Analysis and Construction of Questionnaires.

Course materials for:

424202: MTO-02-Pre-master

424522: MTO-D

424242: Construction and Analysis of Questionnaires

By Dr. Maurits Kaptein

2013-2014

Table of Contents

Concise overview of the course

Lecture	Topic	Details
1	Introduction	Why questionnaires?
2	Association & Correlation	What are correlations?
3	Association & Correlation	Transformations & Combinations
4	Reliability	Model of Classical Test theory Types of Reliability
5	Validity	Evaluating Validity
6	Design and Formulation	Formal methods and importance
7	Design and Formulation	Formulating items
8	Factor Analysis	PCA important concepts
9	Factor Analysis	Example + Number of Factors & Rotation
10	Factor Analysis Confirmatory FA	PAF MGM
11	Confirmatory FA	SEM
12	Cluster Analysis	Ward's method
13	Cluster Analysis	K-means
14	Exam preparation	

Course organization

Here I provide an overview of the main organizational issues regarding MTO-D-MAW / MTO-02-Schakel, and Construction and Analysis of Questionnaires. The content and organization of these three courses is almost identical, with some minor exceptions for the Construction and Analysis of Questionnaires course regarding the bonus assignments. Here I give an overview of the course materials, the organization of the lectures, and the exam.

Books

The only obliged book for this course is:

- Julie Pallant, SPSS Survival Manual, Open University Press, Buckingham U.K., ISBN 0-335-21640-4, 2007 (3e editie).

Besides the SPSS Survival manual I will point you to several – non-obliagtory – articles during the lectures.

Lecture notes

This set of lectures notes is the primary resource for this course. It contains all the theoretical materials you need to know, and it contains the assignments for the tutorials. During the lectures I will cover parts of these lecture notes. *All* the material in these lecture notes will be considered known at the time of examination.

Lectures ("Hoorcolleges")

During the lectures I will discuss the lecture notes, provide additional examples, and give room for discussion of the topics. The lectures will also be used to highlight the links between the different topics covered in the lecture notes.

Note that the lectures are not obligatory: If you understand the material in these lectures notes well you do not need to attend. However, the lectures are strongly recommended: here I will detail the techniques, provide a reference, and be able to answer your questions.

The full course consists of 14 Lectures. On page 5 you can see which topic I will cover during each of the lectures.

Tutorials ("Werkcolleges")

During the tutorials the assignments (see final section of these lectures notes) will be discussed. Also, the tutorials give you the opportunity to ask questions about the lectures.

There will be 8 tutorials during this course. The structure of the tutorials will be as follows:

1. The tutor will check whether or not you have made the assignments.
2. The tutor will discuss the assignments and will give the correct answers. These are also available on BlackBoard
3. The tutor will give you a chance to ask questions.
4. Fifteen minutes before the end of the session the tutor will hand out a small, 8-question, multiple-choice test.

For each tutorial you can obtain a grade: If you have made the assignment you receive 2 points. Next, the small MC test at the end of the tutorial allows you to obtain another 8 points. (These are corrected for guessing: since it's 2-choice questions you need at least 4 correct answers to start gaining points!). With 8 questions correct *and* making the assignments you can gain 10 points.

At the end of the term you will receive an average grade for the tutorials: This will be the average of the best 7 grades obtained during the tutorials.

Note that the tutorials are not obligatory! The grade you obtain in the tutorials can help you, but is not necessary to pass the course (see "Final Grade"). However, the tutorials are – like the lectures – recommended. This is a hard course, and the tutorials allow you to practice the material interactively.

Practical (OBLIGATORY)

This course will also consist of 2 obligatory practical. Hence, these you will have to attend!

For the first practical there will be a homework assignment. This homework assignment will be published on blackboard well in advance. You will have to hand in the homework, and subsequently, during the practical, you will have to do an SPSS assignment.

Again, these are obligatory. Not attending the practical will make you fail the course.

The exam:

The exam will be "closed-book": hence you will not be allowed to bring any of your notes. You will receive a number of formula's on a formula sheet. This sheet will be available on blackboard prior to the exam.

The exam be multiple choice and will consist of 50 2-choice questions. On Blackboard you will find a practice exam. The practice exam will be discussed in the last lecture, and the answers can also be found on Blackboard.

Your final grade will be determined as follows:

$G_{tutorial}$ = Tutorial grade (average of best 7).
G_{exam} = Exam grade.
G_{final} = Your final grade.

If (G_{exam} > $G_{tutorial}$){
  G_{final} = G_{exam}
} else, if ($G_{tutorial}$ > G_{exam}){
  G_{final} = 1/3 * $G_{tutorial}$ + 2/3 * G_{exam}
}

In words: If your exam grade is higher than the tutorial grade, only your exam grade will count. If your tutorial grade (the average of your 7 best tutorials) is higher than the exam grade, the tutorial grade will count for 1/3 for your final grade.

Final grades will be rounded to the nearest half. A 5.5 shall not be given.

Introduction: Why should we care about questionnaires?

More than half of all empirical research uses questionnaires and/or structured interviews. Thus, these are very important ways of obtaining data. Questionnaires are used as a measurement tool both in Scientific research as well as in Market and consumer research and thus it is useful for you to know how to evaluate and design questionnaires. That is what this course is about.

Consider the following statement: "The Dutch are *taller* than the Chinese". To check such a statement one uses a measurement rod to measure a sample of Dutch people and a sample of Chinese people and compare these two samples. The measurement rod exists – we all know how long 1 meter is – so this is relatively easy to do.

Now consider a similar statement: "The Dutch are *happier* than the Chinese". To check this statement we use something to measure happiness both amongst a sample of Dutch people as well as a sample of Chinese people. However, one could wonder *how exactly do we measure happiness?*

Since we cannot use a measurement rod to measure happiness, we have to *ask* people about their happiness. The asking is often done using a questionnaire. Questionnaires are very often used in the social sciences and we use them not only to measure happiness, but many other things: Questionnaires are used to measure attitudes, opinions, feelings, knowledge, behaviors, etc. Since questionnaires are so prevalent in the social sciences you should understand how they are build, and you should be able to assess the validity of conclusions drawn from studies using questionnaires.

During this course you will learn how to design questionnaires, and how to deal with responses to questionnaires. You will learn how to design and formulate questionnaires. Furthermore, you will learn how to formally evaluate questionnaires using *reliability* and *validity*, and you will learn how to summarize the answers to questionnaires using *Factor Analysis* and *Cluster Analysis*.

Understanding and being able to create questionnaires is two-fold: On one hand you will be concerned with statistical and numerical evaluations, while on the other

hand you will be thinking about the actual content of the questionnaire: e.g. what exactly do we ask people. We will cover both of these during this course. In the remainder of this introduction we briefly touch upon both issues and discuss some of the major drawbacks of the use of questionnaires before we go into the actual nitty gritty.

Wording and design of Questionnaires

The first thing one has to worry about when creating questionnaires is the actual wording of both questions and answers. If we want to measure happiness (for both Dutch and Chinese people) we could for example choose to ask:

- *"Are you happy?"* 0 – Yes 0 – No

We could also ask

- *"Please rate your happiness on a scale of 1 to 10 where 1 means very unhappy, and 10 means very happy."*
  My happiness is: _____ points.

Or we could ask:

- *Please indicate how much you agree or disagree with the following statements:*

Statement	Disagree				Agree
1. *"I am very happy"*	0	0	0	0	0
2. *"I am satisfied with my life"*	0	0	0	0	0
3. *"I often feel depressed"*	0	0	0	0	0

And we could even ask:

- *"I really like my iPhone"* 0 – No 0 – A bit

Each of the above questions differs in wording, in lay-out, and in answer categories. And, each of the above questions will give you a different response. In the end it is up to you, the researchers, to decide which of these different methods of asking give a good assessment of people's happiness.

Probably the answer to the last question, *"I really like my iPhone"* is not a good indicator of happiness. It seems to measure something else than happiness. This is similar to measuring someone's length using a weighting scale: you do get a measurement, but you are actually measuring something else. This we call a validity problem, and we will discuss these kinds of problems in depth in the chapters on validity and reliability.

The wording of both the questions themselves as well as the answer categories will in the end determine whether a statement like: *"The Dutch are happier then the Chinese"* is actually truthful.

Statistical Techniques to Asses questionnaires

Besides wording the right questions and designing questionnaires, we use many quantitative techniques to examine questionnaires. Here I want to give you a bit of rationale behind the quantitative analysis of questionnaire answers.

Suppose we focus on the happiness question that used multiple statements:

Statement	Disagree				Agree
1. *"I am very happy"*	0	0	0	0	0
2. *"I am satisfied with my life"*	0	0	0	0	0
3. *"I often feel depressed"*	0	0	0	0	0

I will call the first statement X_1, the second X_2, and the last X_3. We could now administer this questionnaire (e.g. have multiple people fill it out). Suppose we obtain the following dataset by 3 people:

Person	X_1	X_2	X_3
1	4	4	4
2	4	4	1
3	2	2	5

Thus, person 1 has a score of 4 on question X_1. Hence, she almost fully agrees with the statement that she is happy.

There are now two general ways in which we look at these kind of numbers to say something about a) the quality of the questionnaire, and b) the final scores of individuals. Since in the end we want to say whether person 1 is happy or not – or perhaps how happy she is, we want to work towards a single summary: a single happiness score.

First, we can look at the *internal consistency* of the questionnaire: are the provided answers internally consistent. In the example above this is already tricky: the first person states to be almost fully happy ($X_1 = 4$) but, she is also often depressed ($X_3 = 4$). This is somewhat confusing. We would expect happy people to not be depressed. Hence we would expect an answer pattern like: $X_1 = 4$ combined with $X_3 = 2$. Looking at these kinds of patterns will tell you something about the quality of the questionnaire. We will look at these patterns more formally when we discuss correlations, reliability, factor analysis, and cluster analysis.

Second, we can look at how to *summarize the scores* on the questionnaire into a single happiness score. Given that we have X_1, X_2, and X_3 for each of the three persons, can we give each of the respondents a single happiness score?

For this second objective we could choose to (e.g.) compute an average score for each person. For person 2 we would get an average of $(4+4+1) / 3 = 3$. For person three we would get an average happiness score of $(2+2+5) / 3 = 3$. That seems weird: these two respondents give totally different answers to the questions, and still receive the same score. We should thus assess whether we can validly make such single-number summaries.

The core of numerical analysis of questionnaires, both for internal consistency as well as for summarizing is understanding correlations: The linear association between 2 or multiple variables. Correlations run from -1 to 1, with -1 being a perfect negative correlation and 1 being a perfect positive correlation. Correlations are key in almost all statistical methods of evaluating questionnaires. Correlations quantify the linear association between variables, and we will discuss them extensively. To see why correlations are important, consider the answer pattern over persons for $X_1 = \{4,4,2\}$ and $X_3 = \{4,1,5\}$ as given in the table. Since these two questions measure something that seems to be opposed – if you are happy you are not depressed and vice versa – we would expect a pattern like: $X_1 = \{5,4,1\}$ and $X_3 = \{1,2,5\}$. Or $X_1 = \{1,1,3\}$ and

X_3={5,5,3}. Thus, high scores on X_1 lead to low scores on X_3. This expectation is not met in the actual data from our three respondents. Especially respondent 1 gives counter intuitive answers. The correlation in the expected case would be -1: a perfect negative correlation. The correlation in our observed case is -.69. This might make us wonder whether the relation between X_1 and X_3 is correct.

Similarly, correlations can be used when thinking about summaries. We saw that person 2 and person 3 shared the exact same average score, despite a large difference in their scores on X_1, ..., X_3. To see whether this happens often, one can look at correlations: if the correlation between two variables is high than those who score high on one variable also score high on the second variable. In such cases an average might be a good summary. If correlations are low (or negative) those who score high on one variable score low on the other and vice-versa. If such is the case averages might not be a good summary. Here is an example:

Person	X_1	X_2	Average
1	4	4	4
2	5	5	5
3	1	1	1
4	2	2	2

Person	X_1	X_2	Average
1	1	5	3
2	4	2	3
3	4	2	3
4	5	1	3

For the scores on the left there is a perfect correlation between X_1 and X_2. The averages are also a good summary of the actual scores. On the right the correlation is negative, and every person has the same average scores despite very different raw scores on X_1 and X_2. Hence, in the second case it might not be a very good idea to summarize the scores.

I hope this gives you some intuition as to why correlations are important when assessing questionnaires. We will discuss correlations in depth in the next chapter.

Drawbacks of Questionnaires

During this course you will learn how to create and evaluate questionnaires. However, before we even begin you have to know that making good questionnaires is very hard, and that the outcomes of your research might depend very much on the questions that you ask. Consider for example the following study by Paynes (1951): Respondents were asked the following:

- *'Do you think the United States should allow public speeches against democracy?'*

When asked like this about 21% of the American Public believed that the US should indeed allow such speeches. Consider the second version of this question:

- *'Do you think the United States should forbid public speeches against democracy?'*

A similar group of people respondent to this second – and in many ways equivalent – question. Now, 39% of the respondents believed public speeches against democracy should be allowed.

Don't forget, your wording matters!

Besides the wording, there are many things people are notoriously bad at when remembering and motivating their own behavior: Do you still know what color socks you were wearing last week? Do you know how many hours of television you watched in September 2012? Do you know why you bought the bike that you own?

Because of this, you should be careful what you use questionnaires for. Questionnaires can be an invaluable tool for assessing psychological traits, opinions, moods, etc. However, questionnaires are often ill suited to measure behavior or behavioral intentions. We will discuss when you can, and when you should not, use questionnaires.

Defining the place of the course in the empirical cycle

Developing and evaluation questionnaires is only a part of the so-called "empirical cycle". De Groot (1961) and Runkel & McGrath (1971) identify the empirical cycle running from 1) formulating a research problem, to 2) formalizing a theory, to 3) collecting observations / measuring. After which, one starts again with one.

It has to be noted that in this course we focus pretty much solely on point 3. That's a pity because actually thinking about problems and formulating theory is more fun. Also, in real-life, these things would go together. On the other hand,

focusing on observation and measuring only gives us the ability to really dig into questionnaires without being too worried about theories and research questions.

Since we primarily focus on measuring, we often talk about a *questionnaire* or *scale* – and we sometimes do not even mention what it should measure (happiness? depression? etc.). Then we consider *items* $X_1, \ldots, X_j$ on that scale. These are the actual questions. And we often consider answers of persons $i=1, \ldots, n$ on these same items. Thus we can then denote X_{ij}: the question of the i-th person on the j-th scale. We will use some mathematical formalism, because its often shorter. Please don't be scared about this, and always think back: what does the X mean, what does the i mean, what does the j mean? Etc. From there you will become at ease with the mathematical analysis of questionnaire responses. Sometimes we will omit the i or the j subscript if it is clear that we are talking about a specific person or item. At some points it might seem like we are loosing reality, since you will see only symbols and numbers. But, this course is always about actual measuring, within the actual empirical cycle.

Association & Correlation

All techniques to analyze questionnaires that we discuss in this course are – in one way or another – about *association* between variables; association is the foundation! Thus, in the first lectures we will discuss associations, and (one of its) mathematical formalizations: correlations.

Nearly all techniques (except cluster analysis) assume that *correlation* is an appropriate measure for the association between two variables. Therefore we will first extensively consider correlations and her characteristics / peculiarities

However, we will first start by introducing so-called *measurement levels*. Informally, these measurement levels describe "how much information" a measurement contains. We introduce these measurements levels since correlation(s) assume that variables are measured on *interval scale* or *ratio scale.* Thus, you should know what this means.

Measurement and measurement scales

Measuring is formally defined as: *"Assigning values to objects on the grounds of a characteristic of these objects."* So, for example, if we measure your happiness we assign a value – for example 9 – for the characteristic happiness to you, the object. Whether 9 means happy or not remains to be seen, but at least we have performed a measurement.

Researchers often speak of measurements of different scales: again informally meaning the information captured in a measurement. There are four measurement scales with increasing information in values:

1. *Nominal*: classification only. No order.
2. *Ordinal*: ordering objects
3. *Interval*: values of ordered classes are appropriately interpretable; ratios of differences
4. *Ratio*: values of ordered classes are appropriately interpretable; ratios of differences *and* ratios of the values themselves

The example in the following table shows how each of these compare. Consider the following outcomes of a swimming contest:

	Nominal	Ordinal	Interval	Ratio
Ana	"E"	1 (gold)	0:00	4:00
Klaas	"G"	2 (silver)	0:30	4:30
Tom	"B"	3 (bronze)	2:00	6:00

On the *Nominal* scale all that matters is that there are 3 people that we can identify as different people {E, G, B}. This gives no information about the order, but just identifies three distinct values.

On the *Ordinal* scale {gold, silver, bronze} we can also tell apart the three people (as in the nominal scale), but now there is an ordering: gold is better than silver, which is better than bronze. However, the distance between gold and silver could be very small (only 30 seconds), while the difference between silver and bronze can be very big (1 minute and 30 seconds). Thus, the ordinal scale gives an order, but does not make explicit the size of the differences.

On the *Interval* scale we see that Ana, the quickest swimmer, has a score of 0 seconds. Klaas is 30 seconds slower, and Tom is 2 minutes slower. We can now really see the difference and see that the difference between Ana and Klaas is smaller then between Klaas en Tom. By the way, note that we can also see the order, and identify distinct values. The interval scale thus contains all earlier scales.

The *Ratio* scale contains all earlier scales but now adds a shared starting point. Ana swam for 4 minutes, Klaas for 4.30, and Tom for 6. We can now say that it took Tom 1.5 times the time it took Ana to finish the swim. This we could not have said using the interval scale. The fixed starting point (shared null point) allows us to make statements like "John is twice as fast as Peter". This we cannot do using an interval scale.

The measuring scale of variable is crucial in statistics to define which statistical techniques/tests can be used. Correlation assumes that variables are measured on *interval scale* or *ratio scale* (However, more general measures of association for other measurement levels do exist). The techniques we discuss in this course are not based

on *one* correlation (between two variables), but often on *multiple* correlations (more variables). However, we will first formalize and inspect a single correlation.

Correlation

The correlation *r* between X_1 and X_2, often denoted r_{x1x2} is a measure for the linear association between two variables. Informally you can think of: "To what extend can I predict X_2 if I know someone's score on X_1?

Correlations run from -1 (a perfect negative correlation, to +1, a perfect positive correlation. Before we look at the actual formula let's look at three examples:

Example 1: A perfect negative correlation. $r_{x1x2} = -1$.

X_1	5	4	1	3	2	4	5	1
X_2	1	2	5	3	4	2	1	5

Example 2: A perfect positive correlation. $r_{x1x2} = 1$.

X_1	5	4	1	3	2	4	5	1
X_2	5	4	1	3	2	4	5	1

Example 3: A positive correlation. $r_{x1x2} = .7$.

X_1	5	4	1	3	2	4	5	1
X_2	5	3	2	1	3	2	4	1

We can see that for a positive correlation high scores on X_1 are 'coupled' to high scores on X_2. For a negative correlation this is reversed: high scores on X_1 lead to low scores on X_2. The plots of the different correlations between X_1 and X_2 look like this:

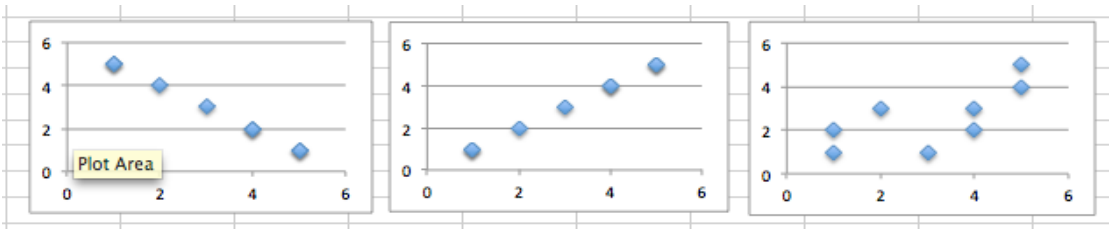


Fig 1.: Correlations. $r_{xy} = -1$ (left), $r_{xy} = 1$ (middle), $r_{xy} = .7$ (right).

You can also see that for the third plot – a correlation that is not perfect – the pattern between X_1 and X_2 (or x and y) is much less clear.

Formulas: Deriving Correlations

One way to understand correlations is to look at the formulas for how to compute them. This will also get you up to speed with thinking more mathematically about the answers X_{ij}.

We first consider a single item, and thus we omit the subscript *j*. To compute a correlation we start by computing the mean (average) score of people *1, ..., i=n* on a question:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{1}$$

To compute the mean $\bar{x}$ we add the score of all *n* people – thus we add $x_1 + x_2 + ... + x_n$, denoted $\sum_{i=1}^{n} x_i$ -- and then divide by the total number of people, *n*. The mean gives a measure of the *central tendency* of scores on the question X.

Once we have computed the mean, we can compute the variance of the item X:

$$\text{var}(x) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1} \tag{2}$$

For the variance we sum the squared distance of individuals scores x_i to the mean score $\bar{x}$ and divide this by n-1. The variance gives an indication for the "spread" of the scores. For example, if ten people all fill out a score of 3 on item X. Thus, $X_1=3$, $X_2=3$, ..., $X_{10} = 3$. Then it is easy to see that the mean is 3, and the difference of each person to the mean is (3-3) = 0. Hence, the variance of X will be 0.

If on the other hand scores are very far away from the mean because some people score very low while others score very high, then $(x_i = \bar{x})^2$ will be high. Thus if people differ a lot in their scores *var(x)* will be high.

The variance of *x* is in units of *x* squared: this is done so that both negative as well as positive deviations from the mean lead to the same additional "spread". However, it is often convenient to work with a measure of the spread in units that are equal to the units of *X*. This is the Standard Deviation of *X*: s_x and it is nothing more than the square root of the variance:

$$s_x = \sqrt{\text{var}(x)} \tag{3}$$

Up till now we have looked only at the central tendency and the spread of a single item. Let's introduce a new item that we call *y*. We can now define the co-variance – the shared "spread" – of x and y:

$$\text{cov}(x,y) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{n-1} \tag{4}$$

Covariances run from –Inf to Inf. To compute the co-variance of *x* and *y*, *cov(x,y)*, we multiply the deviation of an individual score on x_i from the mean of x, $(x_i - \bar{x})$, with his or her deviation from the mean on y_i, $(y_i - \bar{y})$. Next, we add these multiplications over persons *1, ..., i=n*, and divide by *n-1*.

Note that if a person *i* scores higher than the mean on *x* and higher than the mean on *y* the product of the two deviances will be positive. The same is true if a person's scores are lower than the mean on both variables since negative times negative is positive. Thus a consistent pattern of people scoring either high or low on both variables leads to a positive covariance. If however a person scores higher than the mean on *x*, but lower than the mean on *y,* he or she will contribute to a negative covariance.

The final step for computing correlations is to "standardize" the co-variances. Here we make sure that the correlations run from -1 to 1.

$$r_{xy} = \frac{\text{cov}(x,y)}{s_x s_y} \tag{5}$$

By dividing the shared "spread" – the covariance – by the product of the two individual standard deviations we standardized our measure of shared spread.

I hope that gives you some intuition of what correlations mean, and how to compute them. There are faster ways to compute correlations however. The following formula gives the correlation directly from the raw scores and the averages, thus without separately calculation the variances and covariance:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (6)$$

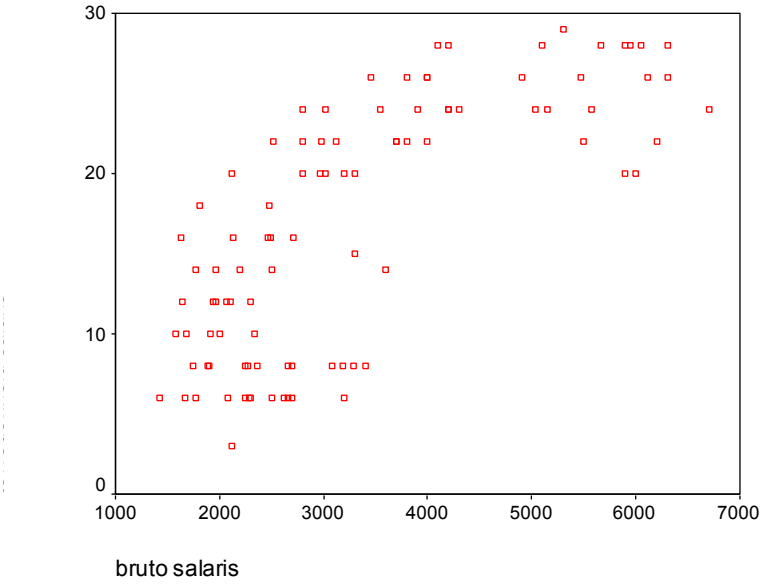
Obviously, Equations (4) and (5) give the exact same result. Make sure to try out the above formulas for yourself and confirm that the examples in the beginning of this section are indeed correct.

Example: Correlations between salary and satisfaction:

Now that we have seen how to compute correlations we want to get a better feel for how this number behaves. Let's look at a plot of *N=95* people of whom we know the following:

- *X*: bruto salary per month in Dutch guilders (its an old example)
- *Y*: satisfaction with salary ("tevredenheid salaris")

The table below the plot gives the mean and standard deviations of each.



Descriptive Statistics

	N	Mean	Std. Deviation
bruto salaris	95	3321.05	1420.86
tevredenheid salaris	95	16.83	7.78
Valid N (listwise)	95		

I hope you can already see from the Figure that there is a positive correlation between *x* (salary) and *y* (satisfaction): Higher incomes seem to be more satisfied. One way to check whether this is true is to see how many people contribute to a positive correlation, and how many contribute to a negative one. You can remember the following rules:

1. people who score *above* the mean on x *and above* the mean on y contribute to positive correlations
2. people who score *below* the mean on x *and below* the mean on y contribute to positive correlations

But:

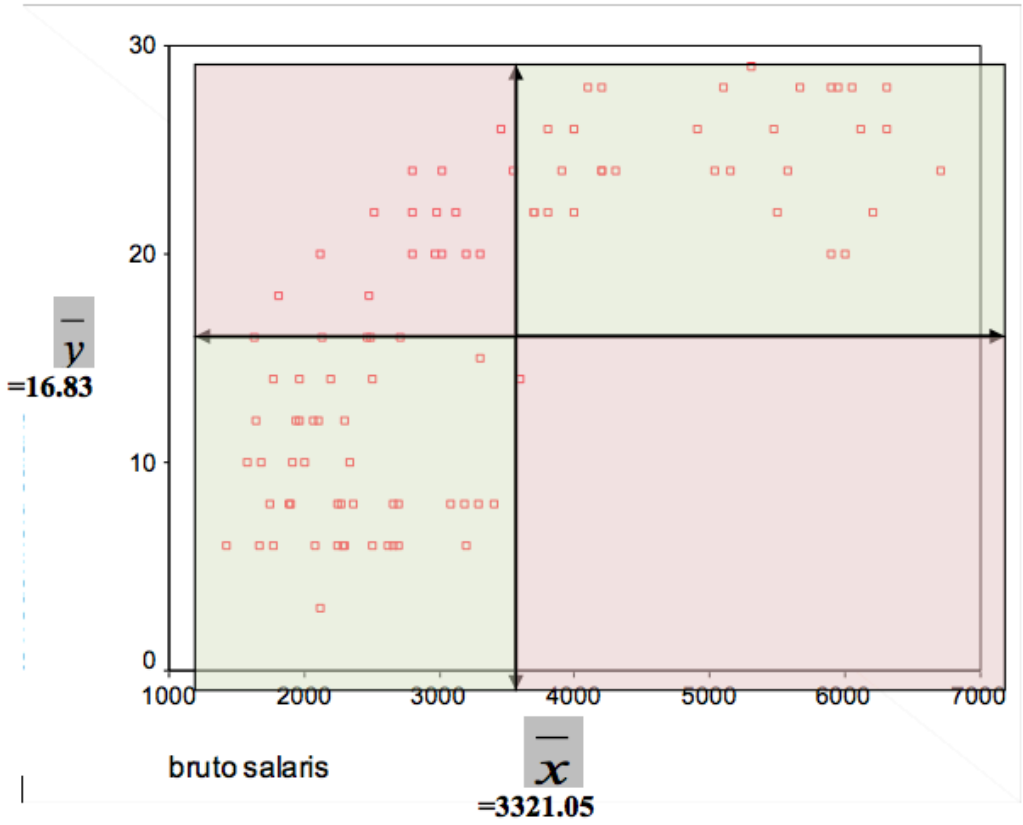
1. people who score *above* the mean on x *but below* the mean on y contribute to negative correlations
2. people who score *below* the mean on x *but above* above the mean on y contribute to negative correlations

In this example we thus have (using SPSS to count the number of people in each cell):

		tevr	
		<16.83	>16.83
sal	< 3321.05	45	14
	> 3321.05	2	34

- 79 positive contributions to correlation
- 16 negative contributions to correlation

We can also plot the number of positively and negatively contributing people:



This provides a quick way for you to guess correlations: if many people contribute to a positive correlation than it is likely positive, otherwise it is likely negative.

In reality the correlation for the above example is, $r_{xy} = .740$. (With covariance, *cox(x,y) = 8179,179*). It is thus indeed positive, but not perfect.

Transformations and Combinations of variables

The concepts of *linear transformation, standardization* and linear *combination* are important for all techniques for analyzing questionnaires. However, these concepts are quite hard. Basically our interest is in the behavior of summaries of items (such as the mean of *x,* or the variance of *y*, when we (linearly) "transform" the original score $x_1, \ldots x_n$. Lets see what this means:

Linear transformation

A linear transformation means that we compute a new score out of an old score. For example, we compute someone's length, v_i, in meters from his or her length in centimeters, x_i:

$$v_i = 1/100 * x_i + 0 \quad (7)$$

Or in general, we linearly transform scores x_i to v_i using:

$$v_i = ax_i + b \quad (8)$$

with constants *a* and *b*. We can now wonder about the mean of v as a function of the mean of *x*, or about the variance of *v* as a function of the variance of *x*. Below are the mathematical rules for such transformations:

$$\bar{v} = a\bar{x} + b \quad (9)$$

$$\text{var}(v) = a^2 \text{ var}(x) \quad (10)$$

Standardization

One very special linear transformation is called *standardization.* Here $a = 1/s_x$, and $b = -\bar{x} / s_x$. It is however easier to remember from the following formula:

$$v_i = \frac{x_i - \bar{x}}{s_x} \quad (11)$$

Applying rules of linear transformation leads to: $\bar{v} = 0$ and var(v) = 1. Thus, after standardization a variable has a mean of 0, and a variance (and standard deviation) of 1. We will often standardize variables before applying statistical techniques because it makes the means and the units of measurement of different variables comparable.

Effect of linear transformations on cov and correlations

We can also examine the effect of linear transformations on the covariances between variables. Suppose we transform x to v using:

$$v_i = a\, x_i + b \quad (12)$$

and we transform y to w using:

$$w_i = \mathrm{c}\, y_i + d \tag{13}$$

we can now wonder about the covariance between v and w. This is relatively simple:

$$\mathrm{cov}(v,w) = ac\ \mathrm{cov}(x,y) \tag{14}$$

We can also wonder about the correlation between v and w, r_{vw}, in terms of the correlation between x and y. By noticing that the correlation is the standardized covariance – as introduced earlier – you can deduce the following:

$$r_{vw} = r_{xy} \tag{15}$$

Hence, you can either standardize the covariance to obtain the correlation, or you can compute the covariance between standardized scores.

To summarize this qualitatively: The…

- covariance changes when the measuring units of x and y change
- correlation does not change when the measuring units of x and y change:

Correlation is a *standardized* covariance!

Linear combination

A linear combination is a sum of linearly transformed variables:

$$v_i = a_1 x_{1i} + a_2 x_{2i} + a_3 x_{3i} + b \tag{16}$$

with constants a_1, a_2, a_3, b. That is even more tricky than just a simple transformation. However, also here we can express the means and variances of the new variable v as a

combination of the means and variances of the old variables. The mean can be expressed as follows:

$$\bar{v} = a_1 \bar{x}_1 + a_2 \bar{x}_2 + a_3 \bar{x}_3 + b \tag{17}$$

and the variance is given by:

$$var(v) = a_1^2\, var(x_1) + a_2^2\, var(x_2) + a_3^2\, var(x_3) + 2a_1a_2\, cov(x_1,x_2) + 2a_1a_3\, cov(x_1,x_3) + 2a_2a_3\, cov(x_2,x_3) \tag{18}$$

This latter term can be made much more general by noting that $cov(x_1, x_1) = var(x_1)$. The general form of Equation 15 is:

$$var(v) = \sum_{j=1}^{J} a_j^2 \operatorname{var}(x_j) + \sum_{j=1}^{J} \sum_{\substack{k=1 \\ j \neq k}}^{K} a_j a_k \operatorname{cov}(x_j, x_k) \tag{19}$$

$$= \sum_{j=1}^{J} \sum_{k=1}^{K} a_j a_k \operatorname{cov}(x_j, x_k)$$

and allows you to compute the variances of linear combinations of *1, ... J*, items.

I will actually not ask you to compute the variances of linear transformation at the exam. However, you should remember the following conclusion:

- Conclusion: (co)variance of a linear combination of variables can be determined using the variances of and covariances between the original variables.

Possible values of Correlations and Covariances

One extremely important thing to keep in mind – which is why I am stressing it in a separate section – is to remember upper and lower bounds of all the quantities we work with:

- Covariance can take on any value – it depends from the values of original scales.

- Minimum correlation is –1 when all points lie on a straight descending line, e.g. $y_i = -2\,x_i + 3$ than $r_{xy} = -1$

- Maximum correlation is +1 when all points lie on a straight ascending line, e.g. $y_i = 0.5\,x_i - 6$ than $r_{xy} = 1$

- No correlation: r=0, when points does not form a straight line.

Next to the possible values, it is very important to know how correlations – since these are the building block for the quantitative analysis of questionnaires – change when the data changes, when distributions changes, etc. Below I give a number of examples. Make sure to become very familiar with correlations!

Effect of distributions on possible values of a correlation

When *X* and *Y* have exact the same distribution, r_{yx} can range between –1 en +1. However, the more the distributions of *X* and *Y* differ, the more limited are the possible values of r_{xy} . For example, the less the number of categories of the variables, the more limited are the possible values of r_{xy}. In the most extreme case, when one of the variables has a variance of 0 then the correlation will 0 (actually, when the variance is exactly zero, the correlation will be undefined…)

Let's look at an example of how the distribution influences the final correlation. Suppose a very simple case, we measure *X*, which only has two levels, *X=0* or *X=1*, and we measure *Y*, also with two levels *Y=0*, or *Y=1*. If we measure 50 people, the margins of the table below give the distribution of answers over *X* and over *Y*:

	Y = 0	Y = 1	
X = 0			45
X = 1			5
	5	45	50

The question is, why do these distributions limit the correlation? This is easy to see once you start filling in the table, while keeping the distributions constant. Suppose we want a very large positive correlation: we then want many people to score {*X=0* and *Y=0*} or {*X=1* and *Y=1*}. These scores contribute to a positive correlation. However, since we cannot mess with the margins, we can at max fill out the following:

	Y = 0	Y = 1	
X = 0	*5*	*40*	45
X = 1	*0*	*5*	5
	5	45	50

We would love to have more people in the {0,0} or the {1,1} cells, but we could not because of the margins. If you compute the correlation of the above table you will find that it is r_{xy} = .111! Hence, because of the margins – the distributions of *X* and *Y*, the maximal correlation you can find is limited.

If we want a very negative correlation we can fill out the table differently, again respecting the margins. We now want as many people as possible in the cells {0,1} and {1,0}:

	Y = 0	Y = 1	
X = 0	*0*	*45*	45
X = 1	*5*	*0*	5
	5	45	50

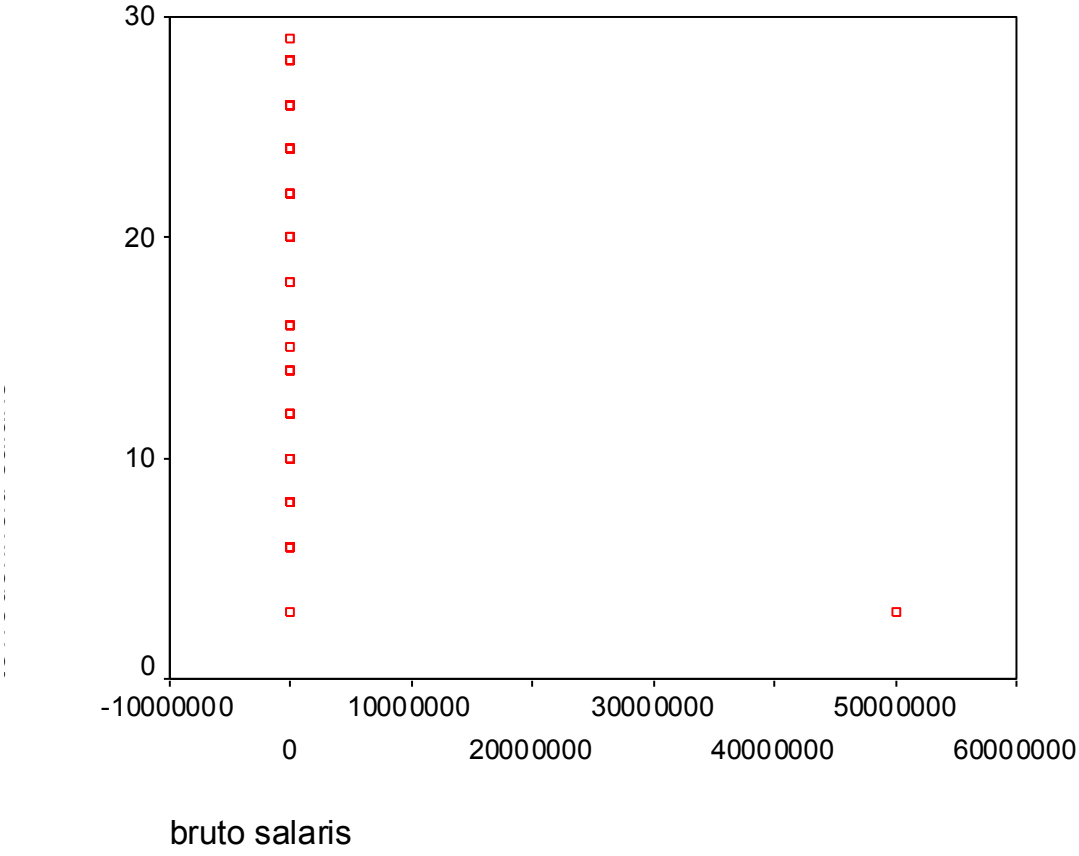
This gives a correlation of r_{xy} = -1.

From this example you should note that:

1. Correlation is generally not a good measure for association between two dichotomous variables
2. When you use techniques that use correlations, make sure that variables have multiple values (>4)
3. Always look at the variances and distributions of variables before correlating them!

Effect of outliers

Outliers – scores of individuals which are very far away from the scores of others – can have a very large effect on correlations. Consider for example the correlations between salary and satisfaction we looked at earlier. The original correlation was .740. Now, we add a very dissatisfied "Bill Gates" who earns 50 million a month, but has a satisfaction score that is only 3. This is depicted below:



If we now compute the correlation again, we obtain a correlation of r_{xy} = -0.179

Would Bill have been happier, things would have been very different. Suppose he makes 50 mln per month, and his satisfaction is 29. We then find a correlation r_{xy} = 0.159.

So remember:

- The contribution of one point to the correlation can be so large that the correlation is almost solely determined by this value
- Always check for outliers!

Effect of non-Linear association

Correlation is a measure for *linear association*. The maximal correlation is 1 for line with positive slope, and the minimal correlation is –1 for line with negative slope. However, a correlation equal to 0 does not imply the absence of association: E.g., a parabola, $y = x^2$ has perfect association between x and y, but $r_{xy} = 0$.

Restriction of range

By *restricting the range* of one variable – thus selecting only a subset of the people – the correlation usually gets lower (closer to zero). Restriction of range means systematically selecting participants based on their scores on one of the two variables. Consider the salary and satisfaction example again. The initial correlation is r_{xy} = 0.740. When we selection only the high incomes ($x > 3321$) then $r_{xy} = 0.336$. This is much lower. When we select only the low incomes low incomes ($x < 3321$): r_{xy} = 0.357. Again this is much lower. This happens because you limit the variance on the item on which you select the participants.

By the way, randomly selecting participants will not necessarily lead to lower correlations. Also, when the true relation between X and Y is strongly non-linear – in which case correlations would not be a good summary – restriction of range can lead to higher correlations. Make sure you understand this latter point.

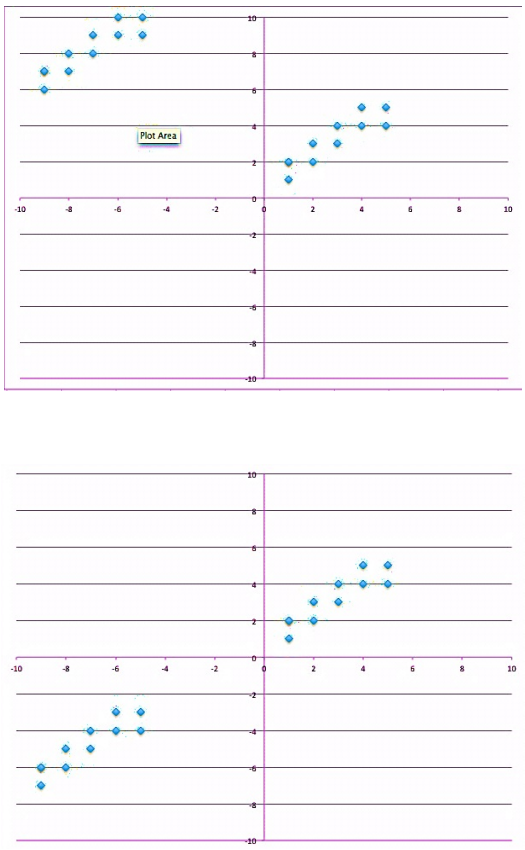
Effect of measurement error

Error, or noise in your measurements can influence the correlation between two variables. We will discuss this in more detail when we talk about reliability and validity. However, for now know that we separate:

1. Systematic measurement error: With systematic error all scores are "moved" in one direction. Basically the scores are linearly transformed: $v_i = a\, x_i + b$. This has no effect on correlations.
2. Unsystematic (or random) measurement error: With unsystematic error correlations will be closer to 0. The actual pattern of scores (e.g. those who score high on x also score high on y) will be distorted by the error. In the limit the error will totally obfuscate the pattern and $r_{xy} = 0$

Merging groups

A group of respondents can consist of multiple subgroups. The correlation in each of the subgroups can differ from the correlation in the total group. To see why please consider the following plots:



Here we see two plots, in both of which two groups are displayed. It is clear that in each of the subgroups of 10 people the correlations are positive, $r_{xy} > 0$. However, combining the two groups does not necessarily lead to a positive correlation. In the first plot (top) the combined correlation is actually negative, $r_{xy} < 0$. In the second plot (bottom) it is positive, and it is even stronger than the original correlations in the subgroups.

Basically, one can "rotate" one group with respect to the other. Thus, the positions of the *means* of the subgroups partially determine the correlation of the total group. In the top plot, the mean of group 1 is lower on *X*, and higher on *Y* than that of group 2. Such a relative position indicates a more negative correlation when combining the groups. In the bottom plot the mean of group 1 is both lower on X as well as on Y: this would indicate a positive correlation.

Testing the correlation

Correlations in a sample are never exactly equal to 0. As you have learned for means in MTO-B / earlier courses, we could do an hypothesis test on the correlation. An often-used test is used to see whether it is likely to observe the correlation that you found in the sample give that in reality the correlation ρ is exactly 0. Thus we test the following:

$$H_0: \rho = 0 \qquad H_1: |\rho| > 0$$

For this test we can compute a t statistics, just as you have done previously to compare means. The t statistic for a correlation for the above null-hypothesis is given by:

$$t = r\sqrt{\frac{n-2}{1-r^2}} \tag{20}$$

where the degrees of freedom, $df = n-2$ are all you need to look up the p-value. SPSS will provide you with hypothesis tests of correlations as well when you compute a correlation table. This course however is <u>not</u> about statistical significance testing.

Bivariate regression

Many of you will have experience with (linear) regression. Here the aim is to predict a *criterium y* as good as possible using one or multiple *predictor(s)* $x_1, \ldots x_k$.

In the simple case we try to predict scores on y (let's say someone's weight) using scores on x (let's say someone's length). We do so by "fitting" a line to the observations: we look for the best linear description of y based on x. This mathematically looks like:

$$y_i = a\, x_i + b$$

The "best" solution to this problem is one in which a is

$$a = r_{xy} \frac{s_y}{s_x} \tag{21}$$

This gives for our salary / satisfaction example: *Satisfaction* = 0.00405 *Sal* + 3.377 where 0.00405 = 0.740 * 7.78/1420.86.

This is of interest primarily because you should know that when the variables are standardized, both s_x and s_y are equal to 1. Hence, when variables are standardized, then $a = r_{xy}$. Regression weights are unstandardized (partial) correlations.

Multiple correlation and explained variance

In case of *multiple regression* the *multiple correlation (MC* or *R)* shows the *correlation* between the *criterium* and the best predicting *linear combination* of *predictors.*

Suppose we use both salary as well as salary^2 (salary squared) to predict satisfaction. We can then compute the multiple correlation of the linear combination of predictors (*Satisfaction predicted = a1 * Sal + a2 * Sal^2 + b*) and inspect the correlation between the predicted satisfaction and the true satisfaction. We than obtain (using SPSS) the following multiple correlation coefficients for the two models:

MC of the bivariate regression with Sal is	0.740
MC of the multiple regression with *Sal* and *Sal^2* is	0.766

The *square of the multiple correlation* shows the "proportion of the variance of the criterium variable" that is *explained* by the regression equation (linear combination of predictors), which is called the *proportion explained variance.* Thus, for the latter regression with both Sal as well as Sal^2 the proportion of variance explained is: $MC^2 = R^2 = 0.766 * 0.766 = .586$; 58.6% of ones salary satisfaction is explained by salary.

Causality

Association (or correlation) *does not* imply causality. There is for example a correlation between weight and length. However, the following sounds bad: *'When I*

eat a lot now, and as a result of that gain 3 kilograms of weight, then my height will increase with 3 centimeters'

What happens is the long people are generally heavier – thus length leads to weight, but not vice versa: you cannot (after a certain age) eat and hope to grow taller.

Because correlation does not imply causality, the term *explained variance* is confusing (explaining is easily associated with causality). So actually a term such as *predicted variance* would be better.

Classical Test theory (CT): Reliability

Introduction to Measuring and Test(s)

Before discussing more modern (and complex) ways to think about scores on questionnaires we will first discuss *Classical Test Theory*. Classical test theory formalizes the observations X_{ij} we obtain by administering a questionnaire to people $i=1, ..., n$, with questions $j=1, ... J$.

Our aim is to measure something: meaning, *to assign a value to an object based on the characteristics of the object*. The object often is a person, and the value we want to assign is a specific score on (e.g.) a trait such as personality, Need for Cognition, Altruism, etc.

Since, in the social sciences we can often not measure the characteristics directly we often work with a test: A systematic classification or measuring procedure. The test is often a questionnaire with multiple items used to measure a single trait. Questionnaires can be used to measure several traits at once, each using different items. For now we will first focus on measuring a single trait using multiple items $1, ... j$. The items together are set to form a scale.

A big part of this course will concern characteristics of scales: how do we create a scale (word the items for example), and how do we (statistically) evaluate the scale. We will use the framework of Classical Test Theory to evaluate the scale statistically (see next section). We will further focus on Reliability of a scale, and Validity of a Scale. This chapter introduces Classical Test Theory and discusses Reliability. Validity we will discuss in the next chapter.

Altruism Scale Example

Let's make the above statement concrete using an example. Suppose we want to measure a trait called "Altruism". There is no direct way – a measurement rod or something – to measure altruism. Thus, we make a scale that is composed of multiple items:

Please fill out the following five items:

		Completely disagree				Completely Agree
X_1	My own interest is mostly more important to me than the interests of others	0	0	0	0	0
X_2	I enjoy it if I can do other people a favor	0	0	0	0	0
X_3	If I do something for someone, I do want something back for it.	0	0	0	0	0
X_4	I do not really think about others' interests	0	0	0	0	0
X_5	I often take others' problems to heart	0	0	0	0	0

Here the scale that intents to measure altruism consists of 5 items. Each is scored on a five-point scale (note the confusing use of the word scale here for both the full test, as well for the answer categories for each item).

Eventually, our intention well be to come to a score X_i for each subject i which is based on her or his scores on all of the 5 items $1, ..., J=5$. Before we will do so we have to see whether these scores can be combined logically, and how we should do so. Classical test theory provides us with a structured way of thinking about this.

Exercise: Based on the above questions, do you think computing an average of the 5 items would lead to a good score on the test? Why?

The Model of Classical Test Theory

Classical test theory provides a way to think about the scores $X_1, ..., X_j$ obtained for a test. The basic idea underlying classical test can be represented by the following formula:

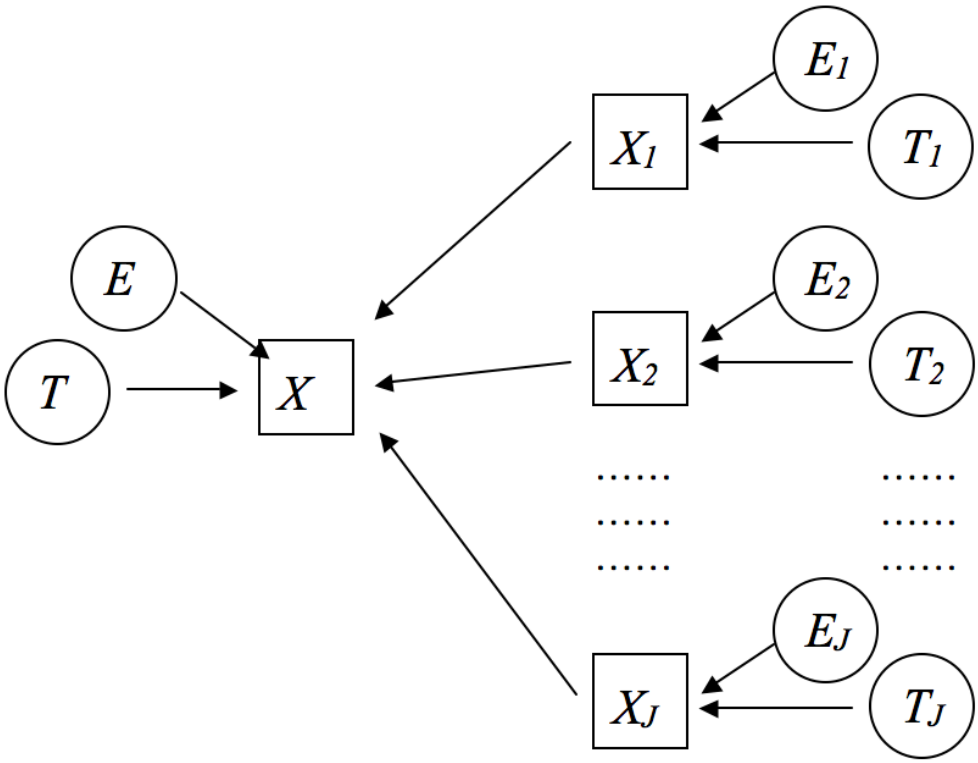
$$X_{ij} = T_{ij} + E_{ij} \quad (1)$$

In words: the score of a person i on item j (X_{ij}) is equal to his or her true score T_{ij} and some error E_{ij}. The following definitions you should know:

- We call X the observed or manifest scores.
- We call both T and E theoretical or latent scores

Keep in mind that we only observe X. The theoretical true score T and the error E are not observed directly. However, we assume that these form the underlying truth that leads to the observed score of X.

To come to a score on a test, classical test theory focuses on the som score of individual items, $X = X_1 + X_2 + \ldots + X_j$. This final sumscore – which is the actual score on the test – is then decomposed into a systematic part: $T = T_1 + T_2 + \ldots + T_j$ and an unsystematic part: $E = E_1 + E_2 + \ldots + E_j$. Graphically we can represent this using the following figure:



Here, manifest variables (the X's) are presented using a square, while latent variables are presented using a circle.

Informally one can think of the model of classical test theory as asserting that whatever we measure (using multiple items) is composed of someone's true trait (for example how Altruistic you really are) and some imprecision in the measurement named *error*. We obviously would like to have very small errors: if the error is 0,

then: $Xij = T_{ij} + E_{ij} = T_{ij} + 0 = T_{ij}$. This would mean that we directly observe the true scores!

Assumptions

In reality, there will be some (measurement) error. However, the crux of classical test theory is to make assumptions about the error. By making a number of assumptions we can show that if we sum over many observed scores to obtain a final score (thus $X = X_1 + X_2 + ... + X_j$) that the final observed sumscore on the test X will be equal to T, the sum over all true scores.

Let's see what kind of assumptions we need for this to work, and see if they are reasonable. The assumptions are:

1. Over persons, we assume the average error to be 0: $\overline{E} = 0$. This is very interesting since (recalling the facts about linear combinations) leads us to conclude that: $\overline{T} = \overline{X}$. Informally this basically says that errors go in all kinds of directions, but not systematically in one specific direction. Averaging over all errors E_i gives 0.
2. Over persons, the error E is independent from everything that E is not a part of. Thus: $r_{EY} = 0$, where Y is a variable of which E is not a part. Note that a special case of this is: $r_{ET} = 0$. This last part means that the errors are not related to the true scores.

You can think of these assumptions about the error as formalizing that the errors are *unsystematic*: they are not related to the true scores, and they average out to 0. Thus, the sum $X_i = X_{i1} + X_{i2} + ... + X_{ij}$ gives us a good idea of T_i.

We also assume this to happen over different persons (as specified above), thus $X = X_1 + X_2 + ... + X_i$ (now dropping the subscript j for the items). X now is the sum score on the test over all individuals. The assumptions above allow you to derive the variance of X since X is a linear combination of E and T:

$$VAR(X) = VAR(T) + VAR(E) \quad (2)$$

This is easily seen since $r_{ET} = 0$. Thus, the variance of X – the spread in test scores of individuals – is the sum of the variance of T which is the actual spread in true scores and the variance of the error scores E. There is no covariance term since this is zero ($r_{ET} = 0$ means $COV(E,T) = 0$).

Note that the above is a theoretical model, and the assumptions might not hold in reality. However, they are fairly reasonable given the idea of unsystematic error. And, this idea allows us to define reliability.

Reliability

Before giving the statistical definition of reliability, we will first give the definition in words: Reliability can be thought of as the stability of a measurement instrument (e.g. a test): *It is the extend to which test scores of individuals remain constant in constant situations.* For example: If we measure someone's length using a measurement rod, the rod is consistent (and thus reliable) if every time we measure the length (and the person did not grow or shrink in the meanwhile) the measured length is exactly the same.

Statistical Definition of Reliability

From our model of classical test theory we can now formally define the reliability:

$$R_{xx'} = \frac{VAR(T)}{VAR(X)} = \frac{VAR(T)}{VAR(T) + VAR(E)} \tag{3}$$

Which states that the reliability, denoted $r_{xx'}$, is equal to the variance (spread) in true scores, so the actual differences between people, divided by the variance in observed scores. However, the second expression is more intuitive: The reliability of X is equal to the proportion of true spread in scores, VAR(T), of the total observed spread, VAR(T) + VAR(E).

To gain some intuition please consider two extreme cases: first, there is a case when the variance of the errors is very very large compared to the variance of true scores. In this case apparently there is a lot of noise / error / imprecision in the measurement and the reliability $r_{xx'}$ tends to 0. If, on the other hand the measurements are very precise, and thus the variance of the errors is 0, then expression 3 tends to 1 since $VAR(T) / VAR(T) = 1$.

Reliability thus quantifies variance of true scores in relation to the error variance. Small error variance gives a reliable measurements $r_{xx'}$ => 1, while large error gives a unreliable measurement $r_{xx'}$ => 0. Note that reliabilities will always be bounded by 0 and 1, and that 1 means perfectly reliable, and 0 means perfectly unreliable.

Four ways of estimating Reliability

Now that we have defined reliability formally, we can start to compute it. However, we have a small problem: we never really observe T or E, we only observe X. So, we don't really know VAR(T) and VAR(E) separately. We only know VAR(X)…

However, not to worry! We have great ways of estimating $r_{xx'}$ without having direct access to VAR(E) or VAR(T). The key to all of these methods is a property of correlations that we discussed earlier: If there is a lot of unsystematic error in a measurement of a variable X, then the correlations of that variable with another variable Y go down. The more unsystematic error, the smaller the correlation!

Similarly, we can use this fact to estimate reliability: if we measure X, and then measure $X_{t=2}$ again, then we can correlate the two: r_{xx2}. If r_{xx2} is close to 1, then apparently there was little error in the measurements. If r_{xx2} is close to 0, then apparently there was a lot of error. Hence, the correlation between X and $X_{t=2}$ is an estimate for the reliability $r_{xx'}$!

All methods of estimating reliability rely on correlations, and the fact that correlations decrease as soon as unsystematic error increases. However, we can think of several ways of correlating different scores of X:

1. *Test-retest reliability:* Test re-test reliability estimates reliability using the same test administered at two points in time. We obtain $X_{t=1}$ and $X_{t=2}$. The correlation between these is used as the reliability estimate.
2. Parallel test reliability: Here we use two different, but parallel tests. Both test measure the same treat, and we correlate the two tests. We thus get $X_{version=1}$ and $X_{version=2}$, and their correlation is used to estimate the reliability of the test.
3. Split half reliability: Split half reliability is based on the idea that a test is composed of multiple items $X_1, \ldots X_j$. We can now compute a sumscore over the first half of the items $X_A = X_1, \ldots X_{j/2}$, and a sumscore over the

second half: $X_B = X_{j/2+1}, \ldots, X_j$. We can now correlate X_A and X_B to estimate the reliability. One thing to keep in mind is that the longer the test, the more reliable the scores (e.g. because than the errors of individual scores are more likely to actually average out to 0). Since the split half method is based on only half the test, usually we use the Spearman-Brown formula to recompute $r_{xx'}$ for the full length of the test. (See further down: "Reliability and Test Length").

4. The Internal consistency method: Since it is not directly clear which halves we should you for the split-half method, the internal consistency method computes the *"average of all possible split half reliabilities"*. More on this in the next section.

Internal Consistency Method: Cronbach's Alpha

The most common measure of reliability in psychological test construction is the internal consistency method. This method computes the *"average of all possible split half reliabilities"*. This is estimated using *Cronbach's α.* (Named after Cronbach, who first wrote about this method).

Before we dive into the definition of *Cronbach's α* and ways to compute it, first note that *Cronbach's α* in practice is an underestimate of the actual (population) reliability. This is caused by the fact that the split-half method (and thus also *Cronbach's α*) is based on the assumption that all items $X_1, \ldots Xj$ are equivalent. This is in reality not the case, and will lower our estimates.

Cronbach's α is not actually computed by averaging over split-half reliabilities. Rather, it is computed directly from a covariance matrix of the covariances between $X_1, \ldots X_j$. The formula for *Cronbach's α* is:

$$\alpha = \frac{J}{J-1}\left(1 - \frac{\sum_{j=1}^{J} \text{var}(X_j)}{\text{var}(X)}\right) \tag{4}$$

To explain how to compute this by hand, consider the following covariance matrix for our 5-item Altruism scale:

Covariance Matrix

	ITEM1	ITEM2	ITEM3	ITEM4	ITEM5
ITEM1	.6565				
ITEM2	.1005	.3946			
ITEM3	.1932	.1276	.5993		
ITEM4	.1834	.1520	.1906	.4904	
ITEM5	.1581	.1319	.0731	.1653	.7239

- Now, $\sum_{j=1}^{J} \text{var}(X_j)$, is the sum of all diagonal elements, and thus 6565+.3946+.5993+.4904+.7239 = **2.8647**

- var(X) is the total variance of X, which is the sum of all of the elements of the covariance matrix: 2.8647 + 2 x (.1005 + .1932 + .1834 + .1581 + .1276 + 1520 + .1319 + .1906 + .0731 + .1653) = **5.8161**

- And thus: $\alpha = \frac{5}{5-1}(1-\frac{2.8647}{5.8161}) = 0.6343$

Note that *Cronbach's α* is used very often in all of social sciences. In practice it is used both to see whether items $X_1, \ldots X_j$, indeed all measure the same construct, and it is used as a mark of quality of the questionnaire. Below we will discuss how *Cronbach's α* is used in practice. However, first we will discuss some features of reliability in general that you should keep in mind.

Properties of Reliability

Before we discuss how reliability is used in practice, we first discuss several properties of reliability.

Reliability and Variance of True Scores

First, note that the variance of true scores limits the possible reliability. In the limiting case, if *VAR(T) = 0*, then *VAR(T) / VAR(X)* will be 0. This is something to keep in mind for example when you select a group of individuals who score high (or

low) on a test. You will then restrict the range of *X*, and likely restrict the range of *T*, thus leading to a lower reliability. Another way to think of this is to consider that correlations go down when you restrict the range of a variable. If the correlation goes down, for example between X_1 and X_2, then the reliability will go down.

Reliability and Variance of Error Scores

As noted earlier, large error scores will decrease the reliability. Thus, if you get more error (or noise) in your measurements, the reliability will go down.

Reliability and Test Length

As hinted on when discussing split half reliability, the reliability of a test goes up once the test goes longer, (e.g. J becomes larger). The relationship between test length and reliability can be expressed formally:

$$r_{KK'} = \frac{K\, r_{XX'}}{1+(K-1)r_{XX'}} \tag{5}$$

Here $r_{xx'}$ is the reliability for a test of length J, and $r_{kk'}$ is the reliability for a new test of length K*J. This is called the *Spearman-Brown* correction for test length.

Let's give a quick example. The reliability (using *Cronbach's α*) of the J=5 altruism scale was estimated at .6343. If we would increase the length of this scale to 20, then K = 20/5 = 4. Now we can estimate the reliability of the new test:

$$r_{kk'} = (4 * .6343) / (1 + (4-1) * .6343) = \mathbf{.8895}.$$

When computing split-half reliability we use a correction of K=2 since the real test is twice as long.

We can also use Spearman-Brown to compute how the reliability would go down if we decrease a test in length. If we decrease some other test in length from J=10 to J=8 items than K = 8/10, and we can fill in the formula again.

Judging Scale Reliability using Cronbach's Alpha

We have now covered all the theoretical material, and will now see how *Cronbach's α* is used in practice. It is used for 2 things:

1. To assess the reliability of a scale (e.g. of a somscore of $X_1 + X_2 + ... + X_J$)
2. To assess the contribution to the reliability of individual items to a scale.

Luckily, we do not have to compute *Cronbach's α* ourselves, we can do so using a computer program such as SPSS. During the practical you will see how this is done.

Here, we will consider the following scale measuring "political dimension" which was filled out by *N=2461* respondents. Here are the item names and descriptions:

Q67.1	Labor Union harder politics
Q67.2	Workers battle for equal positions
Q67.3	Class distinctions smaller
Q67.5	Government interferes with salaries
NEW	Difference between high and low salaries smaller

When we compute *Cronbach's α* for this scale in SPSS we obtain *Cronbach's α = **.731***. We now want to know whether this is "good" or not. There are several rules of thumb that are used:

1. When we want to use the scale to draw conclusions about groups of people a *Cronbach's α < .6* is considered insufficient. A .6 < *Cronbach's α < .7* is considered sufficient, and *Cronbach's α > .7* is considered very good.
2. When we want to use the scale to draw conclusions about individual people a *Cronbach's α < .7* is considered insufficient. A .7 < *Cronbach's α < .8* is considered sufficient, and *Cronbach's α > .8* is considered (very) good.

Given this classification, and if we want to use the above scale to draw conclusions about the political dimension of countries, we would conclude that the scale is "very good".

Note that these are rules of thumb! And thus, they are mostly wrong, but often useful. You should know that a *Cronbach's α* close to 0 is very bad, and a *Cronbach's*

α close to 1 is very good. Psychologist often consider (in practice) *Cronbach's* $\alpha > .8$ to be sufficient. However, there is no need to memorize exact cut-offs for the exam.

Now that we have concluded that the scale overall is good, we can also use *Cronbach's* α to determine whether individual items contribute meaningfully to the scale. If they do not, we can delete them. For this we look both at the correlation matrix, and the "Item-total" statistics which are provided by (e.g.) SPSS:

Correlation Matrix

	Q67.1	Q67.2	Q67.3	Q67.5	NEW
Q67.1	1.0000				
Q67.2	.4891	1.0000			
Q67.3	.3281	.4743	1.0000		
Q67.5	.2372	.2983	.4067	1.0000	
NEW	.1885	.2631	.4356	.4482	1.0000

Item-total Statistics

	Item-Total Correlation	Alpha if Item Deleted
Q67.1	.4249	.7144
Q67.2	.5421	.6650
Q67.3	.5827	.6521
Q67.5	.4723	.6927
NEW	.4538	.6989

Here, often the following rules are used:

1. The "Item-Total" correlation of each item should be higher than .3. This basically means that the correlation of an item (say Q67.1) to the sumscore of all the other items should be higher than .3. In this case it is .4249.

Since for each item the Item-Total correlation is higher than .3 this is no reason to delete items.

2. "Alpha if item deleted" should be higher than alpha. This means that if you would remove an item from the scale (say Q67.1) and compute *Cronbach's α* again for the 4 items that you are left with (Q67.2, Q67.3, Q67.5, NEW), the new *Cronbach's α* is .7144. This is lower than the .731 found for the full scale, and thus the reliability would decrease if we would remove this item. This is a reason to keep it. In this case, *Cronbach's α* would go down no matter which item we delete, and thus we would not delete any items.
3. We look at the content of the items and see if it fits. This is not a statistical argument, but rather a substantive argument. In this case this is hard to evaluate since you cannot see the actual items.

Using the above rules we would decide that the scale itself is good (since the overall reliability is high) and we would conclude that each of the 5 items contributes to the scale (and we would thus not delete any).

Convergence and Divergence

Here I briefly discuss two more criteria that are often used to assess a scale. These are also directly related to validity (next chapter), but I will briefly introduce them. Suppose you have two tests, one measuring Altruism, and one measuring Political Dimension. We can than talk about convergence of items which means that items correlate high to their own scale (e.g. the Item-Total correlation is high, and *Cronbach's α* –if-item-deleted is lower than *Cronbach's α*). Thus, convergence means that items indeed correlate highly to the other items that are supposed to measure the same construct.

Divergence implies that items on one scale have a small correlation to items of the other scale: you would want the Altruism items to correlate low with the Political Dimension items. Why? Well because if the correlation is 1, then they measure the exact same thing and its ridiculous to give the measurement two different names.

Both convergence (high association to items in own scale) and divergence (low association to another scale) are desirable properties of items (and of scales as a

whole, more on this later). We will discuss this more when we discuss validity, and also when we discuss factor analysis.

Validity

After discussing Reliability (informally: whether or not a score is consistent), we will now discuss validity. Validity is *tenability*: *Validity concerns whether or not your are measuring what you intend to measure.* Or, even broader, validity concerns whether or not, given the research setup, measurements, and statistical analysis the conclusions drawn from a research project are valid: (e.g.) do they actually hold true.

Validity as such is a very broad concept. Validity can concern the total setup of a research project, the statistical analysis that is done to derive the conclusions, or merely the measurement instrument.

Here we will focus primarily on three kinds of validity, *criterion oriented validity*, *construct validity*, and *content validity*. These are the types of validity that are primarily used to evaluate measurement instruments. However, before we dig into these types of validity, we will first briefly discuss some other types of validity that you could encounter.

Different kinds of validity

Before digging into the types of validity that are most useful for evaluating measurement instruments its good to give a general overview. I consider two types of validity:

1. **Validity of scientific statements in general**

This type of validity concerns: Statistical conclusion validity, internal validity, construct validity, external validity, etc. etc. You will cover these in MTO-E/MTO-03 MAW (course by Dr. John Gelissen). These types of validity all concern the interpretation research results in general. Hence, they are always relevant and thus also apply to tests/questionnaires.

2. **Validity of measuring instruments**

For the validity of measuring instruments we will consider *criterion oriented validity*, *content validity*, and *construct validity*. Criterion oriented validity concerns the relationship of a test score with scores on another test (often a behavioral measure), which we will call the "criterion". Content and construct validity both

concern the actual questions and wording used in a questionnaire. I will discuss these in more detail below. However, first I will discuss the (conceptual) relationship between reliability (covered in the previous section) and validity.

Relation reliability and validity

Both reliability and validity are desirable characteristics of a test: we would like tests to be both reliable and valid. In layman terms: we want the test to be consistent (reliable) and measure what it intends to measure (valid).

But errors or noise, and therefore a lack of reliability of a test, limit the validity of a test. Informally this is easy to understand: if your test measures a lot of noise (and thus is unreliable), it does not measure what you intent to measure. Since mostly tests are created to measure (e.g.) traits, not noise, a test that measures noise does not measure what it intends to measure.

The phenomenon that an unreliable test leads to an invalid test is called "attenuation". We will formalize this in the context of *criterion oriented reliability* and you will learn how you can (theoretically) correct for this phenomenon.

It can be said that reliability and validity are related concepts in a way that *reliability is a necessary but not sufficient condition for validity.* Note that this is not true the other way around: an invalid measure can still be very reliable. So the following is generally true:

- We can have tests that are both reliable and valid (those are the ones we want).
- We cannot have tests that are unreliable but still valid (impossible).
- We can have tests that are invalid but very reliable.

For example, if we would use an intelligence test that would give different IQ score to the same people when applied in same situations than the test would be neither reliable nor valid because its results could not be trusted to measure anything. On the other hand, if the same test would be reliable, then it could be possible that the test is valid (that it indeed measure intelligence), but it still can be the case that it reliably measures something else than intelligence.

Criterion oriented validity

We will now more formally discuss criterion oriented validity, and in the process we will discuss "attenuation".

Criterion oriented validity indicates if a test is a good predictor of behavior outside the testing situation. For example, consider an intelligence test. We will call the score on the intelligence test *X*. To assess the criterion oriented validity of *X* we need a criterion. Let's choose someone's high school performance as the criterion: we believe that those with a high intelligence should do well in high school, while those with a low intelligence should perform poorly. We will call the high school performance score *Y*. We can now formally define criterion oriented validity as the correlation between the intelligence test score *X*, and the performance in school *Y*:

$$\textit{Criterion oriented validity} = R_{XY}$$

Criterion oriented validity is often used in the context of predicting behavior outside of the testing situation: is the test indeed a good predictor for the actual behavior (school performance in the example) that it is supposed to measure?

Note that criterion oriented validity is measured using a correlation coefficient: the same correlation coefficient as we discussed extensively in previous sections. Thus, all facts that you know about correlation coefficients will also hold in the case of criterion oriented validity: (e.g.) a lot of noise in either *X* or *Y* will decrease R_{YX}, a non-linear relationship will decrease R_{XY}, a very small variance in either *X* or *Y* will decrease R_{XY}, and a restriction of range in either *X* or *Y* will decrease R_{XY} (etc., etc.).

(Correction for) Attenuation

Reliability and (criterion oriented) validity are related through a phenomenon called attenuation. Imagine that, apart from measuring errors, *X* (the test score) and *Y* (the criterion score) correlate perfectly. Therefore – in terms of our model of classical test theory – true scores on *X* (which I will call X^T) correlate perfectly with true scores on Y (Y^T): $r_{XTYT} = 1$. This is the true criterion oriented validity. However, it might be distinct from your observed criterion oriented validity.

Since, in practice, true scores are unknown, you can only calculate the correlation between observed scores on X and Y: r_{XY}. Now, given that r_{XY} is a correlation, it will adhere to all facts we know about correlations. One of these facts is that if either of

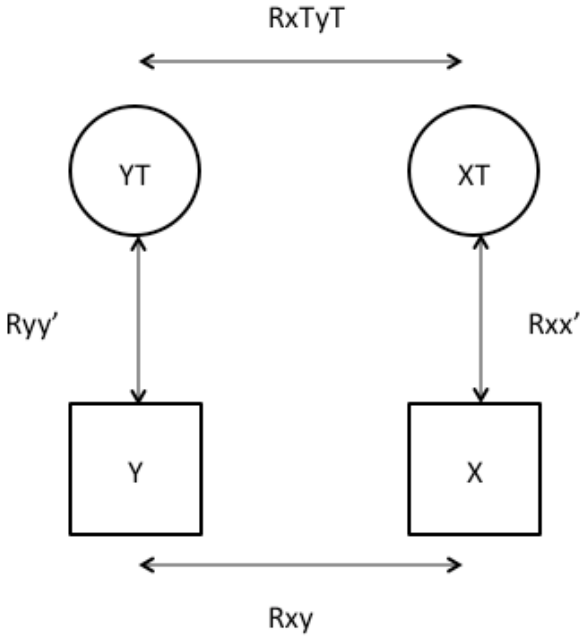
the variables (*X* or *Y*) is measured with noise, then the observed correlation R_{XY} will go down. So in practice, if either *X* or *Y* is not perfectly reliable ($R_{XX'} < 1$ or $R_{YY'} < 1$) than $R_{XY} < 1$. This is called attenuation.

In general, we can summarize this result:

$$R_{XY} \leq R_{XTYT} \tag{1}$$

Make sure you thoroughly understand the above Formula!

If the two reliabilities (RXX' and $R_{YY'}$) are known (or can be estimated), then we can actually estimate R_{XTYT} from R_{XY}. The easiest way of thinking about this is given by the following figure:



We can now describe the exact relationship between r_{XY} and r_{XTYT} (by thinking of this as a pad-model):

$$r_{XY} = r_{X_T Y_T} \sqrt{r_{XX'}} \sqrt{r_{YY'}} \tag{2}$$

Or, equivalently:

$$r_{X_T Y_T} = \frac{r_{XY}}{\sqrt{r_{XX'}}\sqrt{r_{YY'}}} \quad (3)$$

This allows us to "correct for attenuation": We can compute R_{XTYT} if we know R_{XY}, and have estimates of $R_{XX'}$ and $R_{YY'}$.

You should familiarize yourself with questions like:

1. *What is the maximum value that R_{XY} can take when $R_{XX'} = .64$?* You can solve this question by thinking about the maximum values of R_{XTYT} and of $R_{YY'}$. If the true criterion oriented validity is perfect then $R_{XTYT} = 1$. If *Y* is measured without any noise, and hence its reliability is perfect, than $R_{YY'} = 1$. Thus, R_{XY} = 1 * sqrt(.64) * sqrt(1) = .8.

2. *What is the estimate of $r_{XT,YT}$, when $r_{XY} = 0.4$, $r_{XX'} = 0.6$ and $r_{YY'} = 0.9$?* Here we use formula (3) and fill it out: 0.4 / (sqrt(.6) * sqrt(.9)) = .54.

3. *What is the maximum value of R_{XY} when $R_{XTYT} = .8$*. This we can do even without filling out the formula: we know that $R_{XY} \leq R_{XTYT}$. These two are equal if and only if $R_{XX'} = R_{YY'} = 1$ (when the reliability is perfect of both X and Y). In this case we would observe the maximal R_{XY} and thus the maximal $R_{XY} = R_{XTYT} = .8$.

To conclude: Because low reliability lowers the validity, you could say that: 1) reliability is a necessary but not sufficient condition for criterion oriented validity, and 2) because of imperfect reliabilities low r_{XY} are common

Construct and content validity

We will now turn to construct and content validity. Lets first define them:

1. *Construct validity*: Indicates whether a test is a good measurement of the underlying theoretical construct.

2. *Content validity*: Indicates whether the items that are used to measure the construct indeed cover the content of the construct.

Lets discuss both using an example. Suppose we want to measure intelligence, than defining intelligence is the first step to construct validity: which traits / measurements jointly form intelligence? Is it mathematical reasoning and language perception? Or is visual ability also a part of intelligence? These types of questions determine whether or not the construct validity is sufficient. Subsequently, content validity is one step more concrete: Does a question like 1+1=? properly measure mathematical ability? Jointly these are ways to evaluate a test.

We will more formally define construct validity based on three criteria:

1. Is the content domain of the construct sufficient?
2. Is the internal structure of the construct investigated?
3. Is the nomological network specified and tested?

Good content domain of the construct

The content domain of a construct consists of a description of characteristics that the construct is about: *"I want to measure X" -> describe X.*

The items of the test/questionnaire have to be a good representative of the content domain of the construct. The concept of content domain is related to the content validity of the scale: content validity of scale is ok when the content domain of the construct is specified in a good way and the items are specified accordingly.

There are several ways to decide if the content domain of a construct is sufficient. Based on theory, literature research, research on measuring instruments (e.g. other scales) of the same construct, or verification by experts. The content domain has *boundaries* that can be used to judge if items are within or outside a certain content domain. Furthermore, the content domain has *structure* that can be used to distribute items across the content domain.

Unfortunately, many content domains cannot be described in detail in order to provide good basis for item construction or evaluation of content validity (e.g. intelligence, sense of humor, etc.). Thus, assessing the content domain of a construct is always subjective.

Internal structure of the construct is investigated

From the description of the content domain can follow that the construct does or does not consist of multiple separate (sub)constructs (see the intelligence example above). The internal structure of a construct refers to the possible division of a construct in multiple sub-constructs. Research on internal structure is done in part by using statistical techniques such as factor analysis (and a bit of internal consistency analysis – Cranach's α). We will cover some of these techniques in later lectures.

Nomological network of the construct sufficient

The most formal way of assessing the construct validity of a test is to assess its nomological network. This method consists of examining the relations (correlations) between different tests that measure the same construct and different tests that measure different constructs. There are three basic ideas:

1. As constructs get more similar, the correlations between tests of these constructs should be higher
2. Two tests that measure the same construct should have a high correlation (*convergence*)
3. A test should have a low correlation with tests that measure other non-similar constructs (*divergence*)

The nomological network of a construct is sufficient when relationships between the construct and other constructs meet the above basic ideas. Familiarize yourself with these ideas and establish that these are indeed reasonable properties of a test.

The nomological network of a construct can be formally tested using the multitrait-multimethod (MTMM-) matrix by Campbell & Fiske

Multitrait-multimethod matrix (MTMM matrix)

Let us consider several constructs, ideally two: a similar and a non-similar construct. We will call the total number of constructs *I*. Furthermore, let us use different tests, ideally two, to measure these constructs. We will coin the number of tests *J*. When can then create what we call the MTMM matrix by considering the correlations of each of the constructs, measured using the different tests. We thus obtain a matrix of I*J correlations.

		J1		J2	
		I1	I2	I1	I2
J1	I1	B			
	I2	F	B		
J2	I1	C	D	B	
	I2	D	C	F	B

Lets make this concrete using an example. Suppose we measure the following two constructs (I):

Construct 1: English ability (E)
Construct 2: Math ability (M)

And suppose we use two different methods of testing (J):

Measure 1: IQ-test (verbal [IQ-e], nonverbal [IQ-m])
Measure 2: High school grades (Eng, Math)

We now have I*J = 2*2 = 4 variables. And, we can complete our MTMM matrix for these constructs and methods;

		Grades		*IQ*	
		Eng	*Math*	*IQ-e*	*IQ-m*
Grades	*Eng*	*B*			
	Math	F	*B*		
IQ	*IQ-e*	*C*	D	*B*	
	IQ-m	*D*	C	*F*	B

Make sure you interpret each of the possible correlations. For example:

- B: correlation between tests of same construct, same method. This is the reliability of the English grade.
- C: correlation between tests of same construct, different methods
- F: correlation between tests of different constructs, same method
- D: correlation between tests of different constructs, different methods

Logically, we would think that B > C, B > F, B > D. However, more interesting are the following inequalities (which should all hold according to the MTMM matrix). We will also introduce convergence and divergence in this framework:

- C > 0 (*convergence*): The English ability scores using the two different methods correlate (hopefully >> 0). Thus, it does not matter which test we use, we obtain similar scores on English ability. That is preferable.
- C > D (*divergence*): The correlation between English scores using one test or the other is higher than the correlation between the English score and the Mathematics score. This is also preferable.
- C > F (more *divergence*): The correlation between English ability scores obtained using different methods is higher than the correlation between English and Math scores using the same method.

In general note that the basic idea is that you would like a construct to correlate with similar construct measures differently (*convergence*) and not correlate with other constructs – irrespective of the method (*divergence*). Both convergence and divergence are thus desirable properties.

Note that convergence and divergence can also be applied on item level: An item converges when it correlates highly with own scale (corrected item-total correlation). An item diverges when it correlates (much) less with other scales. Again, both properties are desirable. Convergence and Divergence will be discussed in more detail in the practice sessions.

Note: Answer tendencies

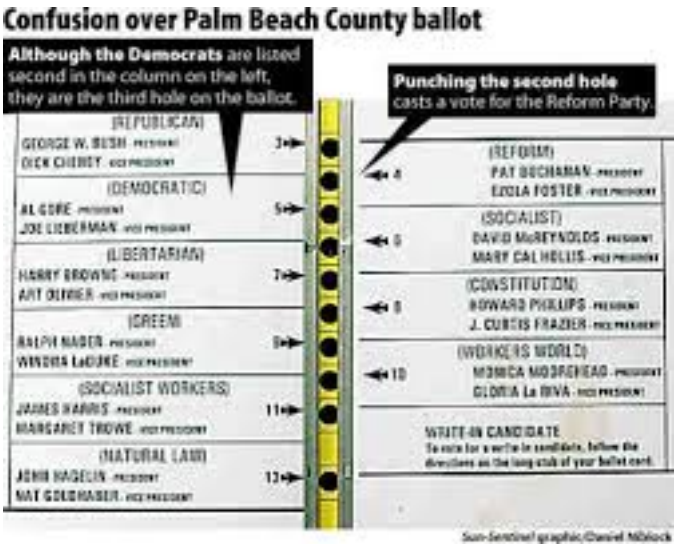
A final note on validity concerns so-called "answer-tendencies": the tendency of respondents to fill out a questionnaire in specific way due to (e.g.) fatigue or perceived social norms. These types of tendencies are a large threat for validity: when

answer tendencies are present than the scores on items are not only determined by the trait that we try to measure but also by confounding factors.

Introduction to questionnaire construction

Up to now we have been discussing a number of formal ways to think about scores of individuals on psychological scales. What we have left out up till now is how to actually create these scales: we have not yet covered the "art" of creating surveys. This next section of the lectures concerns the lay-out and formulation of questionnaires. But first, I want to be very clear about its importance: You can be a great analyst, doing your reliability magic (and more advanced techniques which we will discuss later), but in the end participants respond to what they see on their sheet of paper or on the screen. Making sure that people understand what they are looking at, understand the questions, and can answer them consistently is key to collecting meaningful measures.

To stress its importance, consider the simple survey intended (in the USA) to measure voting preference. During the 2000 USA presidential election Florida used the following "survey" to illicit the votes:



Position one contains "George Bush", and position 2 on the left contains "All Gore". Respondents were supposed to punch one of the black holes in the middle to cast their vote. However, many punched hole number 2, trying to cast a vote for "All Gore". By design however, the second hole referred to "Pat Buchanan", the first name on the right. Bush won Florida (and Pat Buchanan had a surprisingly high score in this state) and subsequently won the country wise election. If only the design of the survey had been different we might have had a different world by now…

Surveying in all its forms

We first consider some of the numerous forms in which surveys (or questionnaires) arrive. A questionnaire can be seen as a special case of a test. A survey comes from "to survey", with a meaning of: "administration of a questionnaire, also called structured interview", or "gathering information from statements of questioned persons, in answer to a problem formulated in advance".

We make a distinction between *structured* and *open* interviews. The first one has a predefined structure, while the second evolves as the surveying is carried out. Mostly, structured surveys are used in quantitative research, while open interviews are usually used in qualitative research. In this course we cover only structured interviews administered on paper or online. We distinguish the following types of surveys:

- On paper vs. orally
- On the telephone
- CAPI & CATI (computer assisted personal interviewing & computer assisted telephonic interviewing
- Group wise vs. individual
- …

More types exist, but the aim of our course is not to give a full overview of all the different methods. Issues of reliability and validity that we discussed in the previous sections will be central for any type of survey.

When to use a survey

When would you use a survey? Surveys can be useful whenever you want to gather data about attitudes, opinions, feelings, thoughts, knowledge, or behavioral tendencies. Surveys (basically an automated way of doing structured interviews) are often cheap, can be done quickly, and can be administered to large groups of people. As such they have advantages over personal interviews. However, they also have drawbacks: people might not always be able to properly recall the things you are asking for, and you might encounter *non-response*: some potential participants might not be willing to participate.

While surveys are cheap and easy, it is very important to consider that surveys might not always give you the correct answers. Non-response (as introduced above), might lead to selection bias: It might for example be the case that those who take drugs are unwilling to participate in a questionnaire about drugs, while those who don't are willing. This would lead to a wrong estimate of the proportion of people who take drugs.

However, there are more troubling examples. LaPiere, in 1934, published an article on "attitudes versus actions". LaPiere spend two years traveling the USA with together with a Chinese couple and visited 251 hotels. They were turned away once during the trip. However, after the trip LaPiere surveyed the same hotels and asked whether or not they would admit people of Chinese race: 128 of the hotels responded and 92% answered "No". LaPiere thus showed that the attitudes (no admit Chinese people) were very distinct from the actual behavior. This signifies that surveys are not a good way to measure future behavior.

At a different level surveys might also be cumbersome: the order of questions, the design of the survey, and the surrounding context might all influence the responses of participants. Dan Arielly for example examined how people respond to the following question:

"What percentage of Afrikan nations is part of the United Nations? Is this more or less then 10%?"

Most of the participants answered "more", and the participants were subsequently asked to estimate the percentage. The average estimate was 25%.

Dan Arielly than asked a similar question to another group of participants:

"What percentage of Afrikan nations is part of the United Nations? Is this more or less then 60%?"

Now, most of the participants answered "less". The average estimate however was 45%! The wording and framing of the question thus had a big effect on people's answers!

Each time you create a questionnaire you will have to think about possible difficulties. Remember that many of us are unable to recall what color socks they

where wearing last week: so recall of many detailed things is likely to be wrong. All of these issues will be in play when designing a questionnaire. In the final section of this chapter I will give a number of "rules of thumb" for good questionnaire construction. These you should always try to adhere to. However, for now remember that surveys, while easy and cheap, will not always lead to the correct answers.

Before discussing the "rules of thumb", we will first discuss a number of formal ways to create scales (measures of a predefined psychological construct).

Formal scale construction

The psychological literature has been concerned with formal methods to create psychological scales. These methods can be (loosely) distinguished based on the following criteria:

- The goal of the construction method
- The method of item construction
- The method of scale construction
- The overall judgment of the scale

The psychological literature distinguishes the following six methods:

1. **The Rational method:** The rational method is aimed to optimize the impression of the scale when evaluated by experts (face validity). Experts are used to create the individual items (based on their expertise) and experts put together the items to make the scale. Often reliability analysis will be used ad-hoc, but the primary means by which the scales are made is by the judgment of experts. This method is often used, but it is very informal and gives little guarantees of reliability and validity.
2. **The Prototypical method:** The aim of the prototypical method is to represent the central elements of the construct one tries to measure well in the final scale. Items are often generated by respondents and selected through "act-nomination": the respondents self-select the items they feel are part of the construct. Experts who put together the items suggested by participants construct the scale. Usually, common measures for reliability and validity are used to evaluate the scale.

3. **The Internal method:** The internal method has the aim of creating an item set with very high reliability. There is generally no theoretical background, and items can be created in all kinds of ways (experts, respondents, etc.). However, the scale is created using reliability and factor analysis (which we will cover later). Items that are unreliable are deleted. This is often used, but because of a lack of theory it often leads to a set of very homogeneous items: these are likely to correlate high and thus be reliable. The validity might however not be very good.
4. **The External method:** The external method aims to optimize the criterion oriented validity. Items can be created in all kinds of ways, but they are selected for inclusion based on their correlation with some criterion. All kinds of items could correlate high, and often the external method thus leads to a very diverse (heterogeneous) set of items. The reliability of these items might be low.
5. **The Construct method**: The construct method aims to "optimize" the nomological network (which we discussed in the previous sections). Here, theory goes before any analysis, as apposed to the internal method. Items are created with the underlying constructs in mind, and are grouped into a scale accordingly. This is a very theory driven approach and hence only applicable if the theory is sufficiently developed. Divergence and convergence are specifically tested.
6. **The Facet method:** The facet method aims to optimize content validity. There is no necessary theoretical background, although it is useful if theory is available. The facet method mainly sets itself apart by the use of "facets" for item construction: an anxiety scale might for example focus on facets of when, where, and what, leading to items like:
    a. "I am fearful when I hear a loud noise in a public place"
    b. I am fearful when I see a sudden movement in a public place
    c. I am … when I hear a loud noise when I am home
    d. I am … when I see a sudden movement when I am at home.

    This way each of the facets of the question can be interchanged.

The most important thing to remember is that there are multiple, theoretically supported, ways of developing an item set and a final scale. Different aims will force

you to make different choices. There is no one single "best" way to create a scale. I can only recommend the following guidelines:

- If possible, items should be tuned to / deducted from theory
- Pay notice to the content validity (e.g. by facet method)
- Evaluate items using reliability and validity as discussed earlier.

Formulating items

We now turn to the hard part: the actual formulation of items. Everybody thinks that he/she can formulate items just like that (à la rational method). This however is a **very big** misunderstanding. Formulating items in a good way is very difficult, maybe even impossible; there are no common rules, at most rules of thumb

Why is it so difficult to formulate items?

There are several reasons why item development is very difficult. We mentioned some in the introduction of this chapter, but here are a few more:

- Respondents often fail to understand questions as intended
- There is often a lack of effort, or interest, on the part of respondents
- Respondents can be unwilling to admit to certain attitudes or behaviors
- Respondent's memory or comprehension processes can be hindered in the (often stressful) process of filling out a survey.
- You, the interviewer might fail: e.g. you might have a tendency to change wording, etc. etc.
- Respondents commonly misinterpret questions. Even the common words such as 'usually', 'generally', 'people', 'children', and 'weekday', elicit a wide range of different interpretations (Belson, 1981)
- Answers to earlier questions can affect respondents' answers to later questions.
- Respondents often answer questions even when it appears that they know very little about the topic
- Cultural context often has an impact
- Etc. etc.

There are a number of studies (besides LaPiere and Arielly) that clearly show possible difficulties:

- Denning in 1944 showed that already after ten days, repeating a question about the age of a respondent might lead to different answers.
- Marquis showed in 1970 that respondents often fail to represent all relevant medical conditions when asked for.
- Respondents' attitudes, beliefs, opinions, habits, interests often seem to be extraordinarily unstable.
- Opinions change, sometimes fast. For example, when asked in beginning and in the end of a survey answers on *same question* were different for 17% of respondents (Gritching, 1986).
- Payne (1951) asked two groups of people the following question(s):
  'Do you think the United States should allow public speeches against democracy?'
  or:
  'Do you think the United States should forbid public speeches against democracy?'
  In the first case 62% thought these speeches should be allowed, in the second case 46% believed they should be allowed.

These – and many other – demonstrations of variability of responses should make you critical of the outcomes of questionnaire studies. Small changes in wording of specific items often have a large effect, and there is no single good formulation of an item. However, guided by theory and our machinery of validity and reliability we can try to do as good a job as possible. And, we have a number of rules of thumb.

Rules of thumb for item formulation

Here I present a list of rules of thumb for item formulation. Make sure you understand each of them. And, if you ever design a survey by yourself, make sure to check each of them! So might sound very trivial, they are still often done wrong in real surveys. So, please do make sure to always check all of these!

1. Respondents should be able to understand the meaning of all used words!

2. Use simple words:
    a. Wrong: *How do you like to recreate?*
    b. Better: *What do you like to do in your spare time?*
3. See to it that the interpretation is unambiguous:
    a. Wrong: *To my opinion, the ratio manager – subordinate is:*
    b. Better: *To my opinion, the communication about work-related business between manager and subordinate is:*
4. Be concrete: Refer to place and time, ask for quantities, dates or data
    a. Wrong: *Do you like fruits?*
    b. Better: *Did you eat an apple yesterday?*
5. Avoid vague words.
6. Try to avoid using 'often', 'sometimes', 'regularly', in the answer alternatives as well!
    a. Wrong: *I am often willing to work on one project for a long time*
    b. Better: *I am willing to learn for an exam daily over the course of a month*
7. Avoid double questions, so do not use 'and' or 'or'-questions:
    a. Wrong: *Our team has a goal in which the vision, the assignment and the values of the team are clearly visible*
    b. Wrong: *The business and the costumers are the most important. Nevertheless I pay enough attention to the employees*
    c. Better: *I pay enough attention to my employees*
8. Avoid double negatives.
9. Avoid not, no, nobody, nothing, never, etc. in question.
    a. Wrong: *I have never seen nobody nowhere in Tilburg that does not wear neither red nor black jeans (a bit of exaggeration, I admit;)*
    b. Wrong: *I do not like to hitchhike on a holiday*
    c. Better: *I find hitchhiking on a holiday unpleasant*
    d. Wrong: *Whether you are a boy or a girl does not matter on our school*
    e. Better: *Boys and girls are treated in the same way on our school*

10. Make short questions. If you need an introduction put it in a separate text box, not in the question itself.
    a. Wrong: *When you have available at you school your own or a shared pc and you barely or never use this pc, what is your reason for that?*
    b. Better: Only answer the next question when you have available on your school your own or a shared computer, but you barely or never use this computer. So, if you do not have available a computer on your school, you can skip this question. When you do have a computer on your school and you regularly use it, you can also skip this question.
    Why do you barely or never use the pc?
11. Avoid suggestive questions:
    a. Wrong: *Do you also have the well-known mensa-hunger feeling in the evening?*
    b. Better: *I find the portions that the mensa serves too small*
12. Do not assume prior knowledge:
    a. Wrong: *The course content was well-constructed*
13. Write abbreviations in full:
    a. Wrong: *Do you preserve old OS's and programs?*
    b. Better: *Do you preserve old operating systems and programs?*
14. Create an equal number of indicative and contra-indicative questions, but without using negatives like 'not'
    a. Wrong: *I do not go into subjects thoroughly*
    b. Better: *I attend to matters superficially*
15. Position the most important part of a question at the end:
    a. Wrong: *I feel relaxed and full of confidence during exams*
    b. Better: *During exams I feel relaxed*

I hope this list helps when creating a scale yourself. However, a questionnaire does not just consist of items, it also concerns answers. Below the rules of thumb for creating answer alternatives.

Rules of thumb for formulating answer alternatives

1. Make alternatives exhaustive
    a. Wrong: How long did you watch television yesterday?
    6 hours or more
    5 hours to less than 6 hours
    4 hours to less than 5 hours
    3 hours to less than 4 hours
    2 hours to less than 3 hours
    1 hours to less than 2 hours
    ½ hour to less than 1 hour
    10 minutes to less than ½ hour

    Why is this wrong? Well, what if you watched less than 10 minutes…

2. Make alternatives mutually exclusive
    a. Wrong: What did you think of surfing the internet?
    Difficult *Fun* *Not fun*
    b. Better: Do you think surfing the internet was difficult?
    No *Yes*
    Do you think surfing the internet was fun?
    No *Yes*

    Why is this wrong? Well, what if you found it both difficult and fun?

3. Use open-ended questions with caution. You should use these only if there is no other possibility, e.g. when you are really not sure about all answer alternatives. Open-ended questions are harder to process and analyze. However, open-ended (follow up) questions can help in the interpretation of deviant responses in close-end questions.

4. Ask for exact numbers, times, places. When the respondent knows the precise answer, you can better ask an open question, e.g. age, how many hours someone works per week, etc.

5. Put all alternatives in logical order. Order alternatives from – to + or the other way around, but be consistent.
    a. Wrong: (1) Totally disagree, (2) Agree, (3) Disagree, (4) Totally agree, (5) Neutral

6. Try not to use 'do not know'. Only use 'do not know' for knowledge questions 'do not know', 'not applicable', 'no answer' categories are a treat to the rather-lazy-than-tired-people. However, do provide the option not to answer a question: you rather have a missing data point then an incorrect answer.

7. Rather not use multiple categories. Multiple category questions are questions in which the respondent may answer multiple answer alternatives. The statistical analysis and content interpretation is a lot harder for multiple-answer questions than for single-answer questions

8. Use the same alternatives for all items. If you need to switch answer alternatives makes this clear in the design of your survey. This makes it easier for respondent to answer questions and it minimizes errors made by respondent

9. Consider the number of answer alternatives. The literature makes a distinction between *bipolar* (negative to positive items) and *unipolar* items (0 to 'many'). Bipolar questions that consist of an odd number of alternatives need to have a neutral middle and be symmetric. Bipolar questions that consist of an *even* number of alternatives need to be symmetric without a neutral middle and are considered "forced choice" questions since you cannot indicate the absolute midpoint. Research shows that scales with 5 answer categories often have the highest reliability. However, when asking about undesirable behavior it is often better to use

7 categories of which the two end-points are extreme: respondents are likely to avoid the extremes.

Finally, a number of advantages and disadvantages of open versus close ended questions can be identified. *Open ended* questions:

- Allow respondent to express themselves in their own words
- Do not suggest answers
- Avoids format effects
- Allow identification of motivational influences and frames of references

However, *close ended* questions:

- Lead to answers can be meaningfully compared since respondents use same answer alternatives
- Produce less variable answers
- Require recognition instead of recall, which makes them much easier to answer
- Are much easier to code and analyze

Note that open ended questions, are often necessary in the early stages of the development of a questionnaire to develop appropriate close ended questions.

Rules of thumb for scale design

Besides proper item construction and proper answer category construction, a lot can be done using a good layout. This is more of an art then a science, but please recall the very bad layout that lead to president Bush being elected: this is not trivial!

A few general remarks can be made:

- Make sure to separate visually the instruction texts and the actual questions

- Put multiple similar items (especially those with identical answer categories) together in blocks. For paper surveys this can be done using spacing, while for online surveys different blocks of questions can appear on different pages.
- Make sure to use a table layout if you have multiple items with the same answer categories. This saves space, and makes it easy for respondents to link the answer categories together.
- Provide clear headings when you move from one subject to the other. You have to take respondents by the hand and guide them through your survey. Again, for online surveys this can be done by using different pages.
- Make the logic of questions clear: if you have conditional questions (e.g. *"Only answer question 2 if you answered "Yes" to question 1"* make sure that the condition is clear visually. For paper surveys you can use indentation and lines connecting the answer on one question to the follow up question. Online you can make surveys dynamic: question 2 will only be presented to respondents who answered "yes" to the first question and will not be displayed otherwise.

Again, developing good questionnaires is an art, but one that can be trained by inspecting and critically evaluating existing questionnaires. In any case, make sure you always *pre-test* your survey: have a number of respondents who are in the target group of your research fill out the survey, and interview them afterwards. Ask them about things that were unclear and ask them to suggest changes. Do not use their data in the actual analysis: the pre-test is merely done to improve the survey!

Concluding comments

Creating a good survey, one that is sufficient in all aspects (reliability, validity, norms, item formulation, formulating alternatives, assessing procedure etc.). is a tremendous amount of work that can take many years. Often we will not have the time to develop and validate our own scales. Therefore, we should resort to validated scales that already exist in the literature if possible: *always see if whatever you are trying to measure has been measured before.*

On a final note, be aware of the fact that you can include “check” questions in your own survey. For example, you can ask specifically for social desirability: ‘Are you always honest?’, or ‘Do you always participate fully?’. Check questions are an active field of study, especially since monetary rewards can be obtained by filling out online surveys: Here, a computer might fill out your survey, and you want to screen such bogus data. This is an interesting topic of study, but we will not discuss it in detail.

I hope the above rules of thumb – and the discussion of formal methods to create scales – will help you to create better surveys in the future. A great way to train yourself is to evaluate existing scales using the rules of thumb that are presented here.

Explorative Factor analysis (FA)

We have focused on reliability and validity, and discussed the "art" of questionnaire construction. It is now time to dig into more sophisticated techniques to analyze questionnaires. We will first discuss so-called "Exploratory Factor Analysis". This is a data-reduction technique: so it can be used to summarize data. The technique was introduced in 1904 by Spearman and is used widely in the social sciences, mathematics, and machine learning (well, in every field really).

During this course we will not dig into the actual mathematics behind factor analysis, but we will discuss exploratory factor analysis in two ways: First, we will discuss a very small numerical example and I will try to explain some basic intuition behind factor analysis. Second, we will discuss the major concepts involved in "real" factor analysis. Finally, we will explore how we can do factor analysis using SPSS, and I will discuss a number of extension of the initial intuition.

Factor analysis will be useful when we measure multiple (sub)constructs. Exploratory factor analysis can be used to determine the number of "underlying" constructs measured by a set of items. These underlying constructs (or factors) can then be used to summarize the data. FA is a different technique than CT in order to summarize the scores on items in a smaller number of scores, each of which measures a certain concept. With CT we summarized items into a single (sum)score, with FA we can summarize items into multiple (weighted) factors: by doing this we can identify the multiple sub-constructs.

The goal of factor analysis is to describe associations (correlations) between a (sometimes very large) number of observed variables by a smaller number of components / factors. Formally the technique aims to:

- To explain as much as possible of the variance of the variables with the as few as possible number of factors / components (explained variance)

- To reproduce as good as possible the correlation matrix with the as few as possible number of factors /components (reproduced correlation matrix)

Note that these two statements are identical. In the context of questionnaires the goal of FA is to assigning items to groups of items, each of which measures a (sub) concept.

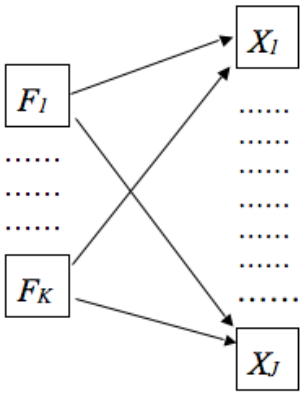
The literature distinguishes between explorative and confirmative factor analysis. With explorative factor analysis we explore how many sub constructs are measured by our item set. With confirmatory factor analysis we test an assumption about a specific number of sub constructs defined a-priori. We will discuss exploratory factor analysis first.

In general, it is useful to set the notation of FA apart from CT, because it helps to introduce the notation that we will use throughout. With CT theory we looked at the reliability of items X_1, …, X_J. (basically, this was based on the correlations between items). If the reliability was satisfactory, we would compute a linear combination of X_1 to X_J by computing the sum score: $X = X_1 + X_2 + \ldots + X_J$. We considered the sum score X the *score* of a respondent on that construct. X was thus a summary of X_1 to X_J.

We slightly change this view when we consider factor analysis. Here we explain the associations between items using (a smaller number of… – but we get to that) factors. We regard:

$$X_j = a_{1j} F_1 + a_{2j} F_2 + \ldots + a_{Kj} F_K \qquad (1)$$

In words, each item X_j is regarded as a linear combination of *Factor scores* F_1, … F_K, and their associated *loadings* a_{1j}, … a_{Kj}. Graphically this looks like this:



I will give some intuition behind the derivation of the factors scores and the weights below. However, for now note that if we can write the score of an item as a linear combination (using a *factor loading* and a factor score):

$$X_j = a_{1j} F_1 \tag{2}$$

than we can also write the factor score as a linear combination of the item and a *regression weight*

$$F_1 = b_{1j} X_j \tag{3}$$

and this will also hold for multiple items / factors. The scores on the factors can now be used to summarize the scores on the items. I will explain below how the summarization actually works, but for now note a) the notation, and b) the fact that the summary (the factor score) will not just be a sum score but rather a linear combination with weights that might be distinct from 1. Hence, not every item will be equally important in the final summary.

Some intuition behind factor analysis

In this section I will give a small numerical example to give you some intuition behind factor analysis. This example is not exact: the factor scores and weights that I show are not computed the way they are actually computed by SPSS. However, it hopefully gives you some basic framework to think about factor analysis. And, it hopefully makes the notation introduced above more clear. Finally, I hope the example shows the reasoning behind factor analysis as a summarization technique.

Suppose we collect a dataset of 5 persons who each answered 3 items, each on a 5 point scale. The data looks as follows:

ID	X1	X2	X3
1	2	2	3
2	3	4	2
3	4	4	1
4	5	5	3
5	4	4	3

We are now asked to summarize this dataset. One thing we should obviously do is look at the correlations: The correlation between X_1 and X_2 is .92, the correlation between X_1 and X_3 is -.15, and the correlation between X_2 and X_3 is -.05. This should already highlight to you that X_1 and X_2 seem to measure the same thing (high correlation), while X_3 measures something else (low correlation with X_1 and X_2). Hence, a good summary of the above dataset would probably need 2 factors (one for X_1 and X_2, and one for X_3).

This is exactly what factor analysis will tell us. By rewriting the score on X_1, …, X_J to factor scores we can start summarizing the dataset. Here is a *possible* factor solution:

	X			F			A			
ID	X1	X2	X3	F1	F2	F3		F1	F2	F3
1	2	2	3	2	3	0	x1	1	0	0
2	3	4	2	3	2	1	x2	1	0	1
3	4	4	1	4	1	0	x3	0	1	0
4	5	5	3	5	3	0				
5	4	4	3	4	3	0				

On the left are the measures scores for each individual, in the middle the factor scores *F* for each individual, and on the right a possible set of factor loadings (these are not the actual factor loadings, I am just using them to give you some intuition). The basic factor analysis trick is that we can reconstruct the *X* scores of each individual by their factor scores and the factor loadings. For example the score of person 1 on X_1:

$$X_{11} = a_{11} * F_{11} + a_{12} * F_{12} + a_{13} * F_{13} = 1*2 + 3*0 + 0*0 = 2$$

Or the score of person 2 on x2:

$$X_{22} = a_{21} * F_{21} + a_{22} * F_{22} + a_{23} * F_{23} = 1*3 + 0*2 + 1*1 = 4$$

Note that this always works: we can always create a set of factor scores and factor loadings that perfectly reproduce all the actual scores. As long as we use as many factors as observed items we can fully reproduce the dataset. Also note that there are

infinitely many solutions to this decomposition: there are (infinite) combinations of factor loadings and factor scores that would allow you to recover the original dataset.

By now you might wonder: "Well, that's all nice and all, but how does this summarize my dataset?" Well, this is where the real factor analysis trick comes in: The factors are not just any decomposition, they are a very specific decomposition in which (*informally*) most of the information in the dataset is placed in the first factor, whatever is left in the second, whatever is left still in the third, and so on and so on. By the time the number of factors is equal to the number of items ($J = K$) the dataset will be reproduced perfectly. The trick however is to not use all the factor scores: $K << J$.

Lets recheck our example and compute, using the factor scores on only 2 of the factors the predicted scores on X, *X'*:

F		
F1	F2	F3
2	3	x
3	2	x
4	1	x
5	3	x
4	3	x

A			
	F1	F2	F3
x1	1	0	.
x2	1	0	.
x3	0	1	.

X'		
x1	x2	x3
2	2	3
3	**3**	2
4	4	1
5	5	3
4	4	3

The score of person 5 on item 3 is computed using 4 * 0 + 3 * 1 = 3 (underlined). If you look at the scores X' and compare them to the original X table you will see that we have managed to replicate it in full, but for the score of person 2 on X_2 (in red). So, we have now used only 2 factors (instead of 3 items), and we can summarize – almost perfectly – our dataset! This can be done because X_1 and X_2 have a very high correlation. This is the basic trick behind factor analysis as a summary tool.

By the way, I hope you also notice that if you look at the factor loadings, *A*, you see that F_1 has a loading of 1 for both X_1 and X_2, while factor three has a loading of 1 for X_3 and zero's otherwise. This is the first step to meaningful interpretation of factor analysis: F_1 apparently explains both X_1 and X_2, and thus X_1 and X_2 measure a common construct. X_3 measures something else, and this is explained by F_2.

This was a very informal description of factor analysis as a summarization technique. I hope this gives you some intuition. Things to remember are:

- If we use as many factors K as items J ($K = J$) then we can fully recreate our dataset. Always.

- Once we select a number of factors K smaller then J ($K < J$) we can use factor analysis to summarize our data.

- Summarization works because items are correlated. If all the items J are uncorrelated factor analysis will not be able to summarize the data well.

- Factor analysis allows you to see "blocks" of correlated items. You can see which blocks of items go together by looking at the factor loadings (as presented in A).

All of this was clear in the example. From here onwards the example breaks down because the decomposition that I choose to make factor analysis intuitive is not the actual one used by factor analysis. The actual factor solution based on the example data (before rotation – but we will get to that topic later) looks like this:

F

F1	F2	F3
-2,49	0,39	0,14
-0,19	-0,58	-0,52
0,68	-1,65	0,27
1,70	1,02	0,03
0,30	0,81	0,07

A

F1	F2	F3
0,69	0,16	0,70
0,70	0,00	-0,70
-0,15	0,98	0,00

You will not need to know how to compute these. However, here you can see the same things: X_1 and X_2 are well explained by F_1 (high factor loadings, underlined in the left table), while X_3 is explained by F_2 (red).

However, you will not be able to re-compute the actual X-values based on this because of a number of differences with the example:

- Actual factor analysis is done on standardized scores

- The aim is to recreate the covariance matrix (since once we have standardized the scores the actual raw score does not matter anymore, we only care about the differences and similarities in scores (the variances and covariances).

- Again, the aim of factor analysis is: To explain as much as possible of the variance of the variables with the as few as possible number of factors / components (explained variance)

So, what we attempt to summarize using factor analysis are not the raw scores. We try to summarize the covariance matrix. We thus do not talk about reproduced scores (as we did in the first example) but rather about the reproduced correlation matrix (which are standardized covariances). We will now discuss the things you need to know about factor analysis, before digging into how its done in SPSS.

Concepts of factor analysis that you should know.

Here is a list of the most important concepts of factor analysis and their interpretation. You need to know these to be able to understand the SPSS output:

- The "*component (or factor) matrix*" which contains the *component (or factor) loadings*. This is analogues to *A* in our example and contains the loadings a_{jk}. It is often called the component matrix because Principal Component Analysis (PCA) is a special case of Factor analysis where the factors are usually referred to as components. We will discuss PCA and Principal Axis Factoring (PAF, another method of factor analysis). However, we will first stick to doing PCA in SPSS.

- The "*communalities*" which are denoted h_j^2. This a descriptive statistic of the factor solution. Note that communalities are indexed by *j*, and thus say something about the item X_j. The communality is the part of the variance in X_j that is explained by the factor solution *F*. Note that if $K = J$, than $h_j^2 = 1$. If $K < J$ and thus the factor solution summarizes the data, then h_{j2} is general smaller then 1.

- The *unicity* denoted b_j^2. This is very simple, and its just $1 - h_j^2$: it is the part of the variance in X_j that is not explained by the factor solution.

- The *eigenvalue* denoted λ_k. This is indexed using a *k*, and thus is a property of factor F_k. It quantifies the variance – out of the total variance – that is explained by factor *k*. Note that since the X scores are standardized the total variance that can be explained is J (J items times 1). The factor solution (or the decomposition) always has the following property $\lambda_1 > \lambda_2 > \lambda_3 > \lambda_4 .... > \lambda_k$. This is the formal description of the fact that was stated in the example: the first factor contains most of the information (and thus explains most of the variance), then the second, then the third, etc.

- Once you know the eigenvalue λ_k it is easy to compute the *proportion of variance explained by a factor*. Suppose we have a set of 6 items (J=6), and the eigenvalue of the first factor is 3 ($\lambda_1 = 3$), then the first factor explains 50% of the variance. (3 out of a total of 6).

- Similarly, we can quantify the *proportion of variance explained by the factor solution*. Here we sum the eigenvalues of the factors we end up selecting (e.g. $\lambda_1 + \lambda_2 + \ldots$), and then see how much of the total variance is explained by these factors together. Suppose $\lambda_1 = 3$ and $\lambda_2 = 1.5$ when we summarize a dataset of 6 items (J=6) into 2 factors. The total variance explained by the factor solution then is (3+1.5) / 6 => 75%. Note that if you choose as many factors as items (J = K), then the sum of eigenvalues ,$\lambda_1 + \lambda_2 + \ldots + \lambda_k$, will be J, and 100% of the variance will be explained.

- When we select a factor solution, we can also look at the *reproduced correlation matrix*. As we did in the example, we can reconstruct, based on the factor solution, the old scores. Since factor analysis concerns standardized scores, all of the information is captured in the correlation matrix. Thus, we can reproduce the correlation matrix. If $J = K$, then we can perfectly reproduce the correlation matrix. If $K < J$, than we are summarizing the data and likely we will not fully reproduce all the

correlations. We can compare the observed correlations with the reproduced correlations to see if our factor solution (our summary of the data) represents the data well.

- The *residual correlation matrix* is a more formal way of looking at the difference between the observed correlations and the reproduced correlations: it is the difference between the two. So, the residual correlation matrix contains in each cell the difference between the actual observed correlation and the reproduced correlation based on the factor solution.

The current section might have raised some question: I have presented a number of things you have not yet encountered. And, you might have open questions such as: "How many factors should I use to summarize my data?" and "How do I interpret the factor solution?". We will cover all of these, and more, by thoroughly discussing an example of a factor analysis in SPSS.

Factor analysis in SPSS

The dataset we use for this example concerns 300 individuals who all filled out 6 items (thus, J=6). Here is the correlation matrix of the items:

Correlation Matrix

		X1	X2	X3	X4	X5	X6
Correlation	X1	1.000	.449	.443	.296	.314	.326
	X2	.449	1.000	.446	.312	.264	.250
	X3	.443	.446	1.000	.279	.258	.282
	X4	.296	.312	.279	1.000	.467	.516
	X5	.314	.264	.258	.467	1.000	.497
	X6	.326	.250	.282	.516	.497	1.000

We can now try to use factor analysis to reduce the 6 items into a smaller set of factors, $K < J$.

Note that if you look at the correlation matrix you see that X_1 to X_3 seem to correlate high, $r > .4$, and the same is true for X_4 to X_6. However, these two "blocks" have a low correlation with each other, $r \sim .3$. This should already indicate to you that this dataset might be summarized using two factors: one for items X_1 to X_3, and one

for X_4 to X_6. It might not always be easy to actually see such a structure from a correlation matrix, but it always is a good idea to check this.

The output of a factor analysis (PCA)

We now run a factor analysis (PCA) on this data using SPSS. In the practical you will see exactly how to do this yourself. One of the most important outcomes of the factor analysis is the following table:

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2,802	46,700	46,700	2,802	46,700	46,700
2	1,081	18,015	64,715			
3	,578	9,639	74,354			
4	,560	9,335	83,690			
5	,522	8,693	92,382			
6	,457	7,618	100,000			

Extraction Method: Principal Component Analysis.

This table shows (under "Initial Eigenvalues") the eigenvalues of the factors using the mathematical decomposition in which most of the variance is explained by the first factor, then by the second, then by the third, etc. It is clear that indeed the eigenvalues go down: 2.801 > 1.081 > .578, etc. etc. Note that if you add all the eigenvalues of the 6 factors (and thus when $K = J$), the sum is equal to 6. This means you explain all of the variance in the items when you use a 6 factor solution.

The percentage of variance explained by the factor can be found in the *% of variance* column, and you can see there that the second factor explains 18.015% of the total variance. The *cumulative %* column shows how much variance is explained by a factor solution of size *k*: if you would choose 3 factors, then you would explain 74.354% of the variance.

Interpreting the 1 and 2 factor solutions

The table above – while columns 2 to 4 give you the eigenvalues for all the possible factors – is actually the output produced by SPSS when you select only 1 factor to summarize the data. This is why only for "component" 1 the last three

columns are filled out. In the next section we will discuss how many factors you should choose, but let us first try to understand all of the output when we select a single factor.

When we only select a single factor our model is very simple:

$$X_j = a_{j1}F_1$$

Thus, the score on a single factor F_1, and the factor loading a_{j1} summarize each item X_j. The factor score is itself the linear combination of the items ($F_1 = b_{11}X_1 + \ldots + b_{61}X_6$) that explains most of the variance in X_1 to X_6. Or, equivalently, it is the single linear combination that reproduces the correlation matrix best. SPSS will give you a table containing the factor loadings (the a_{j1}) for the single factor solution:

Component Matrix[a]

	Component
	1
X1	,687
X2	,654
X3	,651
X4	,708
X5	,688
X6	,710

Extraction Method: Principal Component Analysis.

a. 1 components extracted.

Note that, for the simple one factor solution (we will add more complexities later), these factor loadings equal the correlation between the item and the factor. Thus: $r_{X1F1} = a_{11} = 0.687$.

Also note that this table allows you to compute the communalities and the unicities:

- De *communality* h_j^2 of X_j is that part of the variance of X_j that is explained by F: $h_j^2 = r_{XjF}^2 = a_{j1}^2$. For example: $h_1^2 = a_{11}^2 = (0.687)^2 = 0.472$.

- De *unicity* b_j^2 (unexplained variance) of X_j is that part of the variance of X_j that is **not** explained by F: $b_j^2 = 1- h_j^2$. For example: $b_1^2 = 1- h_1^2 = 1 - 0.472 = 0.528$.

Both of these tell you something about the "fit" of the factor solution. In this case all of the items seem to be equally "important" on the factor (the factor loadings are all very similar), and the variance explained in the items by the single factor solution do not differ a lot.

SPSS will also give you the reproduced and residual correlation matrix:

Reproduced Correlations

		X1	X2	X3	X4	X5	X6
Reproduced Correlation	X1	,472[b]	,450	,447	,487	,473	,488
	X2	,450	,428[b]	,426	,463	,450	,464
	X3	,447	,426	,423[b]	,461	,448	,462
	X4	,487	,463	,461	,502[b]	,487	,503
	X5	,473	,450	,448	,487	,473[b]	,488
	X6	,488	,464	,462	,503	,488	,504[b]
Residual[a]	X1		,000	-,004	-,191	-,159	-,162
	X2	,000		,021	-,152	-,186	-,214
	X3	-,004	,021		-,181	-,189	-,179
	X4	-,191	-,152	-,181		-,021	,014
	X5	-,159	-,186	-,189	-,021		,009
	X6	-,162	-,214	-,179	,014	,009	

Extraction Method: Principal Component Analysis.

a. Residuals are computed between observed and reproduced correlations. There are 9 (60,0%) nonredundant residuals with absolute values greater than 0.05.

b. Reproduced communalities

This you can use to see how well specific relations between the items are reproduced. Apparently the relationship between X_1 and X_2 is reproduced perfectly by the factor solution: the residual correlation in 0. However, the reproduced correlation of .488 between X_1 and X_6 is apparently not very correct: it is -.162 off form the observed correlation.

The reproduced correlation between two variables when you use a simple one-factor solution is equal to the product of the component loadings of these variables on the component F_1. For example: $r_{12} = a_{11}*a_{21} = 0.687*0.654 = 0.450$

The residual correlation is the difference between the data and the prediction: For example: The residual correlation $r_{12} = 0.449-0.450 = 0.000$.

Finally, notice that on the diagonal of the reproduced correlation matrix you find the communalities.

This is all the output you need to understand based on a simple 1 factor solution. (We will get to how many factors to select in a minute). Lets now explore a two factor solution for this same dataset. When selecting a second factor we compute a second linear combination of the items that explains as much as possible of the variance that is left after the variance explained by the first factor has been accounted for. The second factor is $F_2 = b_{12}X_1 + \ldots + b_{62}X_6$. The model now is extended:

$$X_j = a_{j1}F_1 + a_{j2}F_2$$

Thus, the scores on X_1 to X_j are reproduced using two factor scores and their factor loadings.

Note that the default decomposition provided by SPSS – recall that there are an infinite number of possible decompositions – has two properties:

1. The solution is chosen so that factor one explains most of the variance, then 2, then 3 (this we had seen before).

2. The solution is chose such that the factors are uncorrelated: thus $R_{F1F2} = 0$.

When selecting a two-factor solution the total variance explained table looks like this:

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2,802	46,700	46,700	2,802	46,700	46,700
2	1,081	18,015	64,715	1,081	18,015	64,715
3	,578	9,639	74,354			
4	,560	9,335	83,690			
5	,522	8,693	92,382			
6	,457	7,618	100,000			

Extraction Method: Principal Component Analysis.

Note that it is exactly the same as our previous version, but for the addition of the second component to the last three columns.

SPSS will also give the new factor loadings for the 2 factor solution:

Component Matrix[a]

	Component	
	1	2
X1	,687	,375
X2	,654	,467
X3	,651	,464
X4	,708	-,385
X5	,688	-,415
X6	,710	-,432

Extraction Method: Principal Component Analysis.

a. 2 components extracted.

Notice that the loadings on the first component have remained the same. The second component explains the variance and correlation that is not explained by the first component. In this table the loading a_{j2} is equal to correlation between X_j and F_2 (r_{XjF2}). For example: $r_{X1F2} = a_{12} = 0.375$

We can again compute the communalities and unicities:

- The *communality* h_j^2 of X_j is that part of the variance of X_j that is explained by F_1 and F_2: $h_j^2 = r_{XjF1}^2 + r_{XjF2}^2 = a_{j1}^2 + a_{j2}^2$. For example: $h_1^2 = a_{11}^2 + a_{12}^2 = (0.687)^2 + (0.375)^2 = 0.613$. Notice that more variance in X1 is explained using 2 factors then with the previous 1 factor solution.

- The *unicity* b_j^2 of X_j is the variance of X_j which is **not** explained by F_1 and F_2: $b_j^2 = 1 - h_j^2$. For the first item this is $1 - 0.613 = 0.387$

You can also find the communalities directly in the SPSS output:

Communalities

	Initial	Extraction
X1	1,000	,613
X2	1,000	,646
X3	1,000	,639
X4	1,000	,649
X5	1,000	,646
X6	1,000	,691

Extraction Method: Principal Component Analysis.

The *Eigenvalue* λ_k, which we could get from the "total variance explained" table can also be computed. Recall that this is the total variance is explained factor k. For this second component that we have added, we can compute the eigenvalue using the factor loadings:

$$\lambda_2 = a_{12}^2 + a_{22}^2 + a_{32}^2 + a_{42}^2 + a_{52}^2 + a_{62}^2 = 1.081$$

This is easily seen since the squared loadings give you the variance explained in an item (e.g. a_{12}^2 gives the variance explained in item 1 by factor 2). If you add all of these for a single factor then you obtain the total variance explained by that factor.

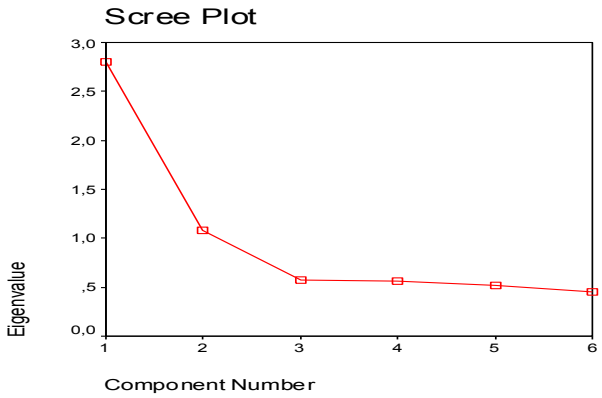
How many factors?

We have now examined a simple 1 and 2 factor solution and we have gotten acquainted with eigenvalues, communalities, and explained variance. However, there is an obvious question as to how many factors to select: If we select $K = J$ factors we explain fully all the variance but we do not have a summary. If we select $K << J$ then we might not capture all the variance well. So, how do we chose K?

There are basically three rules for choosing K. One is a very dumb "rule of thumb", the others make a bit more sense. The dumb rule is to choose $K \leq J/3$. This basically states that you want each factor to (on average) at least summarize 3 items. However, there are smarter rules:

- The Kaiser-Guttman rule: This rule states to choose the number of factors equal to the number of factors with an eigenvalue greater than 1. There is some sense to this rule: if the eigenvalue of a factor is higher then 1, then it explains more variance then a single item does (each item basically adds 1 to the total variance that there is to explain). If the eigenvalue is lower then 1 then the factor (informally) contains "less information" then a single item. Thus, it is not really a summary and you do not want to include it. In our example only F_1 and F_2 have an eigenvalue higher than 1.

- The second rule is also based on eigenvalues, and it is called Cattell's scree test. It is done by looking at a plot of the eigenvalues:



The plot shows the eigenvalue of each component. If you now take a ruler and draw a straight line through the lowest eigenvalues (3 to 6 in this case), then all the factors that have an eigenvalue higher than this line should be included. In this case you would thus select 2 factors.

Rotation

Now you know how to interpret a factor solution, and how to choose a number of factors. However, we have not really discussed any interpretation of the factor solution: how do we know what the factors summarize. This we can do by looking at the factor loadings: this will tell us which items relate strongly to which factor and thus will allow us to interpret the factor. However, the default mathematical solution that SPSS choses for factor analysis (with $\lambda_1 > \lambda_2 > \lambda_3 > \lambda_4 \ldots > \lambda_k$ and $R_{F1F2} = 0$) is not always the simplest to interpret. It is mathematically the easiest to compute, but if we want a good interpretation we can look at different solutions (remember there is a theoretically infinite number of solutions!). We call the examination of different factor solutions *rotation*. Note that you first need to determine the number of items before you start rotating the solution for interpretation!

You will need to understand 3 versions of a factor solution:

1. The standard mathematical solution. This is the one we have just covered, so this one is tackled.

2. The VARIMAX solution. This is a rotation method where the factors are still uncorrelated (thus $R_{FmFn} = 0$) as in the mathematical solution. However the rotation tries to find a simple structure: if the items group together in 2 groups and you have selected 2 factors this solution will attempt to not put as much variance as possible in the first factor, than in the second but it will attempt to use 1 factor to explain one of the groups and the other factor to explain the other group.

3. The OBLIMIN solution. This is a rotation method where alike VARIMAX the solution attempts to create a simple structure. However, additionally the assumption that the items are uncorrelated is relaxed: $R_{FmFn} \neq 0$.

Before discussing rotation examples on the 6 item dataset that we have been using throughout this section it is useful to understand what you do when you rotate after choosing a number of factors. When you chose a number of factor you decide how much of the total variance you want to explain. After you have made this choice, you can redistribute the variance over the factors: you can add less variance to factor 1 and more to factor 2 to get to a solution that is easier to interpret. Suppose a 2 factor solution explains in total 60% of the variance, and factor 1, in the standard solution, explains 40%. You can now redistribute the variance over the factors (for example F_1 contains 32% and F_2 contains 28%) to aid interpretation. Note that the total variance explained by the factors will not change if you rotate: this has been fixed after you decide on the number of factors.

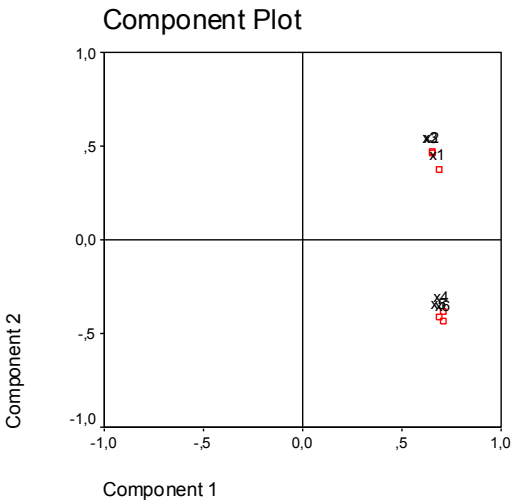
Lets continue our example and see how VARIMAX and OBLIMIN play out. The solution before rotation looked like this:

Component Matrix[a]

	Component	
	1	2
X1	,687	,375
X2	,654	,467
X3	,651	,464
X4	,708	-,385
X5	,688	-,415
X6	,710	-,432

Extraction Method: Principal Component Analysis.

a. 2 components extracted.



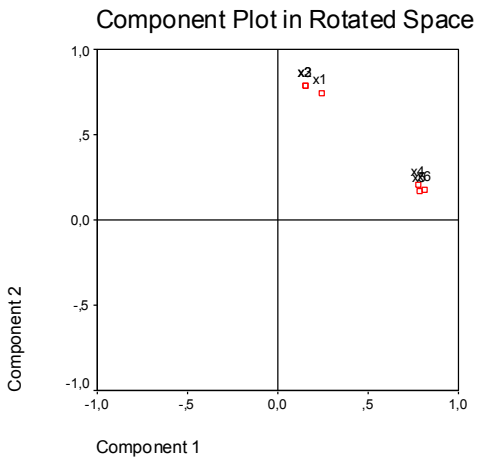
Here we see that factor 1 has only positive loadings, while factor 2 is used to split the two groups of items: recall that X_1 to X_3 correlate high, and X_4 to X_6 correlate high. Basically, in this solution F_1 captures all the joint variance, and F_2 makes the difference between the two "blocks" of items. The VARIMAX rotation looks like this:

Rotated Component Matrix[a]

	Component	
	1	2
X1	,241	,745
X2	,154	,789
X3	,153	,784
X4	,779	,208
X5	,785	,171
X6	,813	,174

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

a. Rotation converged in 3 iterations.



Here the variance is redistributed: Each (rotated) component now explains a different amount of variation in data: 1.989 and 1.895 vs. 2.802 and 1.081. This you can find in the "total variances explained" table. However, to understand what is happening have another look at the plot: Now items X_1 to X_3 score high on factor 2, and X_4 to X_6 score high on factor 1. Thus, we have restructured to solution to make it more interpretable!

You can now directly say that Factor 1 measures X_4 to X_6, while Factor 2 measures X_1 to X_3. This is a much simpler structure. The factor loadings of a rotated solution are often used for interpretation.

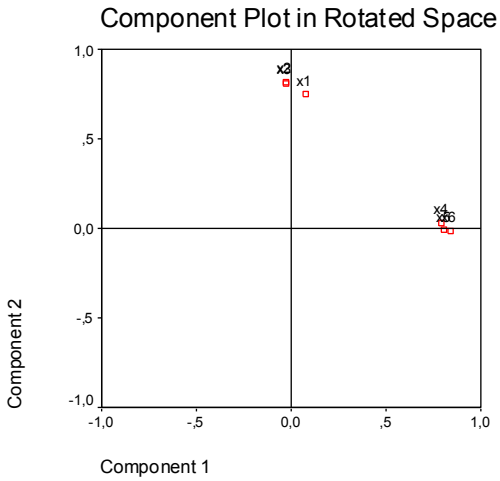
We can also formalize a simple structure: if each item loads on a single factor more than .3, and lower than .3 on the others then the structure is simple. Off course .3 is an arbitrary cut off, but it's a useful rule of thumb.

The OBLIMIN rotation looks like this:

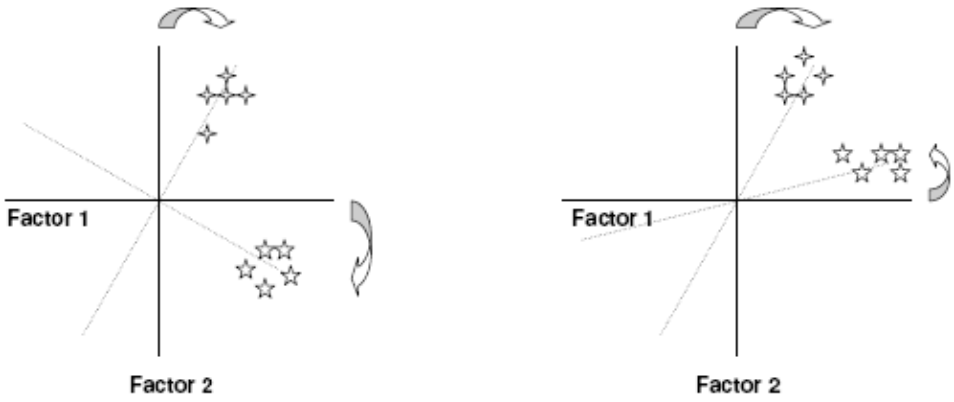
Pattern Matrix[a]

	Component	
	1	2
X1	,076	,747
X2	-,030	,816
X3	-,030	,812
X4	,792	,030
X5	,808	-,010
X6	,837	-,014

Extraction Method: Principal Component Analysis.
Rotation Method: Oblimin with Kaiser Normalization.
a. Rotation converged in 5 iterations.



This is even simpler! To understand what the difference is take a look at this plot showing what the rotations actually do:



With VARIMAX rotation (on the left), we try to rotate the axis so that they are close to the clusters of questions. With OBLIMIN (on the right), we do the same but we

allow the factors to be correlated. Thus, the factors are not orthogonal anymore: they factors do not need to make a straight angle. VARIMAX is an orthogonal rotation since the correlations are set to 0. OBLIMIN is not orthogonal and the factors are allowed to correlate.

Note that in this example OBLIMIN rotation gave the easiest solution. This is often the case, and it is often far-fetched to assume that factors are fully uncorrelated. Hence, I usually prefer the OBLIMIN solution. You should however know the differences between the two versions. If the correlation between the factors in the OBLIMIN solution is very low, you might settle for the easier VARIMAX solution.

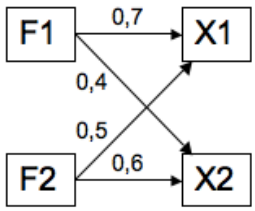
The table below gives on overview of the differences between OBLIMIN and VARIMAX (as compared to the standard mathematical solution):

	VARIMAX	**OBLIMIN**
a_{jk}	different	more different
l_k	different	more different
h_j^2	identical	identical
b_j^2	identical	identical
% expl.var.	identical	identical
R_{prod}	identical	identical
R_{res}	identical	identical

Note that while the factor loadings and the eigenvalues change, the explained variance in both the items as for the full solution do not change: with rotation we are merely redistributing the explained variance over the factors.

Finally, lets be a bit more specific about the changes in a_{jk} when we rotate. We had seen before that $a_{jk} = r_{XjFk}$: The factor loading a_{jk} is equal to the correlation between item J and factor K. This is still true when rotating VARIMAX. The solution will change because of the rotation, but the interpretation is still the same. However, when we move to OBLIMIN, this is not anymore the case. Thus a_{jk} (which is given in the "Pattern Matrix" in SPSS) $\neq r_{XjFk}$ (which is given in the "Structure Matrix").

To see why, we need to get back to the factor analysis model. Suppose we have a very simple 2 item, 2 factor solution. If we now look at the VARIMAX model (top of the figure below) and the OBLIMIN model (bottom) we can see the difference:



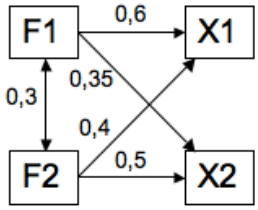
$r_{F1X1} = a_{11} = \underline{0.7}$

$r_{F1X2} = a_{21} = \underline{0.4}$

$r_{F2X1} = a_{12} = \underline{0.5}$

$r_{F2X2} = a_{22} = \underline{0.6}$

Note: (a_{xx} – first index is for variable and second is for factor)



$r_{F1X1} = a_{11} + (r_{F1F2} * a_{12})$

$= 0.6 + (0.3 * 0.4) = \underline{0.72}$

$r_{F1X2} = a_{21} + (r_{F1F2} * a_{11})$

$= 0.35 + (0.3 * 0.5) = \underline{0.50}$

In the second model there is a correlation of .3 between F_1 and F_2. This affects (like in path analysis) the relationship between F_1 and X_1, since a part of the relationship goes "via" F_2. Because of the correlation introduced by OBLIMIN, we need to include this extra path when we compute the correlation between a factor and an item when we are looking at an OBLIMIN solution.

PAF versus PCA

The factor analysis we discussed above was an example of Principle Component Analysis (PCA). In this section we discuss a slightly different technique namely Principal Axis Factoring (PAF). Both are methods of exploratory factor analysis, and both aim to summarize a dataset. The SPSS output (and informal interpretation) of both methods is very similar. However, the underlying model is (very) different. Lets start with the differences (in the text the differences with PCA are in **bold**):

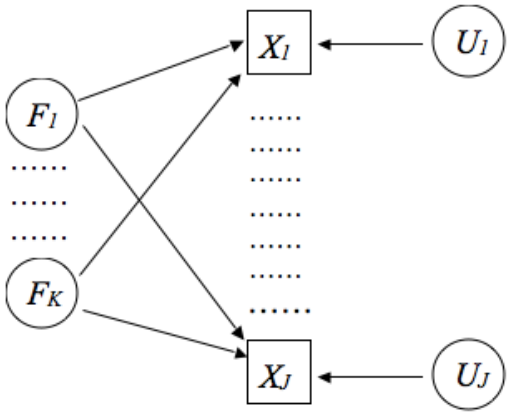
In PAF **Common factors** $F_1, \ldots, F_K$ are *Linear combinations* of **common parts** $\boldsymbol{X_1^*, \ldots, X_J^*}$ **of** observed variables $X_1, \ldots, X_J$:

$$F_k = \boldsymbol{b_{1k}} \boldsymbol{X_1^*} + \ldots + \boldsymbol{b_{J\,k}} \boldsymbol{X_J^*}$$

b_{1k}, …, b_{Jk} we again call the regression weights Similar, variables X_1, …, X_J are regarded linear combinations of **common factors** F_1, …, F_K, plus an **item specific unique factor U_i that represents a non-common part of X_j**:

$$X_j = a_{1j} F_1 + \ldots + a_{Kj} F_K + \boldsymbol{U_j}$$

Here a_{1j}, …, a_{Kj} we again call **factor loadings** (not component loadings this time). Similar to PCA items and factors are standardized. However, the common and unique factors are **latent** as opposed to manifest as in PCA:



The difference between the latent variable specification (PAF) and the manifest variable factor analysis (PCA) becomes most apparent when we split up the variances in X that we observe. We can identify variance that is shared between items, variance that is unique to specific items, and variance that is due to (measurement) error. PCA tries to explain *all* of the variance, while PAF tries to *only explain the variance that is shared between the items*.

While PCA was conducted on the correlation matrix between the items:

1	r_{21}	……..	r_{1J}
r_{12}	1	……..	r_{2J}
……..	……..	……..	……..
r_{J1}	r_{J2}	……..	1

PAF is conducted on the reduced correlation matrix:

h_1^{2*}	r_{21}	……..	r_{1J}
r_{12}	h_2^{2*}	……..	r_{2J}
……..	……..	……..	……..
r_{J1}	r_{J2}	……..	h_J^{2*}

Here the entries on the diagonal are not 1 (which would be the correlation of an item with itself, but it includes all sources of variance) but rather the communalities of the items: the communalities are that part that could be explained by a factor solution and hence is the part of the items that is shared. The real difference in computation is thus that PCA is computed on the observed correlation matrix, while PAF used the reduced correlation matrix. In this way the variance of an item is split into shared variance (captured by the factor) and unique variance which is not captured by the factor solution.

Those were the differences in the model behind PCA and PAF. To show the similarities it is easiest to look at the output of an SPSS PAF analysis on X_1 to X_6 that we used earlier.

Lets first look at the "Total variance explained" table. Note that at the bottom of the table it now states that Principal Axis Factoring is used. Also note that in this new table the last 3 columns (5-7) differ from columns 2-4. This was not the case for PCA. Basically, on the left (columns 2-4), you find the PCA solution. This solution is then used for the PAF solution on the right (columns 5-7). Also note that the variance explained by the PAF solution is lower than that explained by the PCA solution: this is a logical result of the attempt to explain only the shared variance and not the full variance.

Total Variance Explained

Factor	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.802	46.700	46.700	2.276	37.937	37.937
2	1.081	18.015	64.715	.554	9.228	47.165
3	.578	9.639	74.354			
4	.560	9.335	83.690			
5	.522	8.693	92.382			
6	.457	7.618	100.000			

Extraction Method: Principal Axis Factoring.

The factor loadings of a PAF solution are given in the Factor Matrix. As before, these factor loadings (if not rotated using OBLIMIN) give the correlation between a factor and an item.

Factor Matrix[a]

	Factor	
	1	2
X1	.610	.274
X2	.583	.345
X3	.575	.327
X4	.642	-.261
X5	.616	-.268
X6	.665	-.337

Extraction Method: Principal Axis Factoring.

a. 2 factors extracted. 10 iterations required.

We can again compute the communalities using this output: $h_j^2 = r_{XjF1}^2 + r_{XjF2}^2 = a_{j1}^2 + a_{j2}^2$ (For this specific example: $h_1^2 = a_{11}^2 + a_{12}^2 = (0.610)^2 + (0.274)^2 = 0.447$). However, we can also look up the communalities in the SPSS output:

Communalities

	Initial	Extraction
X1	.314	.447
X2	.299	.458
X3	.289	.437
X4	.351	.480
X5	.326	.451
X6	.370	.555

Extraction Method: Principal Axis Factoring.

And finally, as with PCA, we can look at the reproduced correlation matrix:

Reproduced Correlations

		X1	X2	X3	X4	X5	X6
Reproduced Correlation	X1	.447[b]	.450	.440	.321	.303	.314
	X2	.450	.458[b]	.448	.284	.267	.271
	X3	.440	.448	.437[b]	.284	.266	.272
	X4	.321	.284	.284	.480[b]	.465	.515
	X5	.303	.267	.266	.465	.451[b]	.500
	X6	.314	.271	.272	.515	.500	.555[b]
Residual[a]	X1		-8.94E-04	2.588E-03	-2.51E-02	1.096E-02	1.250E-02
	X2	-8.94E-04		-1.45E-03	2.709E-02	-2.47E-03	-2.15E-02
	X3	2.588E-03	-1.45E-03		-4.46E-03	-7.93E-03	1.047E-02
	X4	-2.51E-02	2.709E-02	-4.46E-03		1.237E-03	1.683E-03
	X5	1.096E-02	-2.47E-03	-7.93E-03	1.237E-03		-2.56E-03
	X6	1.250E-02	-2.15E-02	1.047E-02	1.683E-03	-2.56E-03	

Extraction Method: Principal Axis Factoring.

a. Residuals are computed between observed and reproduced correlations. There are 0 (.0%) nonredundant residuals with absolute values > 0.05.

b. Reproduced communalities

In this example the residual correlations are very low, and thus the 2 factor model seems to summarize the data very well. However, the total explained variance of 47.2% is pretty low.

You might now wonder when to use PAF en when to use PCA. As is clear above, the output and interpretation are extremely similar. However, the theoretical ideas are pretty different. In practice, PCA is easier to compute and will work well as a data summary technique. PAF however is “superior” theoretically, and thus often preferred by social science researchers. All surrounding fields use PCA, the social sciences seem to like PAF. Differences in the end are minor.

Assumptions explorative FA

We have almost come to the end of the discussion of factor analysis. To really understand the technique it is important that you practice a lot and try to interpret the output(s). However, before we move to the next topic it is important to discuss a) what the assumptions behind factor analysis are, and b) whey you should and should not use factor analysis.

The assumptions:

Factor analysis (both PAF and PCA) are computed using correlations. This implicitly assumes that correlations are a correct measure to quantify the relationship between the items. Thus, obviously, all of the concerns regarding correlations that we discussed earlier apply also to factor analysis:

- We need an interval measurement level
- The association should be linear
- The distributions of variables should not be too different
- There should be no influential outliers

Thus, in practice, you will have to check whether or not these are actually true in your dataset.

Less explicit are the following assumptions:

- The relationship X_j on the one hand and F_1, …, F_K, U_j on the other hand, is linear: $X_j = a_{1j}F_1 + \ldots + a_{Kj}F_K + (b_jU_j)$

- All unique factors U are uncorrelated: $r_{UiUj} = 0$ for all variables i and j

- All factors F are not correlated with the unique factors U $r_{FkUj} = 0$ for all F_k and U_j

Of which the last two only apply to PAF, since in PCA no unique part of the factors, U, is assumed.

How would you check whether you can do a factor analysis on your data?

Well, you can always perform a factor analysis. However, whether the results make sense and can be interpreted is something different.

Obviously, if all the relationships between items are (e.g.) non-linear, then factor analysis will not work well: factor analysis assumes that correlations are a good measure to describe the associations between items. You can still run a factor

analysis, but its interpretation will be incorrect, and it will not summarize the data well.

Researchers use some rules of thumb to see if the can perform factor analysis. Here are the most popular ones:

- The number of observations that you have needs to be high: $N \geq 100$

- The number of observations that you have needs to be high compared to the number of items: $N \geq 5*J$ (if inter-item correlations are small than there should be at least 10 cases per item)

- The items should be sufficiently correlated:
    - R must contain correlations that are in absolute value greater than 0,3 (Pallant p.180-181)

    - Bartlett's Test of Sphericity must be significant (p-value < 0,05) (Pallant p.180-181). This is a test of the overall correlation matrix.

    - KMO index must be greater than 0,6 (Pallant p.180-181). This is a summary of the correlation matrix

In the end these are just rules of thumb. PCA is used in all kinds of fields, often without these assumptions (Facebook uses PCA to compress images). However, it is a custom in the social sciences to report Bartlett's test and the KMO index.

Confirmatory Factor Analysis

In the previous chapter we discussed exploratory factor analysis (we discussed PCA and PAF as two methods for exploratory factor analysis). With exploratory factor analysis you *explore* the data to see which items group together into a (sub)construct. Exploratory factor analysis can be used to summarize a dataset and interpret possible clusters of items

However, often you might know explicitly the structure of your scale. You might know in advance that items $X_1, \ldots, X_{10}$ are supposed to measure extraversion, while items $X_{11}, \ldots, X_{20}$ measure agreeableness. You thus have a very specific hypothesis about the structure of your constructs. If this is the case then you can use *confirmatory* factor analysis to formally test you assumptions. Here factor analysis is not used to explore the dataset and see which groupings emerge, but rather it is used to test directly a structure that is known in advance.

We will discuss two methods of confirmatory factor analysis: the Multi Group-Method (MGM) and Structural Equation Models (SEM). The first is an informal method but it nicely demonstrates the general idea of confirmatory factor analysis and links this technique explicitly to reliability as defined in classical test theory. The second (SEM) is a formal and very general way of evaluating known structure in a dataset: it can be used for applications that reach far beyond confirmatory factor analysis. We will discuss some of the basic concepts of SEM, but we will not cover all the details: for this you should attend a different course.

Multiple Group-method (MGM)

Lets start with the Multiple Group-Method. The multiple group method is based on the idea that for each (sub) construct you can group the items using classical test theory (e.g. compute cronbach's alpha, and than compute sum scores). Once we have done this we can evaluate whether our assumptions hold by seeing whether:

1. Items that belong to a specific scale indeed correlate highly with that scale (recall convergence, alpha if item deleted, and the corrected item total correlation)

2. Items correlate higher with their own scale then with other scales. If this is not the case then the item might actually "belong" to the other scale and your assumption about the structure is rejected.
3. Items correlate higher with their own scale than scales correlate with each other. This is a measure for divergence.

Let's do a practical example. We will continue with the 6-item dataset introduced as an example for exploratory factor analysis. The correlation matrix looked like this:

Correlation Matrix

		X1	X2	X3	X4	X5	X6
Correlation	X1	1.000	.449	.443	.296	.314	.326
	X2	.449	1.000	.446	.312	.264	.250
	X3	.443	.446	1.000	.279	.258	.282
	X4	.296	.312	.279	1.000	.467	.516
	X5	.314	.264	.258	.467	1.000	.497
	X6	.326	.250	.282	.516	.497	1.000

Now let us assume that we knew a-prior (in advance) that items X_1 to X_3 measured one concept, and X_4 to X_6 measured another concept. Thus the 6 items form to (sub) scales (or measure two (sub) constructs).

The first requirement for MGM is that the scales themselves are good scales, as evaluated using the skills we learned when discussing reliability analysis. Lets check:

Scale $X_{(1)}$: $\alpha = 0.704$

	X_1	X_2	X_3
corrected item-total correlation	0.5244	0.5268	0.5233
α if item deleted	0.6153	0.6111	0.6197

For this scale α is good, the correlations > 0,3, and α decreases if one item is deleted. According to classical test theory this scale is good.

Scale $X_{(II)}$: $\alpha = 0.7448$

	X4	X5	X6
corrected item-total correlation	0.5688	0.5533	0.5920
α if item deleted	0.6634	0.6811	0.6354

For the second scale α is also good, correlations > 0,3, and α decreases if one item is deleted. According to classical test theory this scale is also good.

These two analyses jointly satisfy requirement 1 of the MGM method.

The second analysis concerns the correlations of the items with their own scale. Thus, we compute a sum score for X_1 to X_3 which we will call $X_{(I)}$, the first scale. We also compute a sum score for X_4 to X_6, which we will call $X_{(II)}$, the second scale. Now lets look at the correlations (these are the corrected item total correlations):

	$X_{(I)}$	$X_{(II)}$
X_1	**0.524**	0.383
X_2	**0.527**	0.338
X_3	**0.523**	0.336
X_4	0.373	**0.569**
X_5	0.352	**0.553**
X_6	0.361	**0.592**

We can see that indeed the correlations of items with their own scale is higher than the correlation of items with the other scale. This satisfies the second requirement.

For the third requirement we look at the correlation between the scales:

$$r_{X(I),X(II)} = 0.445$$

This is quire a high correlation, so the two constructs (or scales) are apparently correlated. This might be realistic. However, we should compare the correlation of the

scales with the corrected item total correlation of the items with the scales to check the assumption. Here we see that X_1 to X_3 indeed correlate higher with $X_{(I)}$ than the the correlation between the scales, and the same is true for X_4 to X_6 and $X_{(II)}$. Thus, the third requirement is also satisfied.

This concludes the MGM procedure. We can conclude that we have confirmed our hypothesized structure. If this would not be the case, we might use explorative factor analysis to inform new hypothesis. However, if we want to formally test (e.g. get a *p*-value etc.) of the hypothesized structure we need more than MGM: we move over to SEM.

Structural Equation Models (SEM).

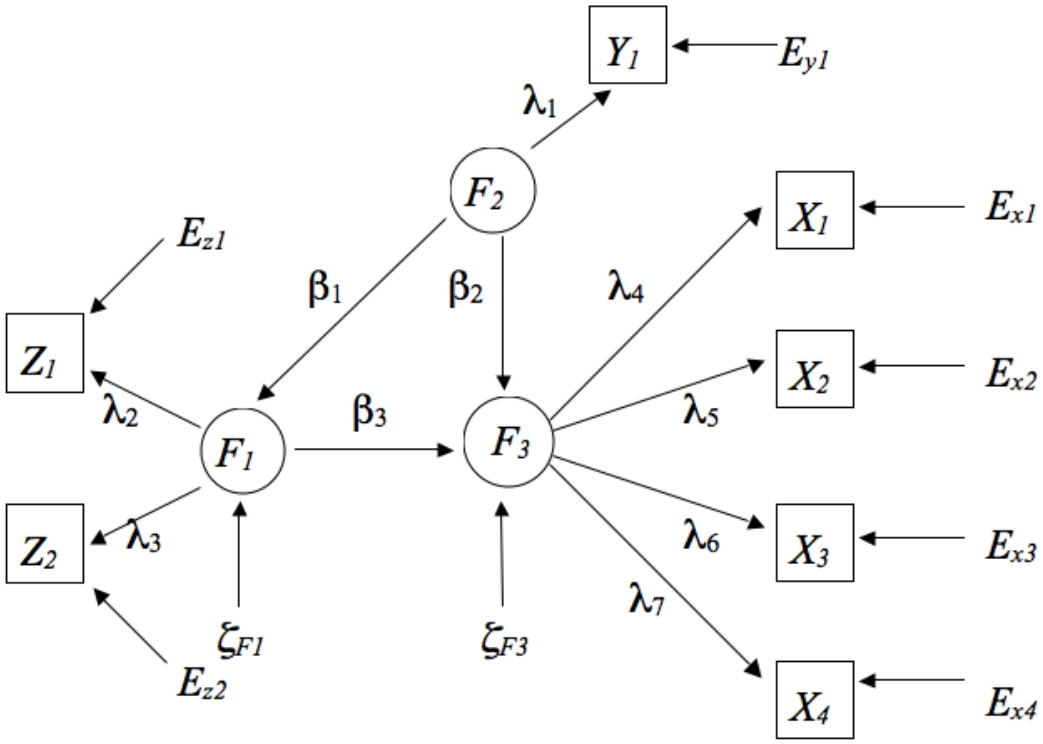
SEM provides a statistical framework for conducting Confirmatory Factor Analysis and other more complex analyses. SEM is also referred to as the analysis of covariance structures. Basically what the method does is check whether the observed data is likely to occur given a pre-specified covariance matrix. The basic steps of a SEM analysis are the description of a covariance structure by the researcher, followed by a test of the fit of the data to this structure. The fit is evaluated using specific "goodness of fit" indices such as Chi-square, the likelihood ration, or information criteria such as AIC and BIC. We will dig into these later.

SEM is a method that is much broader than just confirmatory factor analysis. Many methods that you are used to can be thought of as a special case of a SEM analysis. SPSS will allow you to do some SEM analysis, while specialized computer packages such as Lisrel and AMOS will allow for more applications of SEM. If you are familiar with Path-models, you can think of SEM as a generalization of Path modeling. SEM aims to explain linear dependencies between variables: this is exactly what is described in the covariance matrix.

SEM is powerful, but also dangerous: you need to understand very well what you are doing when interpreting SEM models. It is very easy to have SPSS or AMOS do its computations and derive the wrong conclusions. Thus, if you are unsure about your SEM analysis, always consult an expert. SEM is general, super useful, but also non trivial.

SEM, basic concepts

Let us start with some basic notation and concepts behind SEM models (also called structural models). Again, SEM models concern the analysis of a covariance matrix. However, the relations specified in a covariance matrix can also be specified using a graphical model. A SEM model might look like this:



The following parts can be identified:

- The manifest variables (squares). These are variables that you have observed directly (Z_1, Z_2, Y_1, X_1, …, X_4).
- The latent variables (circles). These are variables that are not directly observed (F_1, F_2, F_3). Latent variables can be split into:
    - Endogenous latent variable and exogenous latent variables. The endogenous latent variables are "predicted by the model" and have an arrow pointing towards them (F1 and F3). These variables are presented together with their error component ζ_F
    - Exogenous latent variables: those latent variables that are not predicted by the model (have no error pointing towards them: F_2). These latent variables do not have an error component.

Often, a distinction is made between the *measurement* model, and the *latent variable* model. The measurement model can be compared to the factor analysis part: it specifies (e.g.) how X_1 to X_4 are used to measure F_3. The latent variable model specifies the relationships between the latent variables F_1, F_2, and F_3. These are the regressions from one latent variable to the other. These can be mixed in a SEM model, and not all SEM models contain both components.

In the figure the latent variable model(s) are:

$$F_1 = \beta_1 F_2 + \zeta_{F1}$$
$$F_3 = \beta_2 F_2 + \beta_3 F_1 + \zeta_{F3}$$

Where F_1, F_2, F_3 are latent variables (circles), β_1, β_2, β_3 are the effects between the latent variables, and ζ_{F1}, ζ_{F3} are the prediction errors.

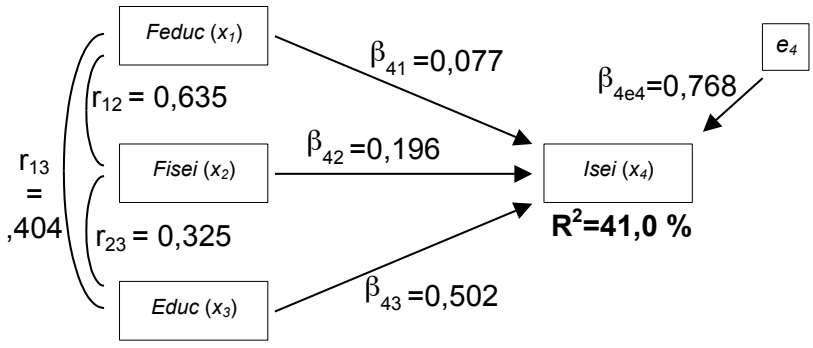
Every latent variable is measured using several manifest variables: X_1, X_2, X_3, X_4 measure F_3, Y_1 measures F_2, and Z_1, Z_2 measure F_1. Indicated by λ_1, …, λ_7 are the effects of the manifest variables on their latent variables. The E's indicate the measurement errors of the manifest variables.

So far for the jargon.

Applications of SEM models

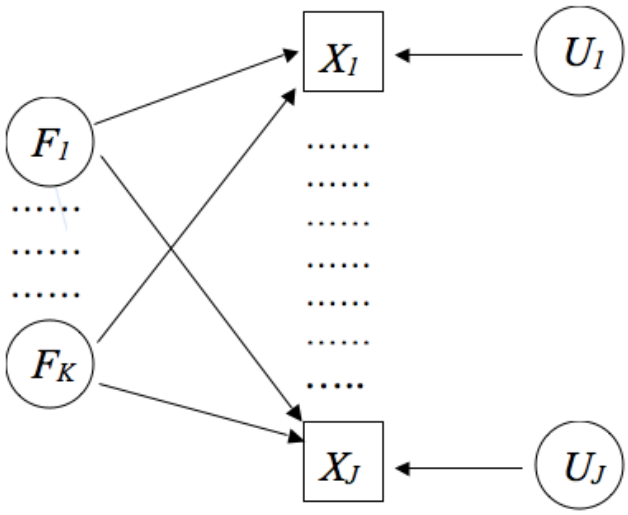
As introduced, SEM models are a very general tool, and actually capture a number of the things you have covered in previous courses. Some specific SEM models you can fit in SPSS, for some you need more specialized software. Some examples:

Linear regression can be thought of as a special case of SEM. For example:



Here all of the variables are manifest (and there is no separate measurement model). This you can easily compute using SPSS. (Note that the β's in this linear regression example above have nothing to do with the β's in the earlier latent variable model.)

Exploratory factor analysis (and off course confirmatory factor analysis – that is why we are discussing SEM anyway) can be thought of as a SEM model. In this case, the SEM model only contains a measurement model. The following is a PAF model:



While we can fit this model (explorative) in SPSS, SPSS will not allow you to specifically test a specific model. The difference is that with a specific model you (manually) set some relations to be present (eg. F_1 -> X_1), and some to be absent (eg F_1 -> X_{10}. SPSS will assume all the relations are present.

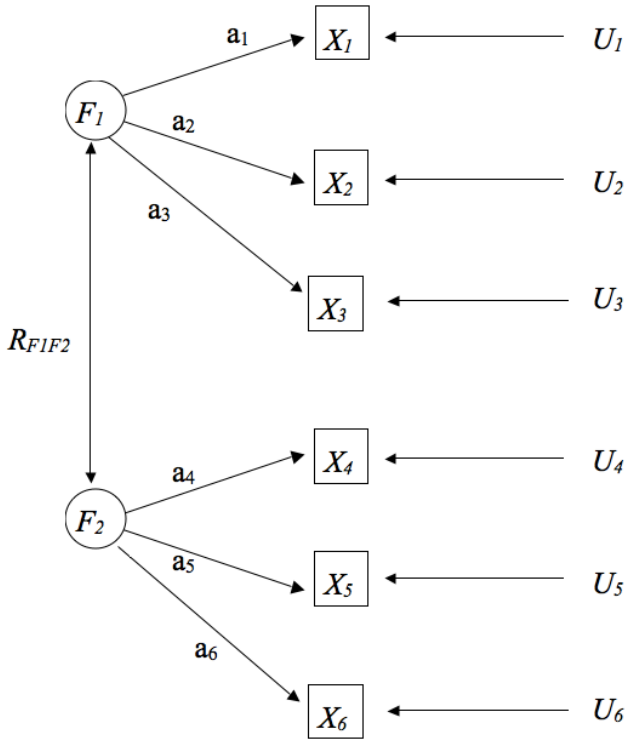
Some models / methods that you have encountered are not special cases of SEM models. Logistic regression for example, is not a SEM model since the SEM assumption of normally distributed errors is not valid for logistic models.

Note that SEM models rely on analyzing the covariance structure of the matrix. Thus, all the caveats we have discussed for factor analysis models apply: There needs to be a linear association between items (and items and factors), and you need a sufficiently large N (e.g. $N > 100$, or $N > 5*J$).

Confirmatory factor analysis using SEM.

We will now discuss a confirmatory factor analysis model that is fit using AMOS. This is to give you some insight in how this is done, and we will discuss the goodness of fit indeces that are most often used to evaluate models. We will also discuss how you would compare competing models. Since we are not going to cover a new software package, we will stick to the interpretation, and we will in this course not cover how you would actually run a model yourself. I just hope this discussion makes that you can at least read papers that use SEM for confirmatory factor analysis, and that you can be critical of their interpretation.

So, lets do an example of confirmatory factor analysis using X_1 to X_6 that we have used before. We are going to assume that X_1 to X_3 represent one latent construct, and that X_4 to X_6 represent a second latent construct. Graphically this model looks like this:



The model states that two latent factors underly the 6 manifest items. Each item has its own unique variance component (PAF), and there is a correlation between the factors (recall OBLIMIN rotation). The structural equation of this model are:

$$X_1 = a_1 F_1 + U_1 \qquad X_4 = a_4 F_2 + U_4$$
$$X_2 = a_2 F_1 + U_2 \qquad X_5 = a_5 F_2 + U_5$$
$$X_3 = a_3 F_1 + U_3 \qquad X_6 = a_6 F_2 + U_6$$

The next step in a SEM analysis is to check the fit of the model. In general the aim is to specify a model that is simple, but still has a good fit. One measure for fit is chi square, χ^2, which is easier to remember as a measure of misfit: the higher the value of χ^2 the worse is the model fit.

A measure for complexity is the "degrees of freedom" of model (often *df*). The larger the degrees of freedom, the simpler the model. So, we should look for models with a low χ^2 value, and a high number of degrees of freedom.

The value of χ^2 you can see in the output of most statistical packages for SEM analysis. The *df* however you can easily compute yourself for confirmatory factor analysis. It is a function of the *true underlying complexity*, and the *complexity of the model that you are fitting*.

The *true underlying complexity* is given by the number of unique elements in the covariance matrix of manifest variables. In our example we have 6 manifest items. The covariance matrix thus contains 6*6=36 entries. However, $COV(X_1, X_2) = COV(X_2, X_1)$, so we have less unique elements. The number of unique elements is given by:

*Unique elements of J*J COV matrix = J (J+1) / 2*

In our example this is 6*7 / 2 = 21.

The *model complexity* is given by the number of parameters of the model. These are basically all the arrows in the model. For our example we have 3 arrows from F_1 to $X_1,\ldots,X_3$, plus 3 arrows from F_2 to X_4, …, X_6. Furthermore we have 6 arrows for the item unique parts, and we have 1 correlation. Thus:

6 (factors to items) + 6 (unique variances) + 1 (correlation) = 13

The *degrees of freedom* in the example now is given by:

$$\underline{Df = \text{true complexity} - \text{model complexity} = 21 - 13 = 8}$$

Note that if df = 0, than the model fit will be perfect: we are not reducing the data in any way. From AMOS we can get the following output:

```
Chi-square = 3.817
Degrees of freedom = 8
Probability level = 0.873
```

This we can use to do a χ^2 test. The null hypothesis of this test is that the model fits the data perfectly. The *p*-value (`probability level` in the output) of the hypothesis test shows that this hypothesis should not be rejected ($p > .05$), and thus we accept that the fit is perfect.

The χ^2 test is one possible test for the fit of the model to the data. However, it has a negative property: If *N* (the number of observations) is large, than the value of χ^2 will be large: indicating misfit. This is a negative property since you would like to be able to accept models if you have more data. However, it is a logical consequence of the way χ^2 is computed: a little "misfit" with lots of observations makes that you are sure that there is some "misfit". The model is not perfect. However, often we do not want to know whether the model is perfect, but rather whether its "good enough".

This motivates another measure for misfit that is widely used in SEM analysis called the RMSEA (Root Mean Square Error of Approximation). We will not dig into the details, but basically low values of RMSEA indicate a good fit. Often a model is considered good if the RMSEA < .05. There also exist an hypothesis test for the RMSEA (in AMOS this is denoted a `PCLOSE`). It test the following hypothesis:

$$H_0: \text{RMSEA population} = 0{,}05 \quad [\text{model fit COV ok}]$$
$$H_1: \text{RMSEA} > 0{,}05 \quad [\text{model fit COV not ok}]$$

Thus, if *PCLOSE* < *.05* (if we take the standard .05 value for alpha of a hypothesis test) then we reject the null hypothesis and conclude that the model fit is not ok.

Besides testing the overall model fit, we can also test the estimated effects of the model. Lets look at the estimated effects for our example data when we use AMOS. The estimate factor loadings are:

	Estimate	S.E.	C.R.	P	Label
X1<---F1	.748	.067	11.13	*	a1
X2<---F1	.702	.066	10.68	*	a2
X3<---F1	.635	.060	10.57	*	a3
X4<---F2	.759	.064	11.91	*	a4
X5<---F2	.681	.060	11.35	*	a5
X6<---F2	.779	.063	12.37	*	a6

The estimated correlation between the factors is:

	Estimate	S.E.	C.R.	P	Label
F1<-->F2	.611	.060	10.16	*	par_13

And the estimated variance components are:

	Estimate	S.E.	C.R.	P	Label
F1	1.000				
F2	1.000				
U1	.624	.077	8.08	*	e1
U2	.637	.074	8.65	*	e2
U3	.539	.061	8.77	*	e3
U4	.585	.070	8.33	*	e4
U5	.556	.062	8.96	*	e5
U6	.532	.069	7.73	*	e6

Note that for each a P is given indicating with a "*" whether the *p*-value of the test whether or not the estimate is equal to 0 is smaller than .05. This would indicate that all estimated effect (e.g.) for the factors to the items are distinct from 0.

While all of these tests are useful to evaluate the model, the fact that a model fits well (e.g. the null hypothesis that the RMSEA < .05 is not rejected) does not mean our model is the true model: there might be other models that fit the data just as well, or even better. The true power of SEM is in model comparisons: we can compare

simple models to ever more complex models. Here we do a χ^2 on the difference between two (nested) models where:

$$\chi^2_{difference} = \chi^2_{simple} - \chi^2_{complex}$$
$$df_{difference} = df_{simple} - df_{complex}$$

Models are nested when the simple model can be obtained by deleting parameters (arrows) from the complex model. In our example we could for example check whether or not deleting the correlation between the factors affects the model fit. The logic here is that a model without the correlation is simpler, and if it still fits the data well than we might prefer the simple model. We could also test whether for example it suffices to simplify the model by including what is called *parallel* items. This means that the factor loadings and the item variance are set to be equal to each other for each scale. Formally:

$$a_1 = a_2 = a_3$$
$$a_4 = a_5 = a_6;$$
$$var(U1) = var(U2) = var(U3)$$
$$var(U4) = var(U5) = var(U6)$$

This model would only have 5 parameters (one for a_1 to a_3, one for a_4 to a_6, one for VAR(U_1) to VAR(U_3), one for VAR(U_4) to VAR(U_6), and one for the correlation between the factors. From AMOS we can obtain the χ^2 of this simpler model: 11.695. This we can use to compare the models:

$$\chi^2_{difference} = 11.695 - 3.817 = 7.878$$
$$df_{difference} = 16 - 8 = 8$$

In the above formula for the df of the simple model are given by *21 (true complexity) – 5 (model complexity) = 16*.

The decrease in model fit is not significant (you can look up the p-value for a χ^2 test with $\chi^2 = 7.878$ and df=8. This indicates that the simple model is not any worse than the complex model and thus would be preferred. By incrementally comparing

different SEM models we can get a better and better understanding of the true model that generated our data.

From the above discussion of SEM you will need to understand the general concepts. We will not cover how to do SEM analysis yourself. You should be able to interpret χ^2 tests and the RMSEA, and understand that we can build better and better models by comparing nested models. That's it.

The SEM warning!

Three things that are often overlooked when using SEM, but are very important, make clear why SEM is both powerful and dangerous:

1. If you look at a graphical model, then the effect of the arrows you do not draw is as big (if not bigger) than those you do draw. If you do not draw an arrow you are setting that covariance to 0. That is a very explicit assumption!
2. Many graphical models that might look very different end up encoding the exact same covariance matrix. This is partly due to the fact that the direction of the arrows is not specifically encoded in the covariance matrix. So, models that might look very different graphically might be the same mathematically.
3. If your model has a good fit (RMSEA < .05), than this does not mean that your model is correct. It only means that the data is not unlikely given the model. However, the data might also not be unlikely given a different model. So it strengthens your confidence that your model is useful but you can never confirm that your model is the only true model.

Again, if you want to do a SEM analysis and are unsure about what you are doing, consult an expert.

Cluster analysis (CA)

Slowly but surely we have arrived at our last topic of the course: Cluster analysis. Cluster analysis is, like factor analysis, a data summarization technique. So, at a really high level its comparable to factor analysis. However, the approach to summarizing data in cluster analysis is distinct from factor analysis. While with factor analysis we tried to group items into (sub)scales, cluster analysis does is usually not used to cluster items, but rather to cluster people. The aim of cluster analysis is to find groups of people who have similar scores on a number of questionnaire items. So, in some way you can think of factor analysis as grouping the columns (questions) in an SPSS dataset, while cluster analysis aims to group the rows (persons). Note that another difference between the two methods is that while factor analysis gives us continuous factor scores for each person, cluster analysis will create discrete scores: each person will be assigned to one specific cluster.

Cluster analysis is an umbrella term form numerous different algorithms which all aim to group together similar units (in the social sciences units are often people). It is used widely in all branches of science. We will cover two examples of methods of two distinct classes of clustering:

- Hierarchical clustering: With hierarchical clustering we start by putting together the two most similar people into a cluster. Next, we put together the second two closest units, etc. etc. All the way until we end up with one giant cluster. We will cover a version of hierarchical clustering called Ward's method.
- Non-hierarchical clustering: With non-hierarchical clustering we choose a number of cluster *k* in advance, and then split up the dataset into *k* groups where the aim is to put individual together in a cluster who are as similar as possible.

Cluster analysis is not valued highly in the social sciences since it is relatively a-theoretical. And, many different clustering methods exists, each often leading to different outcomes. However, despite these difficulties clustering can be useful both for summarizing as well as interpretation.

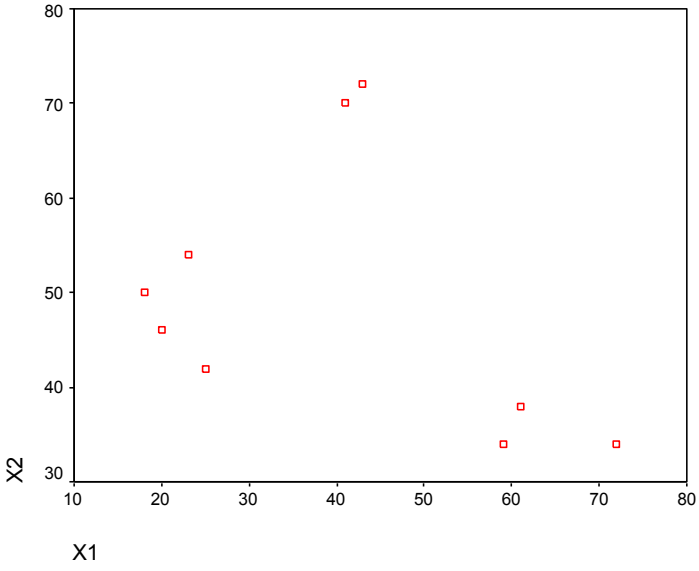
Cluster analysis by example: Ward's method

We will start with a hierarchical clustering method called Ward's method. We will run through a very small example dataset to make all the concepts clear.

We will start with a very small dataset containing the observations of 9 people on two items, X_1 and X_2. The data looks like this:

Person	X_1	X_2
1	18	50
2	20	46
3	23	54
4	25	42
5	41	70
6	43	72
7	59	34
8	61	38
9	71	34

It is hard to see directly from the data how we would group people together. However, if we plot the data the structure is immediately clear:



The data clearly contains three clusters. On the top of the plot are person 5 (X_1=41, X_2=70) and 6 (X_1=43, X_2=72). These cluster together. These two people are close together, and distinct from the others. Similarly for person 1 to 4, and person 7 to 9.

While we can see the clustering clearly in this simple dataset, it is often very hard to see clusters in real datasets. If you are clustering based on more than 2 variables it is hard to plot the data, and if you have 100's of people, the clustering might not be clear immediately. Finally, to have a computer cluster the data we need more than just our eyes: we need to formalize what we mean by close together. For this we need a measure of equality between the persons.

Cluster algorithms can differ on their measure of equality. Here we will discuss Ward's method, and later on we will discuss k-means clustering. These both use the same distance measure. And, both of these methods can be performed in SPSS.

Measuring distance

Ward's method works by starting with *N* clusters (as many as we have people) and subsequently merge 2 clusters that are close together to form a single cluster (step 1). Next, the second closest clusters are merged (step 2), all the way to step *N-1*, at which only a single cluster is left.

Which clusters are closest to each other (and thus should be merged) at each step is relatively simple to define mathematically (although it looks challenging). The clusters for which the increase in so-called SS_w (the Sum of Squares Wards method) is minimized will be merged. Here is the specification of SS_w:

$$\sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{i=1}^{n_K}(X_{ijk}-\bar{X}_{jk})^2$$

Let's dissect this formula. In parentheses it states the difference between X_{ijk} and $\bar{X}_{jk}$. X_{ijk} is the score of person *i* in cluster *k* on item *j*. At the first step, every person will be in his or her "own" cluster (there are as many clusters as people), so this will initially just be the score of person *i* on item *j*. $\bar{X}_{jk}$ is the mean of variable *j* in cluster *k*. So (X_{ijk} - $\bar{X}_{jk}$) is the distance of a person to the cluster mean on a specific item. This distance is squared, since we care about the absolute distance (and since squaring gives

equations that are easier than taking the absolute value). So, we now have the squared distance of individual *i* on item *j* in cluster *k* from the mean score on *j* in cluster *k*. (Try to bear with me…).

Finally, to compute the SS_w, this distance is summed over all the individuals in cluster k, then summed over all items, and than summed over all clusters.

This might have been tricky, so let's examine the steps for our small data example:

At step 0: There are 9 clusters. Everyone is their own cluster, and thus the difference of the score of the person from the mean score on that cluster is 0. Summed over all items and all clusters this gives $SS_W = 0$

At step 1: We want to make 8 clusters, so we should add the two persons together that are the closest to each other. It looks like these are persons 5 and 6 (i=5 and i=6). If we indeed put these two together, then the mean score of this cluster (5 and 6 together) will be 41+43/2 =42 for X_1, and 70+72/2=71 for X_2. Thus $\bar{X}_{1k} = 42$, and $\bar{X}_{2k} = 71$. The total SS_w when putting these two together is:

$$SS_W = [(41-42)^2 + (43-42)^2] + [(70-71)^2 + (72-71)^2] + 0 + \ldots + 0 = \mathbf{4}$$

where the first term is the difference of i=5 on X_1 (=41) with the cluster mean of X_1 (=42). The second term is the difference between i=6 on X_1 (=43) and the mean on X_1 (=42). Next, this is repeated for X_2: 70-71 for i=5, and 72-71 for i=6. Since these are the only two people that are clustered together, all the other people will still be in their "own" cluster (as in step 0), and thus there will be no difference between the mean of that cluster and their own score. All others thus contribute 0 to the total SS_w.

A computer would normally not be able to "see" that *i=5* and *i=6* should be clustered together. What actually happens is that the computer tries all possible options (1&2 together, 1&3 together, etc. etc.) and computes the SS_w each time. It than looks for which clustering the SS_w is the smallest and makes that cluster. You should convince yourself that any other cluster adding two people together will lead to an SS_w that is larger then 4.

Step 2: We now want to make 7 clusters. The two closest are $i = 7$ & $i = 8$. Now $\bar{X}_{1k} = 60$ and $\bar{X}_{2k} = 36$. The total SS_w when adding these two units together is:

$$SS_W = [(59-60)^2 + (61-60)^2] + [(34-36)^2 + (38-36)^2] + 4 + 0 + \ldots + 0 = 10+4 = \mathbf{14}$$

here the first terms are the difference for i=7 to the cluster mean on X1, then i=8 on X1, etc. etc. Just like above. The 4 comes from the previous clustering we made at Step 1. The total SS_w now is 14, and you should again convince yourself that making another cluster than 7 and 8 would lead to a higher SS_w.

Step 3: At step 3 we merge 1 & 2. The total $SS_W = 14 + 10 = \mathbf{24}$

We now skip some steps and get to the last step (step *N-1*):

Step 8: Here we merge two clusters which include multiple people. The first includes (1,2,3,4,5,6) and the second includes (7,8,9): $SS_W = 1487{,}333 + 3371{,}111 =$ **4858,444.**

I hope this example gave you some intuition for how the clustering is done. However, in practice you would obviously not compute the SS_w by hand, you use a computer.

Ward's method in SPSS

Lets use SPSS to compute Ward's clustering for our example dataset. One of the most important tables that you will get in the output is the agglomeration schedule:

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	5	6	4.000	0	0	7
2	7	8	14.000	0	0	6
3	1	2	24.000	0	0	4
4	1	3	58.667	3	0	5
5	1	4	123.000	4	0	7
6	7	9	221.667	2	0	8
7	1	5	1487.333	5	1	8
8	1	7	4858.444	7	6	0

This table shows, at each step in the clustering process, which clusters are merged. In the column "coefficients" you can find the SS_w at each step.

The "Clusters Combined" columns give you insight into which clusters are joined together. In the first step *i=5* and *i=6* are clustered. This gives an SS_w of 4. In the next step 7 & 8 are clustered. This gives a total SS_w of 14. Next 1&2 are clustered.

At step 4 clusters 1 & 3 are clustered. Note that cluster 1 now already contains 2 persons since these were joined at step 3. SPSS will refer to a cluster with multiple people using the lowest person number. You can also see this in the "Stage Cluster First appears" columns: in the first 3 steps these contain only 0's: we are only adding people together. At step 4 you see that cluster 1 appeared earlier, namely at step 3: this is when *i=1* and *i=2* were clustered. The final column indicates when a cluster that has been formed is used again. Here you can (e.g.) see that the first cluster (*i=5* & *i=6*) is only added to the other cluster at a very late stage: stage 7.

The agglomeration schedule thus gives a full overview of the clustering steps. Ward's method is called hierarchical clustering because this stepwise method.

Note that the SS_w in the agglomeration schedule increase at each step. This is required, because clustering will lead to an increase in the distance. However, also note that the increase between SS_w's at subsequent steps increases (or is equal): from step 0 to 1 the SS_w increase by 4. Then, it increases by 10. Next it increases by 10 again. Then, it increases by 34.6. etc. etc. This increase will always keep increasing: if it did not than two other clusters should have been merged earlier.

A second useful table of the Ward cluster analysis output is the cluster membership table:

Cluster Membership

Case	8 Clusters	7 Clusters	6 Clusters	5 Clusters	4 Clusters	3 Clusters	2 Clusters
1	1	1	1	1	1	1	1
2	2	2	1	1	1	1	1
3	3	3	2	1	1	1	1
4	4	4	3	2	1	1	1
5	5	5	4	3	2	2	1
6	5	5	4	3	2	2	1
7	6	6	5	4	3	3	2
8	7	6	5	4	3	3	2
9	8	7	6	5	4	3	2

This table shows at each step which cases (people) belong to which cluster. Here you can see that when there are 8 clusters, cases 5&6 belong to the same cluster. If there are 7 clusters, then 7&8 also belong to the same cluster. Etc. etc.

However, this raises the very obvious question: how many clusters should I choose?

Selecting the number of clusters

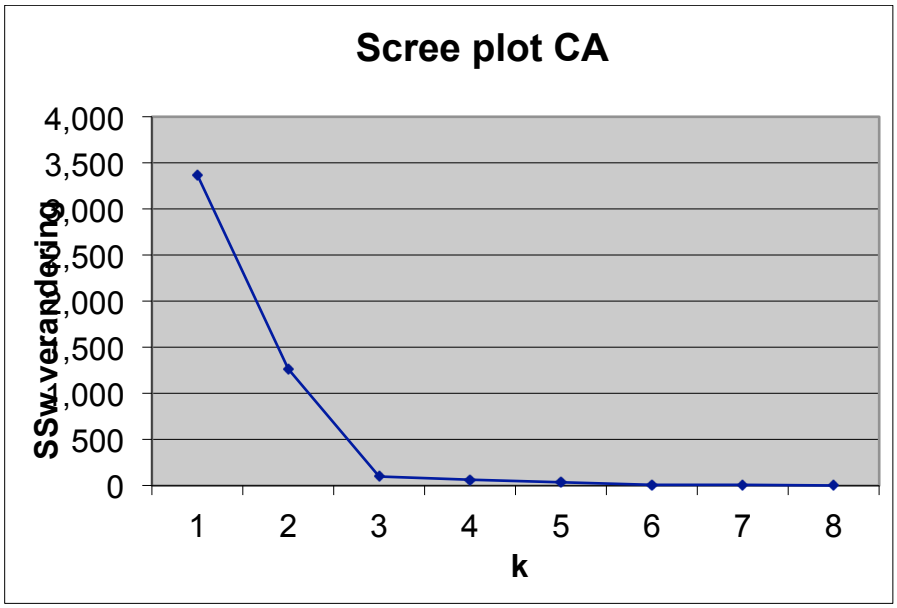
There are basically two ways of determining the number of clusters. Both are based on the SS_w. First, and this is the one I find most useful, you can look at a plot of the *difference* in SS_w as the number of clusters increases. For our example the differences are:

$$SS_{W1} - SS_{W2} = 3371{,}111$$
$$SS_{W2} - SS_{W3} = 1265{,}666$$
$$SS_{W3} - SS_{W4} = 98{,}667$$
$$SS_{W4} - SS_{W5} = 64{,}333$$
$$SS_{W5} - SS_{W6} = 34{,}667$$
$$SS_{W6} - SS_{W7} = 10$$
$$SS_{W7} - SS_{W8} = 10$$
$$SS_{W8} - SS_{W9} = 4$$

If you read these from the bottom up then you see that the steps are small at first: 4 to 10 to 10 to 34, etc. However, when we move from 3 clusters to 2 clusters ($SS_{w2} - SS_{w3}$) there is suddenly a very large increase. This indicates that at this step we are adding together two clusters that are *very* different. This makes sense given our

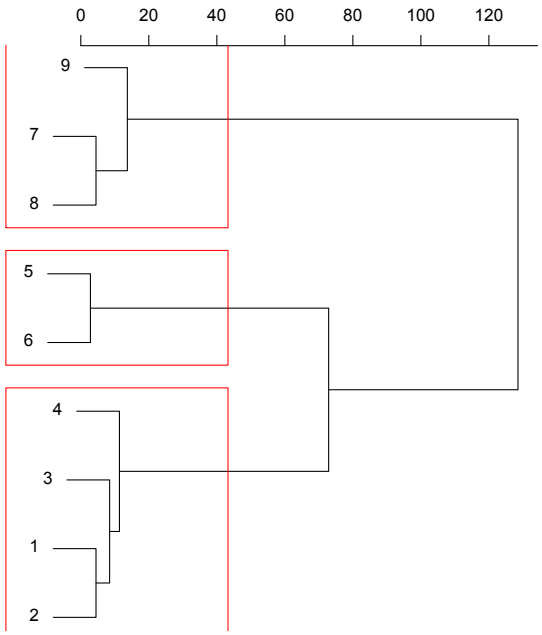
example data: the moment we move from 3 clusters (which are clearly visible) to 2 clusters we are adding together two distinct clusters. This is something you do not want to do: we want to have clusters that are similar to each other, thus we do not want to add two very distinct cluster together. This means we want to stop our clustering right before a large increase in the difference in SS_w.

A plot of these differences looks like this:



You can see that the line is very flat for *k=3* to *k=8*. At *k=2* there is suddenly a large increase. Since we do not want to have this increase we should stop clustering right before the increase: we should stop when we have 3 clusters.

This same logic also applies to a second method of deciding on the clustering: the *dendrogram*. This graphical representation of the hierarchical clustering procedure shows on the x-axis the SS_w and on the y axis there are entries for each of the clusters. It is presented below. Here what we look for is the first time that the horizontal lines in the dendrogram are long. Initially, when persons that are close together are clustered the increased SS_w is small and thus the horizontal lines are short. Then, when we combine dissimilar clusters, the lines become long. If you look from left to right for the first occurance of long horizontal lines, and the count the number of lines, this gives you the number of clusters you should select. In our example the outcome is again clearly 3 clusters.



In the above dendrogram the three clusters are indicated using red squares. Note that dendogram's, while very simple in our example, are often very hard to interpret. If you have many persons, and if the clustering is not so clear-cut as in our example, it is often really hard to tell what number of clusters you should use.

Interpreting the clusters

Interpreting a cluster solution can be done by computing summary statistics of the clusters. Once we have decided we are going to use 3 clusters on our example data, we can add the cluster membership to our dataset:

Person	X1	X2	**Cluster**
1	18	50	**1**
2	20	46	**1**
3	23	54	**1**
4	25	42	**1**
5	41	70	**2**
6	43	72	**2**
7	59	34	**3**
8	61	38	**3**
9	71	34	**3**

Subsequently we can analyze our dataset using the cluster memberships as a new variable. We can for example run a standard ANOVA using the cluster memberships as a factor:

Report

Mean

Ward Method	X1	X2
1	21.5000	48.0000
2	42.0000	71.0000
3	64.0000	35.3333
Total	40.2222	48.8889

ANOVA Table

		Sum of Squares	df	Mean Square	F	Sig.
X1	Between Groups	3104.556	2	1552.278	72.199	.000
	Within Groups	129.000	6	21.500		
	Total	3233.556	8			
X2	Between Groups	1532.222	2	766.111	49.604	.000
	Within Groups	92.667	6	15.444		
	Total	1624.889	8			

The output shows the differences in the means of the clusters, and it shows these differences are statistically significant. Don't get too excited about this: you would expect clear differences between the clusters because that is exactly why you made the clusters the way they are. However, you can use the means of the clusters to describe the people in that cluster (For example, cluster 1 contains people who score low on X_1 and average on X_2). Obviously, once you have the clusters, you can also examine the differences between the clusters on variables that you have not used for the clustering.

An example on a larger dataset

Lets run through Ward clustering on a larger dataset. Here we run Cluster analysis on five scales 'risk aversion' (riskav), 'sensitivity to others' (sens), 'tiredness' (moe), 'extent to which you have secrets' (geheim), 'satisfaction' (tevreden).

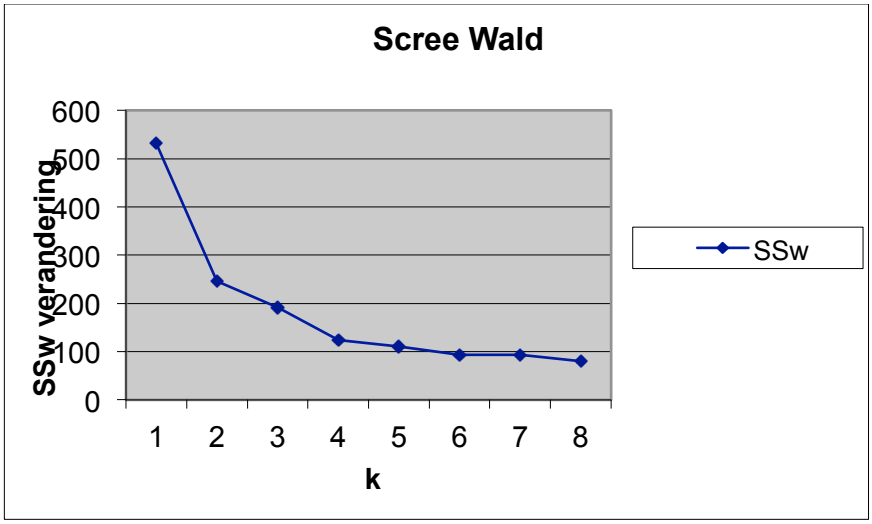
The first step in the actual analysis would be to standardize the scores on these scales. This is very important since if you do not standardize the scores then those items that have a larger range (e.g. run from 1 to 100) will contribute more to the

distance measure (SS_w) than those that have a small range (1 to 10). If you do not standardize then the variable with the large range will have a much large impact on the cluster solution. Here are some overviews of the data (before standardization):

Descriptive Statistics

	N	Range	Mean	Variance
GEHEIM	591	20.00	11.4873	15.057
TEVRED	591	16.00	18.5431	7.360
MOE	591	19.00	10.1303	8.202
SENS	588	50.00	60.8010	77.151
RISKAV	590	5.00	2.6593	1.213
Valid N (listwise)	587			

When we run Ward clustering, we can look at the agglomeration schedule (but it will be very big since we have 591 units to cluster!), but its easier to look straight at the scree plot:



Since the first real increase is from 3 to 4 clusters, we should choose 4 clusters. We can then ask SPSS to add the cluster memberships to our datafile and look at the descriptives:

Report

Ward Method		Zscore(GEHEIM)	Zscore(TEVRED)	Zscore(MOE)	Zscore(SENS)	Zscore(RISKAV)
1	Mean	-.2652772	.3768041	-.1482899	.5631699	-.2422328
	N	214	214	214	214	214
2	Mean	.8146644	-.9016471	.8920021	.0951029	.2047608
	N	165	165	165	165	165
3	Mean	-.2567399	.2605419	-.7313836	-.9653922	-.8580208
	N	112	112	112	112	112
4	Mean	-.4853035	.3795671	-.3692130	-.3070338	1.1888606
	N	96	96	96	96	96
Total	Mean	.0039290	-.0042871	-.0032591	-.0023660	-.0000341
	N	587	587	587	587	587

This makes clear that (e.g.) cluster 1 contains people who score very high on sensitivity to others (compared to the other clusters), while cluster 4 contains people that are very risk averse.

We can also look at differences between clusters on variables we did not consider for the clustering. For example, we can look at differences between clusters on sickness ('ziek'), depression ('depres'), self-awareness ('self-aw'), attention to feelings ('gevoel'), and importance social contacts in a sport ('sport2'):

Report

Mean

Cluster Number of Case	DEPRES	ZIEK	GEVOEL	SELFAW	SPORT2
1	10.2403	5.0516	45.9484	27.4416	11.1415
2	13.7931	6.9569	45.5259	24.5948	11.4722
3	10.1934	5.2028	46.5755	28.2275	11.3851
4	11.4231	5.6538	46.1923	26.2621	11.5571
Total	11.1365	5.5894	46.1346	26.9521	11.3667

ANOVA Table

		Sum of Squares	df	Mean Square	F	Sig.
DEPRES * Cluster Number of Case	Between Groups	1139.478	3	379.826	60.504	.000
	Within Groups	3653.600	582	6.278		
	Total	4793.078	585			
ZIEK * Cluster Number of Case	Between Groups	293.866	3	97.955	31.723	.000
	Within Groups	1800.188	583	3.088		
	Total	2094.055	586			
GEVOEL * Cluster Number of Case	Between Groups	89.912	3	29.971	.740	.528
	Within Groups	23604.456	583	40.488		
	Total	23694.368	586			
SELFAW * Cluster Number of Case	Between Groups	1073.724	3	357.908	18.327	.000
	Within Groups	11326.934	580	19.529		
	Total	12400.658	583			
SPORT2 * Cluster Number of Case	Between Groups	8.770	3	2.923	.341	.795
	Within Groups	3468.217	405	8.563		
	Total	3476.988	408			

It is clear form this analysis that the clusters differ significantly on their depression, sickness, and self-awareness scores.

As a final note before we more over to k-means clustering: Ward's method often works best when the number of units to cluster is relatively small (say $N < 50$). For large datasets, different methods are often faster and more easy to interpret. K-means clustering, which we will discuss next, is a method that works very well on large datasets (however, it is non-hierarchical!)

K-means clustering

K-means clustering is a non-hierarchical form of clustering. Here, you as the analyst decide on the number of clusters before you run the analysis (obviously you can try multiple numbers of clusters, but still, you have to specify them). It is non-hierarchical since it does not aim to add together the closest individuals (or clusters) at each step like Ward's method, but rather it tries to "place" the cluster centers for the number of clusters that you selected in such a way that the distances within the clusters are as small as possible (and the between cluster distances are large). K-means uses the same distance measure as Ward's method (the Sum of Squared differences from the cluster means) to decide on the location of the cluster means.

K-means is an iterative method in the sense that it will try to place the k cluster centers you specified in the plot, and then compute the SS. Then, the algorithm tries to move the cluster centers and see if the total distance decreases. The algorithm will do this until the distance fails to decrease and then it will stop. Once the algorithm stops it knows where the cluster centers are, and which people "belong" to that cluster.

K-means in SPSS

Suppose we take our simple example of 9 people again, and now run k-means clustering in SPSS. We already know from the Ward's analysis that we want 3 clusters so we choose *k=3*.

The first table in the output tells you where SPSS initially placed the cluster centers:

Initial Cluster Centers

	Cluster		
	1	2	3
X1	18.00	72.00	43.00
X2	50.00	34.00	72.00

Next, in a very long table, you can see the iterations of the algorithm:

Iteration History[a]

Iteration	Change in Cluster Centers		
	1	2	3
1	3.225	6.083	.943
2	.645	1.521	.314
3	.129	.380	.105
4	.026	.095	.035
5	.005	.024	.012
6	.001	.006	.004
7	.000	.001	.001
8	4.128E-05	.000	.000
9	8.256E-06	9.282E-05	.000
10	1.651E-06	2.320E-05	4.790E-05
11	3.302E-07	5.801E-06	1.597E-05
12	6.605E-08	1.450E-06	5.322E-06
13	1.321E-08	3.626E-07	1.774E-06
14	2.642E-09	9.064E-08	5.914E-07
15	5.284E-10	2.266E-08	1.971E-07
16	1.057E-10	5.665E-09	6.571E-08
17	2.114E-11	1.416E-09	2.190E-08
18	4.226E-12	3.541E-10	7.301E-09
19	8.434E-13	8.852E-11	2.434E-09
20	1.719E-13	2.213E-11	8.112E-10
21	3.178E-14	5.532E-12	2.704E-10
22	1.005E-14	1.383E-12	9.013E-11
23	.000	3.458E-13	3.005E-11
24	.000	8.644E-14	1.002E-11
25	.000	2.930E-14	3.336E-12
26	.000	.000	1.110E-12
27	.000	.000	3.769E-13
28	.000	.000	1.206E-13
29	.000	.000	4.019E-14
30	.000	.000	2.010E-14
31	.000	.000	7.105E-15
32	.000	.000	.000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 32. The minimum distance between initial centers is 33.302.

At each step in the algorithm the cluster centers move. At step 22 the center of cluster 1 stops moving, and after 32 steps all clusters have stopped moving. Now SPSS can give you the final cluster centers:

Final Cluster Centers

	Cluster		
	1	2	3
X1	21.50	64.00	42.00
X2	48.00	35.33	71.00

Note that, as compared to Ward's method the cluster numbering changed: the numbers are used only on a nominal scale and the size of the numbers thus has no meaning. Also note that the cluster centers for Ward's method and for k-means in this example are the exact same. However, this need not be the case. Because this example is simple and both methods use the same distance measure we obtain the same solution. However, due to the hierarchical structure of Ward, and the non-hierarchical approach of k-means, the solutions will not always be the same. K-means is better able to minimize the distance since it does not have the "legacy" of previously grouped clusters.

Besides the above tables running k-means in SPSS will immediately give you the ANOVA comparisons between the clusters. Just like in the case of Ward's method you can add the cluster memberships to the datafile, and use them in subsequent analysis. Once you move your analysis from people to an analysis of clusters, you have greatly summarized your data.

Final remarks

These lecture notes contain all the material you will need to know for this course. However, they are concise: to really become proficient at creating and evaluating questionnaires you need practice. During the tutorials and the practical you will practice yourself with all the techniques that are discussed in these notes. Make sure you understand the methods, and know how to apply them.

A lot of the statistical methods we discussed rely heavily on the analysis of correlations (or covariances). Make sure that you thoroughly understand these topics.

Good luck on the final exam!

Maurits Kaptein
m.c.kaptein@uvt.nl

Before emailing: please check blackboard first to see whether your question is answered there. Also, for questions about the materials of this course during the course please consult me at the break or directly after a lecture.

Formula sheet:

1) Gemiddelde mean(x) $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

2) Variantie var(x) $\text{var}(x) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$

3) Standaard deviatie s_x $s_x = \sqrt{\text{var}(x)}$

4) Covariantie cov(x,y)

$$\text{cov}(x,y) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

5) Correlatie r_{xy} $r_{xy} = \frac{\text{cov}(x,y)}{s_x s_y}$

6) Lineaire combinatie $V = a_1X_1 + a_2X_2 + \ldots + a_JX_J + b$

Gemiddelde lineaire combinatie $\bar{v} = b + \sum_{j=1}^{J} a_j \bar{x}_j$

Variantie lineaire combinatie $\text{var}(V) = \sum_{j=1}^{J}\sum_{k=1}^{J} a_j a_k \text{cov}(x_j, x_k)$

7) Lineaire combinaties $V = a_1X_1 + a_2X_2 + \ldots + a_JX_J + b$ en $W = c_1Y_1 + c_2Y_2 + \ldots + c_KY_K$.

Covariantie lineaire combinaties $\text{cov}(V,W) = \sum_{j=1}^{J}\sum_{k=1}^{K} a_j c_k \text{cov}(x_j, y_k)$

8) Model klassieke testtheorie $X_i = T_i + E_i$

9) Opsplitsing variantie testscore $\text{var}(X) = \text{var}(T) + \text{var}(E)$

10) Betrouwbaarheid $r_{xx'} = \frac{\text{var}(T)}{\text{var}(X)}$

11) Spearman-Brown $r_{kk'} = \frac{k\, r_{xx'}}{1 + (k-1) r_{xx'}}$

12) Spearman-Brown, herschreven $k = \frac{r_{kk'}(1 - r_{xx'})}{r_{xx'}(1 - r_{kk'})}$

13) Variantie van een testscore $\text{var}(X) = \sum_{j=1}^{K}\sum_{k=1}^{K}\text{cov}(x_j, x_k)$

14) Cronbachs α $\alpha = \frac{K}{K-1}(1 - \frac{\sum_{k=1}^{K}\text{var}(X_k)}{\text{var}(X)})$

15) Standaardmeetfout $S_E = S_X\sqrt{1 - r_{xx'}}$

16) Correctie voor attenuatie $r_{X_TY_T} = \frac{r_{XY}}{\sqrt{r_{XX'}}\sqrt{r_{YY'}}}$

17) Model factoranalyse $X_i = \sum_{k=1}^{K} a_{ik}F_k + b_iU_i$

18) Communaliteit $h_i^2 = \sum_{k=1}^{K} a_{ik}^2$

19) Gereproduceerde correlatie r_{X1X2} als factoren orthogonaal zijn:

$$r_{X1X2} = \sum_{k=1}^{K} a_{1k}a_{2k}$$

20) Door factor k verklaarde variantie (eigenwaarde) $\lambda_k = \sum_{j=1}^{J} a_{jk}^2$

21) Gereproduceerde correlatie r_{X1X2} als factoren gecorreleerd zijn, in het geval van twee factoren F_1 en F_2: $r_{X1X2} = a_{1F1}a_{2F1} + a_{1F2}a_{2F2} + r_{F1F2}(a_{1F1}a_{2F2} + a_{1F2}a_{2F1})$

22) Correlatie tussen factor en item als factoren gecorreleerd zijn, in het geval van twee factoren F_1 en F_2: $r_{X1F1} = a_{1F1} + a_{1F2}r_{F1F2}$

23) Aantal elementen in covariantiematrix (COV): $J(J+1)/2$

24) Aantal parameters in confirmatieve factoranalyse: $2J \leq$ parameters $\leq 2J + K(K-1)/2$

25) Df = aantal elementen COV – aantal parameters

26) C.R. = effect/(standard error effect)

27) (geneste) Modeltest in confirmatieve FA van Model A tegen Model B

chi-kwadraat model B – chi-kwadraat model A
$df = df_B - df_A$

28) $(\text{afstand})^2$ van *i* in *k* op *j* t.o.v. gemiddelde op *j* in cluster *k*

$$(X_{ijk} - \bar{X}_{jk})^2$$

29) <u>som over alle individuen in *k*</u> van $(\text{afstand})^2$

$$\sum_{i=1}^{n_K} (X_{ijk} - \bar{X}_{jk})^2$$

t.o.v. gemiddelde op *j* in cluster *k* =
= bijdrage variabele *j* aan SS_W van cluster *k*
want $(\text{afstand})^2 = (\text{error})^2$

30) <u>som over alle variabelen</u> van bijdragen

$$\sum_{j=1}^{J} \sum_{i=1}^{n_K} (X_{ijk} - \bar{X}_{jk})^2$$

variabele aan SS_W van cluster *k*
$= SS_W$ van cluster *k*

31) <u>som over alle clusters</u> van SS_W

$$\sum_{k=1}^{K} \sum_{j=1}^{J} \sum_{i=1}^{n_K} (X_{ijk} - \bar{X}_{jk})^2$$