. A. S., & Jensen, S. M. (1999). The structure of affect in a non-Western culture: evidence for cross-cultural comparability. *Journal of Personality, 67,* 505-534.

Clark, L. A., & Watson, D. (1995). Constructing validity: basic issues in objective scale development. *Psychological Assessment, 7,* 309-319.

Cohen, J. (1988*). Statistical power analysis for the behavioral sciences (2nd ed.)*. Hillsdale, NJ: Lawrence Erlbaum.

College Board (2007). *SAT percentile ranks for males, females, and total group.* Retrieved from www.collegeboard.com/prod_downloads/highered/ra/sat/SATPercentileRanksCompositeCR_M_W.pdf

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78,* 98-104.

Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation, 10(7)*. Retrieved from pareonline.net/getvn.asp?v=10&n=7

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334.
Cronbach, L. J. (1988). Internal consistency of tests: analyses old and new. *Psychometrika, 53*, 63-70.
Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioural measurements: Theory of generalizability for scores and profiles.* New York: Wiley.
Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology, 16,* 137-163.
Davenport, E. C., Jr., & El-Sanhurry, N. A. (1991). Phi/Phimax: review and synthesis. *Educational and Psychological Measurement, 51*, 821-828.
De Gooijer, J. G., & Yuan, A. (2011). Some exact tests for manifest properties of latent trait models. *Computational Statistics and Data Analysis, 55*, 34-44.
De Raad, B., & Barelds, D. P. H. (2008). A new taxonomy of Dutch personality traits based on a comprehensive and unrestricted list of descriptors. *Journal of Personality and Social Psychology, 94*, 347-364.
Diesfeldt, H. F. A. (1997). De depressielijst. Een instrument voor stemmingsonderzoek in de psychogeriatrie. *Tijdschrift voor Gerontologie en Geriatrie, 28*, 113-118.
Diesfeldt, H. F. A. (2004). De Depressielijst voor stemmingsonderzoek in de psychogeriatrie: meetpretenties en schaalbaarheid. *Tijdschrift voor Gerontologie en Geriatrie, 35*, 224-233.
Digby, P. G. N. (1983). Approximating the tetrachoric correlation coefficient. *Biometrics, 39*, 753-757.
Drasgow, F. (1988). Polychoric and polyserial correlations. In L. Kotz & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences: Vol. 7* (pp. 69-74). New York: Wiley.
Drenth, P. J. D., & Sijtsma, K. (2006). *Testtheorie. Inleiding in de theorie van de psychologische test en zijn toepassingen.* Houten: Bohn Stafleu van Loghum.
Dykstra, R. (1991). Asymptotic normality for chi-bar-square distributions. *Canadian Journal of Statistics, 19*, 297-306.
Ellis, J. L. (1993). Subpopulation invariance of patterns in covariance matrices. *British Journal of Mathematical and Statistical Psychology, 46*, 231-254.
Ellis, J. L. (2003a). *Statistiek voor de psychologie, deel 3*. Amsterdam: Boom.
Ellis, J. L. (2003b). *Statistiek voor de psychologie, deel 4*. Amsterdam: Boom.
Ellis, J. L. (2004). *Statistiek voor de psychologie, deel 2*. Amsterdam: Boom.
Ellis, J. L. (2013a). An inequality for correlations in unidimensional monotone latent variable models for binary variables. *Psychometrika*, advance online publication. DOI: 10.1007/S11336-013-9341-5.
Ellis, J. L. (2013b). A standard for test reliability in group research. *Behavior Research Methods, 45*, 16-24. DOI 10.3758/s13428-012-0223-z
Ellis, J. L., & Junker, B.W. (1997). Tail-measurability in monotone latent variable models. *Psychometrika, 62*, 495-523.
Eysenck, H. J., & Eysenck, M. W. (1985). *Personality and individual differences: a natural science approach.* London: Plenum 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,* 272-299.

Fan, X., & Sivo, S. A. (2005). Sensitivity of fit indexes to misspecified structural or measurement model components: rationale of two-index strategy revisited. *Structural Equation Modeling, 12,* 343-367.

Fan, X., & Sivo, S. A. (2007). Sensitivity of fit indices to model misspecification and model types. *Multivariate Behavioral Research, 42,* 509-529.

Feldt, L. S. (1965). The approximate sampling distribution of Kuder-Richardson reliability coefficient twenty. *Psychometrika, 30,* 357-370.

Feldt, L. S. (2011). Estimating the effect of changes in criterion score reliability on the power of the F test of equality of means. *Educational and Psychological Measurement, 71,* 420-430.

Finch, J. F., & West, S. G. (1997). The investigation of personality structure: statistical models. *Journal of Research in Personality, 31,* 439-485.

Fischer, G. H. (1974). *Einführung in die Theorie psychologischer Tests.* Bern: Huber.

Fischer, G. H. (1995). Derivations of the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models. Foundations, recent developments, and applications* (pp. 15-38). New York: Springer Verlag.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods, 9*, 466-491.

Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment, 7,* 286-299.

Forrest, S., Lewis, C. A., & Shevlin, M. (2000). Examining the factor structure and differential functioning of the Eysenck personality questionnaire revised – abbreviated. *Personality and Individual Differences, 29*, 579-588.

Frisbie, D. A. (1988). Reliability of scores from teacher-made tests. *Educational Measurement: Issues and Practice, 7,* 25-35.

Ghosh, M. (1995). Inconsistent maximum likelihood estimators for the Rasch model. *Statistical and Probability Letters, 23*, 165-170.

Glas, C. A. W. (1988). The derivation of some tests for the Rasch model from the multinomial distribution. *Psychometrika, 53,* 525-546.

Glas, C. A. W., & Ellis, J. L. (1993). *Rasch scaling program*. Groningen: iec ProGAMMA.

Glas, C. A. W., & Verhelst, N. D. (1995). Testing the Rasch model. In G. H. Fischer, & I. W. Molenaar (Eds.), *Rasch models. Foundations, recent developments, and applications* (pp. 69-96). New York: Springer Verlag.

Goffin, R. D. (2007). Assessing the adequacy of structural equation models: Golden rules and editorial policies. *Personality and Individual Differences, 42*, 831-839.

Green, S. B., Lissitz, R. W., & Mulaik, S. A. (1977). Limitations of coefficient alpha as an index of test unidimensionality. *Educational and Psychological Measurement, 37*, 827-838.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika, 10*, 255-282.
Guttman, L. (1954). Some necessary conditions for common factor analysis. *Psychometrika, 19*, 149-161.
Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9*, 139-164.
Hayduk, L., Cummings, G., Boadu, K., Pazderka-Robinson, H., & Boulianne, S. (2007). Testing! testing! one, two, three – Testing the theory in structural equation models!. *Personality and Individual Differences, 42*, 841-850.
Heene, M., Hilbert, S., Freudenthaler, H. H., & Bühner, M. (2012). Sensitivity of SEM fit indexes with respect to violations of uncorrelated errors. *Structural Equation Modeling, 19*, 36-50.
Hendrickson, A. E., & White, P. O. (1964). Promax: A quick method for rotation to oblique simple structure. *British Journal of Mathematical and Statistical Psychology, 17*, 65-70.
Henson, R. K., & Roberts, J. K. (2006). Use of exploratory factor analysis in published research: common errors and some comment on improved practice. *Educational and Psychological Measurement, 66*, 393-416.
Holland, P. W., & Rosenbaum, P. R. (1986). Conditional association and unidimensionality in monotone latent variable models. *Annals of Statistics, 14*, 1523-1543.
Holzinger, K. J. (1924). On scoring multiple-response tests. *Journal of Educational Psychology, 15*, 445-447.
Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*, 179-185.
Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology, 24*, 417-441, 498-520.
Hotelling, H. (1936). Simplified calculation of principal components. *Psychometrika, 1*, 27-35.
Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55.
Hu, L., Bentler, P. M., & Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted? *Psychological Bulletin, 122*, 351-362.
Hutchinson, T. P. (1993). Kappa muddles together two sources of disagreement: tetrachoric correlation is better. *Research in Nursing and Health, 16*, 313-315.
Iacobucci, D., & Duhachek, A. (2003). Advancing alpha: measuring reliability with confidence. *Journal of Consumer Psychology, 13*, 478-487.
Ip, E. H. (2001). Testing for local dependency in dichotomous and polytomous item response models. *Psychometrika, 66*, 109-132.
Izard, C. E., Libero, D. Z., Putnam, P., & Haynes, O. M. (1993). Stability of emotion experiences and their relations to traits of personality. *Journal of Personality and Social Psychology, 64*, 847-860.

Jackson, D. L., Gillaspy, J. A., & Purc-Stephenson, R. (2009). Reporting practices in confirmatory factor analysis: an overview and some recommendations. *Psychological Methods, 14*, 6-23.

Jackson, P. H., & Agunwamba, C. C. (1977). Lower bounds for the reliability of the total score on a test composed of nonhomogeneous items: I: Algebraic lower bounds. *Psychometrika, 42*, 567-578.

Jöreskog, K. G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika, 32*, 443-482.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika, 34*, 183-202.

Jöreskog, K. G. (1990). New Developments in LISREL: Analysis of ordinal variables using polychoric correlations and weighted least squares. *Quality and Quantity, 24*, 387-404.

Jöreskog, K. G. (1994). On the estimation of polychoric correlations and their asymptotic covariance matrix. *Psychometrika, 59*, 381-389.

Jöreskog, K. G. (2003). *Factor analysis by MINRES*. In www.ssicentral.com/lisrel/techdocs.minres.pdf.

Jöreskog, K. G., & Lawley, D. N. (1968). New methods in maximum likelihood factor analysis. *British Journal of Mathematical and Statistical Psychology, 21*, 85-96.

Jöreskog, K. G., & Sörbom, D. (1996). *LISREL 8: User's reference guide*. Chicago, IL: Scientific Software International.

Jöreskog, K. G., & Sörbom, D. (1999). *LISREL 8.30 and PRELIS 2.30*. Chicago, IL: Scientific Software International.

Jöreskog, K. G., & Sörbom, D. (2006). *LISREL 8.80* [software]. Chicago, IL: Scientific Software International.

Junker, B. W. (1993). Conditional association, essential independence and monotone unidimensional item response models. *Annals of Statistics, 21*, 1359-1378.

Junker, B. W., & Ellis, J. L. (1997). A characterization of monotone unidimensional latent variable models. *Annals of Statistics, 25*, 1327-1343.

Junker, B. W., & Sijtsma, K. (2000). Latent and manifest monotonicity in item response models. *Applied Psychological Measurement, 24*, 65-81.

Junker, B. W., & Sijtsma, K. (2001a). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*, 258-272.

Junker, B. W., & Sijtsma, K. (2001b). Nonparametric item response theory in action: An overview of the special issue. *Applied Psychological Measurement, 25*, 211-220.

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika, 23*, 187-200.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement, 20*, 141-151.

Kaiser, H. F. (1961). An note on Guttman's lower bound for the number of common factors. *British Journal of Statistical Psychology, 14*, 1.

Kelley, T. L. (1928). *Crossroads in the mind of man: A Study of Differentiable Mental Abilities.* Stanford: Stanford University Press.

Kenny, D. A. (2015). *Measuring model fit*. http://davidakenny.net/cm/fit.htm
Kistner, E. O., & Muller, K. E. (2004). Exact distributions of intraclass correlation and Cronbach's alpha with Gaussian data and general covariance. *Psychometrika, 69*, 459-474.
Knol, D. L., & Berger, M. P. (1991). Empirical comparison between factor analysis and multidimensional item response models. *Multivariate Behavioral Research, 46*, 457-477.
Kolenikov, S., & Bollen, K. A. (2012). Testing negative error variances: Is a Heywood case a symptom of misspecification? *Sociological Methods & Research, 41*, 124-167.
Kristof, W. (1963). The statistical theory of stepped-up reliability when a test has been divided into several equivalent parts. *Psychometrika, 28*, 221-238.
Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*, 151-160.
Lambert, Z. V., Wildt, A. R., & Durand, R. M. (1991). Approximating confidence intervals for factor loadings. *Multivariate Behavioral Research, 26*, 421-434.
Lawley, D. N. (1940). The estimation of factor loadings by the method of maximum likelihood. *Proceedings of the Royal Society of Edinburgh, 60*, 64-82.
Loevinger, J. (1948). The technique of homogeneous tests compared with some aspects of 'scale analysis' and factor analysis. *Psychological Bulletin, 45*, 507-530.
Lord, F. M. (1980). *Application of item response theory to practical testing problems.* Hillsdale NJ: Erlbaum.
Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison Wesley.
MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modelling. *Psychological Methods, 1*, 130-149.
Markland, D. (2007). The golden rule is that there are no golden rules: A commentary on Paul Barrett's recommendations for reporting model fit in structural equation modelling. *Personality and Individual Differences, 42*, 851-858.
Maydeu-Olivares, A., Coffman, D. L., & Hartmann, W. M. (2007). Asymptotically distribution-free (ADF) interval estimation of coefficient alpha. *Psychological Methods, 12*, 157-176.
Maydeu-Olivares, A., & Joe, H. (2005). Limited and full information estimation and goodness-of-fit testing in 2n tables: a unified approach. *Journal of the American Statistical Association, 100*, 1009-1020.
Maydeu-Olivares, A., & Montaño, R. (2012). How should we assess the fit of Rasch-type models? Approximating the power of goodness-of-fit statistics in categorical data analysis. *Psychometrika*. DOI: 10.1007/S11336-012-9293-1
McDonald, R. P. (1967). *Nonlinear factor analysis* (Psychometric Monograph No. 15). Richmond, VA: Psychometric Corporation.
McDonald, R. P., & Ahlawat, K. S. (1974). Difficulty factors in binary data. *British Journal of Mathematical and Statistical Psychology, 27*, 82-99.
McDonald, R. P., & Ho, M. H. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods, 7*, 64-82.

McGrae, R. R., & Costa, P. T., Jr. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology, 2*, 81-90.
McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*, 30-46.
McIntosh, C. N. (2007). Rethinking fit assessment in structural equation modelling: A commentary and elaboration on Barrett (2007). *Personality and Individual Differences, 42*, 859-867.
Miecskowski, T. A., Sweeney, J. A., Haas, G., Junker, B. W., Brown, R. P., & Mann, J. J. (1993). Factor composition of the Suicide Intent Scale. *Suicide and Life Threatening Behavior, 23*, 37-45.
Miles, J., & Shevlin, M. (2007). A time and a place for incremental fit indices. *Personality and Individual Differences, 42*, 869-874.
Millsap, R. E. (1997). Invariance in measurement and prediction: Their relationship in the single-factor case. *Psychological Methods, 2*, 248-260.
Millsap, R. E. (2007a). Invariance in measurement and prediction revisited. *Psychometrika, 72*, 461-473.
Millsap, R. E. (2007b). Structural equation modeling made difficult. *Personality and Individual Differences, 42*, 875-881.
Mokken, R. J. (1971). *A theory and procedure of scale analysis*. The Hague: Mouton.
Mokken, R. J. (1997). Nonparametric models for dichotomous responses. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 351-368). New York: Springer.
Mokken, R. J., & Lewis, C. (1982). A nonparametric approach to the analysis of dichotomous item responses. *Applied Psychological Measurement, 6*, 417-430.
Mokken, R. J., Lewis, C., & Sijtsma, K. (1986). Rejoinder to 'The Mokken scale: A critical discussion'. *Applied Psychological Measurement, 10*, 279-285.
Molenaar, I. W. (1991). A weighted Loevinger H-coefficient extending Mokken scaling to multicategory items. *Kwantitatieve Methoden, 37*, 97-117.
Molenaar, I. W. (1997). Nonparametric models for polytomous responses. In: W.J. van der Linden & R.K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 369-380). New York/Berlin: Springer-Verlag.
Molenaar, I. W., & Sijtsma, K. (2000). *User's manual MSP5 for Windows* [software manual]. Groningen, The Netherlands: iec ProGAMMA.
Mönks, F. J., Van Boxtel, H. W., Roelofs, J. J. W., & Sanders, M. P. M. (1986). The identification of gifted children in secondary education and a description of their situation in Holland. In K. A. Heller & J. F. Feldhusen (Eds.), *Identifying and nurturing the gifted* (pp. 39-65). Toronto: Huber.
Mulaik, S. A. (1965). Reliability as the upper limit of a test's communality. *Perceptual and Motor Skills, 20*, 646-648.
Mulaik, S. A. (1972). *The foundations of factor analysis*. New York: McGraw-Hill.
Mulaik, S. (2007). There is a place for approximate fit in structural equation modelling. *Personality and Individual Differences, 42*, 883-891.
Muthén, B. O. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika, 43*, 551-560.

Muthén, B. O. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika, 49*, 115-132.
Novick, M. R., & Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements. *Psychometrika, 32*, 1-13.
Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.
Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.
Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika, 44*, 443-460.
Osburn, H. G. (2000). Coefficient alpha and related internal consistency reliability coefficients. *Psychological Methods, 5*, 343-355.
Parry, C. D., & McArdle, J. J. (1991). An applied comparison of methods for least-squares factor analysis of dichotomous variables. *Applied Psychological Measurement, 15*, 35-46.
Pearson, K. (1900). Mathematical contributions to the theory of evolution: VII. On the correlation of characters not quantitatively measurable. *Royal Society Philosophical Transactions, Series A*, 195, 1-47.
Ponocny, I. (2001). Nonparametric goodness-of-fit tests for the Rasch model. *Psychometrika, 66*, 437-460.
Rajaratnam, N., Cronbach, L. J., & Gleser, G. C. (1965). Generalizability of stratified-parallel tests. *Psychometrika, 30*, 39-56.
Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika, 56*, 611-630.
Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Nielson & Lydiche.
Raykov, T. (1998). On the use of confirmatory factor analysis in personality research. *Personality and Individual Differences, 24*, 291-293.
Raykov, T. (2007). Reliability if deleted, not 'alpha if deleted': Evaluation of scale reliability following component deletion. *British Journal of Mathematical and Statistical Psychology, 60*, 201-216.
Raykov, T. (2008). 'Alpha if deleted' and loss in criterion validity. *British Journal of Mathematical and Statistical Psychology, 61*, 275-285.
Reise, S. P., Waller, N. G., & Comrey, A. L. (2000). Factor analysis and scale revision. *Psychological Assessment, 12*, 287-297.
Revelle, W. (2008). *psych: Procedures for personality and psychological research (R package version 1.0-51).*
Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega and the glb: comments on Sijtsma. *Psychometrika, 74*, 145-154.
Robinson, D. L. (2001). How brain arousal systems determine different temperament types and the major dimensions of personality. *Personality and Individual Differences, 31*, 1233-1259.
Rosenbaum, P. R. (1984). Testing the conditional independence and monotonicity assumptions of item response theory. *Psychometrika, 49*, 425-435.
Roskam, E. E., Van den Wollenberg, A. L., & Jansen, P. G. W. (1986). The Mokken scale: A critical discussion. *Applied Psychological Measurement, 10*, 265-277.

Sardy, S., & Victoria-Peser, M.-P. (2012). Isotone additive latent variable models. *Statistical Computing, 22*, 647-659.

Saris, W. E., Satorra, A., & Van der Veld, W. M. (2009). Testing structural equation models or detection of misspecifications? *Structural Equation Modeling, 16*, 561-582.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8*, 350-353.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74*, 107-120.

Sijtsma, K., & Junker, B. W. (2006). Item response theory: past performance, present developments, and future expectations. *Behaviormetrika, 33,* 74-102.

Sijtsma, K., & Meijer, R. R. (2007). Nonparametric item response theory and special topics. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics, vol. 26: Psychometrics* (pp. 719-746). Amsterdam: Elsevier.

Sijtsma, K., & Prins, P. M. (1986). Itemselectie in het Mokken model. *Tijdschrift voor Onderwijsresearch, 11*, 121-129.

Simms, L. J. (2007). The big seven model of personality and its relevance to personality pathology. *Journal of Personality, 75*, 65-94.

Spearman, C. (1904a). 'General intelligence', objectively determined and measured. *The American Journal of Psychology, 15*, 201-292.

Spearman, C. (1904b). The proof and measurement of association between two things. *The American Journal of Psychology, 15*, 72-101.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3*, 271-295.

Spearman, C. (1927*). The abilities of man*. New York: Macmillan.

Steiger, J. H. (1990). Structural model evaluation and modification: an interval estimation approach. *Multivariate Behavioural Research, 25*, 173-180.

Steiger, J. H. (2007). Understanding the limitations of global fit assessment in structural equation modelling. *Personality and Individual Differences, 42*, 893-898.

Stewart, D., Barnes, J., Cote, J., Cudeck, R., & Malthouse, E. (2001). Factor analysis. *Journal of Consumer Psychology, 10*, 75-82.

Stout, W. F. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika, 55*, 293-325.

Stout, W., Habing, B., Douglas, J., Kim, H., Roussos, L., & Zhang, J. (1996). Conditional covariance-based nonparametric multidimensional assessment. *Applied Psychological Measurement, 20*, 331-354.

Swaminathan, H., Hambleton, R. K., & Algina, J. (1974). Reliability of criterion-referenced tests: A decision-theoretic formulation. *Journal of Educational Measurement, 11*, 263-267.

Ten Berge, J. M. F., & Hofstee, W. K. B. (1999). Coefficients alpha and reliabilities of unrotated and rotated components. *Psychometrika, 64*, 83-90.

Ten Berge, J. M. F., Snijders, T. A. B., & Zegers, F. E. (1981). Computational aspects of the greatest lower bound to the reliability and constrained minimum trace factor analysis. *Psychometrika, 46*, 201-213.

Ten Berge, J. M. F., & Sočan, G. (2004). The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. *Psychometrika, 69*, 613-625.

Ten Berge, J. M. F., & Zegers, F. E. (1978). A series of lower bounds to the reliability of a test. *Psychometrika, 43*, 575-579.

Ten Holt, J. C., Van Duijn, M. A. J., & Boomsma, A. (2010). Scale construction and evaluation in practice: A review of factor analysis versus item response theory applications. *Psychological Test and Assessment Modeling, 52*, 272-297.

Thurstone, L. L. (1931). Multiple factor analysis. *Psychological Review, 38*, 406-427.

Thurstone, L. L. (1938). Primary mental abilities. *Psychometric Monographs*, 1.

Thurstone, L. L. (1947). *Multiple factor analysis*. Chicago: University of Chicago Press.

Thurstone, L. L. (1954). An analytical method for simple structure. *Psychometrika, 19*, 173-182.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38*, 1-10.

Van den Wollenberg, A. L. (1982). Two new test statistics for the Rasch model. *Psychometrika, 47*, 123-140.

Van der Ark, L. A. (2007). Mokken scale analysis in R. *Journal of Statistical Software, 20*, 1-19.

Van der Ark, L.A. (2012). New developments in Mokken scale analysis in R. *Journal of Statistical Software, 48(5)*, 1-27. Retrieved from http://www.jstatsoft.org

Van der Ark, L. A., Croon, M. A., & Sijtsma, K. (2008). Mokken scale analysis for dichotomous items using marginal models. *Psychometrika, 73*, 183-208.

Van der Ven, A. H. G. S., & Ellis, J. L. (2000). A Rasch analysis of Raven's standard progressive matrices. *Journal of Personality and Individual Differences, 29*, 45-64.

Van Loveren-Huyben, C. M. S., Van der Bom, J. A., & Bronts, P. A. J. M. (1988). *Handleiding voor de BPS. Beoordelingsschaal voor Psychische en Sociale problemen in het verzorgingshuis*. Deventer: Van Loghum Slaterus.

Van Onna, M. J. H. (2002). Bayesian estimation and model selection in ordered latent class models for polytomous items. *Psychometrika, 67*, 519-538.

Van Zyl, J. M., Neudecker, H., & Nel, D. G. (2000). On the distribution of the maximum likelihood estimator of Cronbach's alpha. *Psychometrika, 65*, 271-280.

Veldhuijzen, N. H., Goldebeld, P., & Sanders, P. F. (1993). Klassieke testtheorie en generaliseerbaarheidstheorie. In T. J. H. M. Eggen & P. F. Sanders (Eds.), *Psychometrie in de praktijk.* Arnhem: CITO.

Verhelst, N. D. (1993). Itemresponstheorie. In T. J. H. M. Eggen & P. F. Sanders (Eds.), *Psychometrie in de praktijk* (pp. 83-178). Arnhem: Cito.

Vermunt, J. K. (2001). The use of restricted latent class models for defining and testing nonparametric and parametric item response theory models. *Applied Psychological Measurement, 25*, 283-294.

Vernon, T., & Eysenck, S. (2007). Introduction. *Personality and Individual Differences, 42*, 813.

Waller, N. G., Tellegen, A., McDonald, R. P., & Lykken, D. T. (1996). Exploring nonlinear models in personality assessment: Development and preliminary validation of a negative emotionality scale. *Journal of Personality, 64*, 545-576.

Warm, T. A. (1989). Weighted maximum likelihood estimation of ability in item response theory. *Psychometrika, 54*, 427-450.

Watson, D., & Tellegen, A. (1985). Toward a consensual structure of mood. *Psychological Bulletin, 98*, 219-235.

Weiss, D. J. (1995). Polychotomous or polytomous? *Applied Psychological Measurement, 19*, 4.

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: current approaches and future directions. *Psychological Methods, 12*, 58-79.

Woodhouse, B., & Jackson, P. H. (1977). Lower bounds for the reliability of the total score on a test composed of non-homogeneous items: II: A search procedure to locate the greatest lower bound. *Psychometrika, 42*, 579-591.

Yalcin, I., & Amemiya, Y. (2001). Nonlinear factor analysis as a statistical method. *Statistical Science, 16*, 275-294.

Yeomans, K. A., & Golder, P. A. (1982). The Guttman-Kaiser criterion as a predictor of the number of common factors. *The Statistician, 31*, 221-229.

Yuan, A., & Clarke, B. (2001). Manifest characterization and testing for certain latent properties. *Annals of Statistics, 29*, 876-898.

Yuan, K. (2005). Fit indices versus test statistics. *Multivariate Behavioral Research, 40*, 115-148.

Zhang, J. (2007). Conditional covariance theory and detect for polytomous items. *Psychometrika, 72*, 69-91.

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle's β, and Mcdonald's ωH: their relations with each other and two alternative conceptualizations of reliability. *Psychometrika, 70*, 123-133.