

## **Design of a Questionnaire Instrument**

***Willem-Paul Brinkman, Delft University of Technology, the Netherlands***

### ***Introduction***

Questionnaires have obtained a rather ambivalent reputation as a research instrument. Although they can be very useful, designing a good questionnaire takes considerable effort and thinking. Two things that are not always around, it seems. Not surprisingly, we all have come across some, to put it mildly, terrible questionnaires. Sitting down for five minutes to write down a number of questions is, of course, unlikely to result in a brilliant questionnaire. Still for the untrained eye, it looks like a questionnaire and again another proof that they should not be trusted. Among Human-Computer Interaction (HCI) professionals, this aversion is also noticeable. For example, when asked to compare usability methods, professionals rated questionnaires and surveys among the bottom of methods they use or have used (Gulliken *et al.*, 2004; Bark *et al.*, 2005; Mao *et al.* 2005). On the other hand, a large group is using them. For example, in a survey among 197 HCI practitioners (Bark *et al.* 2005) 53% of the responders had used a questionnaire to evaluate a product, whereas 97% had used a user test for this. This means that although HCI professionals might sometimes be sceptical towards questionnaires, they do use them and therefore there is a market for good questionnaires. In this chapter I will try to give you an insight in how to design, evaluate, and administer such a questionnaire. I will do this by providing examples from research I have done with my colleagues over the years in the area of mobile HCI. From reading this chapter you should be able to setup a questionnaire, conduct various analyses to check its reliability, its validity, and have an idea how to distribute and analyse the questionnaire.

### ***Constructs, questions and scales***

The text of a questionnaire is in many ways similar to the source code of a program. It is the result of a lengthy design process. It includes instructions, and should be well tested before it can be used. Before we look at the actual design of a questionnaire and its underlying concepts, I suggest we start with looking at the basic general setup of a questionnaire.

### **Questionnaire setup**

The setup follows a normal conversation. You first introduce yourself; explain your motivations etc, next, you have your conversation, which in this case is a question and answer session; at the end, if only out of respect, you ask if the person likes to add something, and you thank the person for his or her time. If you look at a questionnaire you might find this back in the following items:

- *General introduction*; Here you can welcome people and thank them for taking the time for filling out the questionnaire. You also need to explain what the aim is of the questionnaire, what will be done with the data, who is collecting this data, and, if people have questions, who they can contact. Also important is the way in which data will be reported. Will answers be reported anonymously? As people are always busy, it might also be good to give an indication how long it will take to complete the questionnaire.
- *Collecting biographical data*; Although you might be evaluating a mobile phone application, later on in your reporting you might like to mention something about the background of the people that provided the data, for example: age, gender, educational background, mobile phone experience, occupation, country of origin etc. With this data you will be able to give an indication how generalisable your findings are. For example, if I ask a group of retired academic professors to evaluate a mobile phone application, how applicable would these findings be to my 14-year old niece Monica and her teen-age friends?
- *Introduction topic*; If the questionnaire covers multiple topics, it would be good each time to spent a few words to introduce the topic before you present the questions. If the topic is abstract or could be interpreted in different ways, you might include a definition in this introduction.
- *Questions*; Yes, your questionnaire should have questions. Number the questions, and make sure that everything is well laid out. People are less willing to fill out a questionnaire that seems to have been put together in a rush without much effort and consideration.
- *Rounding up*; At the end of the questionnaire, you should consider to include an open-question to give people the opportunity to comment on the questionnaire it self. This information can be very useful. Someone might write down for example that when he changed his mind he put a circle around the right answer and that you should ignore the answer he originally had ticked off. It would not be the first time that I had to throw out data because the answers were ambiguous. Of course, I had only myself to blame, as you should give clear instructions on what to do if people want to change their answer. You might also ask the people to check that they have answered all questions and did not accidentally skip one. Do not forget to thank people again for their time and effort, and finally make sure that people know how they can return the completed questionnaire to you.

## **Formulating questions**

Having an idea about the general setup of a questionnaire, it is now time to pay some attention to formulating a question. A general rule is that a question should be clear and answerable. Here are some other points to consider when formulating your questions.

*Pose one question at a time.* Take the following question: How would you rate the usability and the usefulness of this application? Now imagine that the usability is dreadful but the application is extremely useful. How would you answer this question? It is therefore better to split this question into two separate questions. In this way it is easier to interpret the answer.

*Questions should be unambiguous.* Vague and woolly questions result in answers that are also vague and often do not discriminate. Take, for example, the question: Has mobile technology changed society? This question reads like your average horoscope. You can read into it whatever you like. Does society mean my country, the whole modern civilisation or are we talking about the people in my street? Similarly in what way did society change? Has it become more polite, flexible, technology dependent, etc. With this amount of ambiguity, it is easy for people to agree on an answer, although they might mean completely different things.

*The language and expected knowledge should fit the target group.* I am not just saying here that I might run into some problems with my Dutch questionnaire in the middle of Peking. Instead, the right language also refers to the level of technical terms that I can or should use. Terms like user interface, flow, and cognitive overload might not be suitable for a questionnaire targeting my mother's old theatre group, but be completely appropriate for a questionnaire targeting HCI experts.

*Formulate questions neutrally, avoid leading questions.* With a question like: Do you also hate this ugly mobile design? people might find it difficult to say No. A leading question gives people an indication of what answer is expected; something that should be avoided if you want a truthful answer.

*Avoid unnecessary sensitive questions.* It often is difficult to find people to fill out a questionnaire. Certainly I do not want to scare them away because of unnecessary sensitive questions. Therefore, you have to consider whether you really need an answer to a sensitive question, and whether you can reformulate it into less a sensitive question. Take for example the questions to collect some biographical data. You could ask people to give their company's name. However, people might be uncomfortable with this as their answers might reflect badly on their company. Instead, you might avoid this problem altogether if you ask the type of company they work for.

*Avoid negative, or even worse, double negative questions.* If your questions are confusing people, their answer will not be very useful. For example, which question is easier to understand: (1) Are all the menu options easy to navigate to? or (2) Are none of the menu options not easy to navigate to? The last question simply hurts my brain.

Questions can be split into open and closed-questions. With open questions, you leave the format in which people answer a question up to them. For example, 'How would you describe the usability of this application?' is an open-question, allowing people to write down whatever they like to say on the matter. With a closed-question, or fixed-choice question, you provide people with an answer format. For example, people have to select their answer from a list of answers. Closed-questions have the advantage that they are easier for people to respond to and they are easier to analyse. Open-questions however give you much richer information. Still, people need more time to answer them, and processing the data also takes longer. Often questionnaires consist of a mix of open and closed-questions. Still as the question type

versus sample size matrix (Figure 1) shows the ratio might vary. For example, when I have only access to a small number of participants, but if they are willing to put aside an extensive amount of time, I will go for a questionnaire with mainly open questions. This will give me the best option for understanding the phenomenon that I am trying to study. Instead of a very well-structured questionnaire, I might even go for a semi-structured or even unstructured interview. This will result in a large amount of data from each person. Still, because I am dealing with a small number of participants, I would be well able to code, to analyse and to report this data. Using a large number of closed-questions in this situation is less useful. The numbers do not give in-depth information. Next, with a small sample size it is difficult to make with confidence any generalisations towards the entire population.

In a situation where I have access to a large number of people, I can not expect that they can all set aside an extensive amount of time. I therefore have to streamline the process by using closed-questions that people can answer easily. When answers are presented in a structured format, it is also much easier to process them later on; something that is important if I have to deal with a large sample size, say 100 people. Using a large number of open-questions will simply overwhelm me. Instead of aiming for in-depth understanding, with closed-questions the focus is on systematically summarising the data and if possible trying to generalise it to the population at large.

0% closed 100% open	Emphasis is on in-depth understanding More time per participants	Results in a large amount of data, which is difficult to code, to analyse and to report
Ratio open/closed questions		
100% closed 0% open	Data is missing depth and width and therefore less useful	Emphasis is on generalisation using statistical methods
	Few	Many

Number of participants

**Figure 1: Question type versus sample size matrix.**

Closed-questions have answer options and come with answer instructions. When considering answer options you already have to consider the type of statistical analysis you want to apply. In other words you have to decide on the level of measurement, which can be nominal, ordinal, interval, and ratio (Table 1).

**Table 1: Levels of measurement and their properties.**

Level	Example	Properties	Analysis type
Nominal	Nationality	Distinction	Mode, $\chi^2$
Ordinal	Education (BSc, MSc, PhD)	Ordering	Median, Range, Mann-Whitney U-test, Spearman Correlation
Interval	Intelligence	Ordering with equal interval	Mean, Standard deviation, <i>t</i> -test, ANOVA, Pearson Correlation
Ratio	Weight	Ordering with equal interval and meaningful zero point	multiplication and division (see also interval)

Nominal level is the simplest level. A person is to select one of the options, whereby there is no fixed order between the options, for examples nationality, gender, or make. A question, including answer instructions, and answer options might look like this:

- 1) What is your gender?  
(tick one of two options)
- male
  - female

Answers of an ordinal level can be placed in an order, for example: grade, academic qualification, or generation of the application. Together with the nominal level, ordinal level data is sometimes referred to as categorical data or qualitative data. The latter is somewhat confusing as some people use the term qualitative data to refer to only unstructured data obtained for example with open-questions. An example of a question with answer options on an ordinal level would be:

- 2) How would you rate the artistic quality of the mobile phone design?  
(tick one of the options)
- very low
  - low
  - average
  - high
  - very high

Data on interval or ratio level is often referred to as quantitative data. Answers on interval level can also be placed in an order, but in addition, they have equal intervals between them, for example: intelligence, the date, or temperature in Celsius or Fahrenheit. Data on a ratio level is almost similar to interval level. However, it also has a meaningful zero, for example, age, temperature in Kelvin, or weight. Where on interval level you can calculate the mean, on a ratio level you can go one step further and make multiplication, for example, that people used an application twice as long. However, often researchers tend to take ratio and interval level together. An example question with an answer on an interval level would be:

3) How would you rate the artistic quality of the mobile phone design?  
(put a cross on the line)



Measuring the distance between the starting bar and the cross with a ruler would give you a value representing the rating of the artistic quality.

### Attitude scales

One type of answer scales that you often will use in questionnaires are attitude scales. Attitude is defined as ‘a psychological tendency that is expressed by evaluating a particular entity with some degree of favour or disfavor’ (Eagly and Chaiken, 1993, p.1). An example of an attitude question would be someone’s attitude towards using a special character entry method. Although attitude might be considered as a multi dimension concept, a scale attempts to measure it unidimensional. Popular attitude scales are Thurstone scale, Likert scale, and Semantic differential scale. Let’s start with Louis Thurstone’s method of Equal-Appearing Intervals.

### Thurstone’s method of equal-appearing intervals

The idea of this attitude scale is that people receive a number of statements and for each they have to indicate whether they agree or disagree with it. Although not presented in the questionnaire, each item has a scale value. A person’s attitude score is determined by taking the median or the mean of these scale values for all the items a person has agreed with. Take for example Table 2, which shows a scale to measure people’s attitude towards mobile phones. If a person only agreed with the two statements: “Mobile phones are a necessary evil” and “Mobile phones are a mixed blessing”, the score would be  $(28+65)/2 = 46.5$ .

**Table 2: Items and scale value from the attitude towards mobile phones scale.**

Item	Mean	SD	Scale Value
Mobile phones are convenient	8.4	1.5	90
Mobile phones are a necessary evil	3.8	2.1	28
Mobile phones are the best invention of the 20th century	9.2	2.0	100
A mobile phone is a piece of technology like any other piece of technology	5.6	1.9	53
Mobile phones should be banned	1.7	1.6	1
Mobile phones should be used with caution	4.9	2.1	43
Mobile phones are a mixed blessing	6.6	2.0	65
Mobile phones are wonderful	9.0	1.7	97
Mobile phones ruin the art of conversation	3.3	1.9	22
Mobile phones have their usefulness	8.2	1.7	86
Mobile phones make me stressed	2.6	0.9	13
Mobile phones are dangerous	2.4	1.9	10
Mobile phones give me value for money	7.4	1.8	77

Note: Scale values were obtained after a linear transformation to a 100-point scale. 25 judges (male=21, female=4) with a mean age of 30.5 year (SD = 8.2) rated 61 items on an 11-point scale from very unfavourable to very favourable towards mobile phones.

I created Table 2 in a number of steps. First I determined the attitude I wanted to measure, which was a person's attitude towards mobile phones. The next step was to create a large set of candidate statements that describe specific attitudes people might have towards mobile phones. You could, for instance, ask a group of people to do this. In my case however, I did a little brain storming session with my wife while driving to the beach. When I got back, I also searched on the Internet with queries like "mobile phones are", "mobile phones give" etc. This resulted in a list of 61 statements that were relatively short, containing only one single thought, were not factual, and did not include words such as "never", "always", and "only". The next step was to give these statements to a panel of judges to rate them. I asked 25 students and colleagues at my department to rate the statements on a scale from 1 = 'very unfavourable' to 11 = 'very favourable' towards mobile phones. I made sure to instruct the judges not to rate their agreement with the statement. Instead I instructed them to judge the unfavourableness or favourableness expressed in the statement. The following step was to calculate the mean score and the Standard Deviation (SD) of each statement. The mean scores I used as the raw scale values. For the final scale I selected statements with relative small SD, indicating a high level of agreement among the judges. I also tried to select statements that together represented in more or less equal intervals the range of the possible scale values. This resulted in a set of 9 statements, which I extended by selecting two additional statements at the low side of the scale and two at the high side of the scale. I did not only look at the mean and the SD. I also made sure that the statements that I selected made sense in the context of the other statements. To make the scale values more interpretable I transformed the raw mean score to a score on a 100-point scale. However, remember a Thurstone scale is presented in a questionnaire without these numbers. People are only confronted with the statements in a random order. Besides the Thurstone's method of Equal-Appearing Interval, there is also the Method of Successive Intervals, and the Method of Paired Comparisons. The last one is interesting if you only have a small candidate pool. Instead of rating statement on one scale, judges are asked to compare pairs of statements and to indicate which of the two is more favourable towards the attitude object. You can imagine that with a set of 61 statements, I would have to be very nice to each judge, as he or she would have to do 1830 comparisons for me.

### **Likert scale**

You might not have encountered a Thurstone scale yet. However I am sure that you have come across a Likert scale. For example this is a question with a Likert scale from Fred David's (1989) questionnaire that I used to evaluate the usability of a mobile phone's keyboard (Brinkman *et al*, in press):

Please indicate the likelihood of the following statement: 'I would find the keyboard of the mobile phone easy to use'  
(circle one of the numbers)

Unlikely	1	2	3	4	5	6	7	Likely
	extremely	quite	slight	neither	slightly	quite	extremely	

Likert scales are also often used to ask people to state their agreement with a statement. For example, the following five-point scale allows people to rate their agreement from 'strongly disagree' to 'strongly agree'.

1	2	3	4	5
Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree

Some people also apply a nine-point Likert scale, or use an even-point scale without including a neutral option. A Likert attitude scale exists out of a number of these statements, which like Thurstone scale represent either favourable or unfavourable attitude towards the attitude object. However, unlike Thurstone scale, which is considered interval-level, there is disagreement whether a Likert item should be considered interval or ordinal, the latter especially when the numbers are left out or the options are not visually presented with equal distances. As I mentioned before, this level of measurement determines what type of statistical methods can be applied. To calculate the attitude score the response on individual Likert items are often added up (or averaged). In that case, items should use the same Likert scale and it must seem reasonable that they can be considered to be of an interval-level. The attitude score can again be considered to be of an interval-level. To find a set of Likert items, you first need to create a pool of candidate items. These items must be pilot tested with a relative small group of people. Based on their response you can remove items that do not discriminate, i.e. have small standard deviation, or have low reliability, which I will explain in more detail later on.

A problem with Likert scales is the interpretation of the neutral option as some people might select this if they have no opinion or can not give an answer. For example, a statement might refer to a specific function of the mobile phone application a person has never seen. This of course would distort your results, as it seems to suggest a neutral attitude whereas a person simply had not formed an attitude. If this is likely to happen you might want to offer people a separate option with 'no opinion' or 'not applicable', and ignore their data when analysing people's attitude.

### Semantic differential scale

The last attitude scale I like to discuss is the semantic differential scale. Together with Steve Love (Brinkman and Love, 2006) we developed an instrument to evaluate people's attitude towards the design of a mobile phone. Below are three semantic differential scales used in our instrument to measure, what we have labelled, the sophistication of the design.

	Sophistication							
Childish	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Sophisticated
Silly	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Classy
Novelty	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Business

People are asked to mark a scale between bipolar words (adjectives preferably). With regard to the level of measurement, it is a similar story as before with the Likert scale. Also semantic differential scales have the



problem of people sometimes using the middle options as 'no opinion'. Like the other scales, a fundamental first step in the creation of the scale is to establish a candidate pool of items. One way of doing this is using the Repertory Grid technique (Fransella and Bell, 2004). This interview technique, which complements the Personal Construct Theory (Kelly, 1963), helps individuals to talk about the constructs they use to interpret the world, for example the design of mobile phones. In our study we gave 20 people 3 photos, each of a different mobile phone, and we asked them to say which of the phones were similar and which the odd one out. Next, we asked them to describe and put a label on the two groups of the triad, which resulted in bipolar scales, such as immature – mature, complex – userfriendly, or ugly – appealing. Our participants did this for 10 sets of 3 phone designs. Together these 10 bipolar scales formed a grid. After collecting the grids, we asked people to rate all the designs on the 10 scales they had created.

The following step was to see if the grids developed by each individual had some constructs in common with other participants. One possible way would be to study the semantics of the labels in the 20 grids. However, this approach is limited because individuals might have verbalised the labels of the construct differently, while referring to a similar underlying construct. We therefore applied a statistical approach. The assumption of this approach was that although participants might verbalise construct differently, if they refer to the same underlying construct, they would rate the designs in a similar way. For example, one participant mentioned the construct Femininity - Masculine and rated the phones on this scale as 6, 4, 2, 5, etc, while another participant mentioned the construct Female – Male and rated the phones on this scale as 7, 5, 2, 6, etc. Although the ratings are not exactly the same, a clear correlation in the rating patterns is visible. Factor analysis is a systematic statistical method by which the correlations between all these ratings can be studied. It can establish a reduced number of components that accounts for the variance in the rating of the designs. If individual bipolar scales represent the same underlying fundamental construct, they should correlate highly with the same component of the factor analysis. Table 3 shows part of the results when we entered the 200 bipolar scales in a factor analysis. In selecting the components and filtering out bipolar scales we used the following rules:

- Factor loadings (correlations) below 0.69 were ignored.
- Components should have constructs loading from at least five different participants.
- A clear semantic relationship between the labels of the construct should exist.

**Table 3: Part of the result of a factor analysis on bipolar scales obtained in Repertory Grid analysis.**

Participants	Label	Load
1	Confusing - Appealing	0.77
	Complicated - Simplistic stylish	0.91
	Chunky - Simplistic	0.94
	Large - Easy to use	0.92
2	Uncomfortable to use - Appealing	0.94
	Complex - Attractive	0.92
	Not clear - Clear	0.94
	Gender biased - Gender neutral	0.70
3	Sophisticated - Plain	0.76
4	Complex - Simple	0.70
	Not understandable - Understandable	0.76
8	Uncomfortable in pocket - Comfortable in pocket	0.77
9	Unconventional - Conventional	0.81
15	Modern - Traditional	0.75
	High interactivity - Low interactivity	0.74
17	Modern - Traditional	0.81
	Unique - Simple	0.77
	Futuristic - Current	0.87
18	Fully functional - Minimalistic	0.79
	Futuristic - Modern	0.78
20	Complicated looking - Simple to use	0.86
	Hidden features - Fast access to options	0.90

Source: This table is taken from Brinkman and Love's (2006) study in developing a measure instrument for individual's attitude towards the design of mobile phone.

Looking at Table 3, we decided that the constructs that loaded on this component related to the mobile phone's ease of use. Scales that seemed to suggest this were: confusing – appealing, complicated – simplistic stylish, chunky – simplistic, large – easy to use, comfortable to use – appealing, complex – attractive, not clear – clear, complex – simple, not understandable – understandable, uncomfortable in pocket – comfortable in pocket, unique – simple, and complicated looking – simple to use.

The repertory grid study resulted into 5 groups of bipolar scales. We used these scales to evaluate the design and people's preference. In a second study we examined the reliability and validity of these scales from our candidate pool, which resulted into an instrument with only 3 groups. However, before going into this in more detail, I first like to look at the relation between constructs and individual questions.

### **Constructs, dimensions and items**

I mentioned in the beginning that developing a questionnaire has some parallels with programming. When you develop a program, you first make a design of the program. You start with the main functions or main objects, and gradually you fill in the details. For the design of a questionnaire you do something similar. Until now we have mainly talked about the lowest

operationalisation level of the design. However, if your questionnaire deals with more complex and abstract concepts it is important that you first define these high level constructs and operationalise them into measurable items. Table 4 shows part of the operationalisation of the attitude toward mobile phone design questionnaire. It shows that on the highest level we are interested in people's attitude. As we found out in our study this construct is multidimensional, including dimensions such as: gender, sophistication, and ease of use. For the ease of use, dimension we assumed that people would want an easy-to-use design. For the other dimension, this is an individual choice. Some people might prefer a feminine design, but evaluate the design of a specific mobile phone as masculine. We therefore include the sub-dimension *evaluation* and *preference*. To measure these (sub) dimensions, we developed indicators that we translated each into a single question. In this case we used questions with semantic differential answer scales. However, other types of questions are possible. Note that for more abstract constructs you might even have sub-sub-dimensions.

**Table 4: Example of operationalisation of the construct 'attitude toward mobile phone design'.**

<b>Construct</b>	<b>Dimension</b>	<b>Sub-dimension</b>	<b>Indicator (item)</b>
Attitude towards mobile phone design	Gender	Evaluation (The phone is)	Feminine – Masculine
			Female – Male
		Preference (I like a phone which is)	Feminine – Masculine
			Female – Male
	Sophistication	Evaluation (The phone is)	Childish – Sophisticated
			Silly – Classy
		Preference (I like a phone which is)	Novelty – Business
			Childish – Sophisticated
Ease of Use	(The phone is)	Silly – Classy	
		Novelty – Business	
		Hard to use – Easy to use	
			Complex – Simple
			Complicated – Simple to use

Now you might wonder why we use 15 questions just to measure one thing. Would it not be much easier just to use one simple question something like "What is your attitude towards the design of the mobile phone?" Indeed, it would save everyone a lot of time. However, it has as a number of severe drawbacks, which have to do with the complexity and abstractness of the construct. First of all, would a person understand what we mean with this abstract construct? Even if this would be the case, would the person take everything in consideration when answering this one single question? For example, when answering this question, a person might only think about the ease of use, and might forget that he did not like the design because it looks too feminine. The real fundamental issue is that we often have to deal with abstract mental constructs, something that does not have an actual representation in the physical world. By using multiple questions per

dimension you reduce the chance that a person gives you a wrong or partial answer just because he or she did not understand that specific question or forgot to look at all issues involved. By using multiple questions, you surround your target and attack it from all possible angles. In doing so, you increase your chance of catching it. Or as you like, kill it.

The number of questions that you should use to measure a dimension of course varies with the complexity of a dimension. For example, if you want to know someone's age, a single question "What is your age?" will probably do it. More abstract construct requires more indicators. Some books even suggest as guideline 10 questions per dimension. Still it all depends on the construct.

### ***Reliability and validity***

A central issue for scientific research is reliability and validity. Especially when dealing with vague intangible mental constructs such as happiness, usability, or flow, people might wonder how well a questionnaire captures them. Reliability refers to the consistency in the results of the measurement, while validity concerns whether the questionnaire is measuring what it claims to be measuring. Imagine I come up with a new perfume for attracting only beautiful girls. However, each time I use it, I only attract the attention of the dogs in my neighbourhood (no pun intended!). The validity of perfume's claim would be low or not existing. Still the result of my perfume would be very reliable, as it consistently gives the same result. Notice that the other way around is not possible. A measure can not be valid but unreliable at the same time. If each time I wear my perfume and I attract the attention of different creatures (e.g. dogs, cats, or even sometimes beautiful girls) the perfume is very unreliable, and I certainly can not claim my perfume is only attracting beautiful girls. The same can be said for questionnaires. If it intends to measure the perceived artistic beauty of a mobile phone design, this measure is reliable if it gives me the same results as long a person does not change his or her opinion about the beauty of the design. The measure is valid if it measures what it supposes to measure, for instance the perceived beauty and not the usability or usefulness of a mobile phone.

Researchers talk about a large number of different types of validity. I will focus only on those directly relevant for questionnaire measures, which are: face validity, whether the measure looks valid; content validity, whether the measure captures the full content of the construct; criterion validity, whether the result of the measure agrees with another valid sources; construct validity; whether the measure measures the unobservable, theoretical construct. But before I discuss all these types of validity and how to assess them, I first like to talk about reliability.

### **Reliability**

As I mentioned before, reliability is about consistency in the results of the measurement. One simple way of testing this would be to apply the questionnaire twice under exactly the same conditions and see if you obtain

similar results, by for example calculating the correlation coefficient. This is called *test-retest reliability*. Although this sounds straightforward, it can be tricky as the same conditions, also mean the same person, under the same condition. For example, measuring someone's mood with an interval of a couple of hours might give completely different results. Although in a lesser extent, measuring attitudes, beliefs and intention might also be affected as people can change their mind over time. Still, when the interval is not too far apart you would still expect some degree of correlation. The interval however should not be too short otherwise people might simply try to remember their previous answers.

Another approach to examine the reliability is to look for consistency within a single measurement. Two methods often mentioned in this context, are the *split-half* method and *Cronbach's alpha*. The split-half method splits the questions of a dimension in two, for example odd-numbered questions versus even-numbered questions, or just randomly split. Next, it correlates the scores across the two groups. Table 5 shows the results of a split-half analysis that I performed in SPSS on the original five ease-of-use dimension items. The correlation between the two groups (forms) is 0.441, which is not very high. However, splitting a group into two also reduces the reliability, as reliability is increased with the number of questions. In other words, reliability of two groups is lower than five items together. The Spearman-Brown prophecy coefficient compensated for this. Table 5 shows that the coefficient in our example is 0.612 or 0.619 depending whether the calculation assumes that a group has equal or unequal number of items.

**Table 5: Result of the split-half method on the ease-of-use dimension items.**

Measures		Values	
Cronbach's Alpha	Part 1	Value	.636
		N of Items	3 <sup>a</sup>
	Part 2	Value	.222
		N of Items	2 <sup>b</sup>
Total N of Items			5
Correlation Between Forms			.441
Spearman-Brown Coefficient	Equal Length		.612
	Unequal Length		.619

<sup>a</sup> The items are: Complicated - Simple to use, Difficult - Plain, Hard to use - Easy to use.

<sup>b</sup> The items are: Simple - Complex, Difficult to carry - Easy to carry.

The result of the split-half method changes when you split up the items differently. Why not try all possible combinations and come up with some average? This is what Lee Cronbach might have thought when he was developing his measure. When items have equal variance (e.g. after a z-transformation) the Cronbach's alpha measure becomes:

$$\alpha = \frac{N \times \bar{r}}{(1 + (N - 1) \times \bar{r})}$$

Whereby  $N$  is the number of items and  $\bar{r}$  the average correlation between the items. The Cronbach's alpha for the five ease-of-use dimension items is 0.644. This is relatively low. Criteria of acceptable levels of alpha qualify a value of 0.8 and higher as good, and above 0.7 as satisfactory (Loewenthal, 2001). Our alpha value is therefore not very satisfactory. There are several ways to deal with this. Kate Loewenthal mentions that it is unlikely with a small number of items to get a high alpha level, and you may consider accepting an alpha level of 0.6 if (and only if, she stressed): you have good evidence for the validity, you have theoretical and/or practical reasons for all items in the dimension, and the number of items in the dimension is small (less than about 10 items). Often a much easier way is to look again at the alpha level after an item is removed. Table 6 shows part of the SPSS results of a Cronbach's alpha analysis on the items. As you can see, removing the 'Difficult to carry – Easy to carry' item will result in a Cronbach's alpha of 0.723. This item does not seem to correlate well with the other items, as it only has 0.144 item-total correlation.

**Table 6: Part of the results of a Cronbach's alpha analysis on ease-of-use dimension items.**

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
Complicated - Simple to use	3.18	21.399	.494	.549
Difficult - Plain	3.27	22.073	.407	.587
Hard to use - Easy to use	3.44	20.617	.492	.545
Complex - Simple	3.85	19.531	.520	.527
Difficult to carry - Easy to carry	4.14	24.179	.144	.723

Important when you calculate Cronbach's alpha is that all items measure the construct in the same direction, for example, from low to high ease-of-use. In the original questionnaire people rated the mobile phone on a scale from 'simple to use – complicated'. In the analysis, I therefore reversed the scale and the data. If you forget to do this, the alpha value will decrease and could even become negative. Why? you might ask; because it is based on the average correlation. Before calculating Cronbach's alpha it is therefore always good to examine item-total correlations, as they should be positive.

With reliability, we only have an indication whether the questionnaire provides consistent results of something. The issue of validity addresses this 'something'.

### Face Validity

Face validity refers to whether the measure looks valid. If people look at the mobile phone design questionnaire, would they say: yes, this would measure a person's attitude towards the mobile phone design? There seems however some difference of opinion who you should ask to make this judgement. For example, Neuman (1997) refers to the scientific community, i.e. the experts; while others (Loewenthal, 2001; Rust and Golombok, 1999) refer to people that would potentially fill out the questionnaire. Although both cases concern

acceptance, the reason why it is important is different. The first one refers the consumers of the questionnaire and its data. If they do not accept it, they will not use the questionnaire and will not trust its data. The second one refers to the responders. If they do not accept the questionnaire, they might not fill it out seriously or refuse to fill it out all together. Responders might feel that questions are offensive, inappropriate to ask, or irrelevant. For that reason Rust and Golombok (1999) suggest to evaluate the face validity by asking potential responders whether the questionnaire name and that of the items (scales) are acceptable.

## Content validity

Content validity addresses the question whether the full content of a construct is represented in the measure or are some dimensions left out. Like face validity, content validity is a consensus issue. For content validity, experts have to agree that the construct has been operationalised capturing all facets of the construct. Take for example ISO (1998) standard 9241 – 11 which defines usability as ‘the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use’. If you follow this definition, the questionnaire should include the dimensions effectiveness, efficiency and satisfaction. These dimensions on their term should have indicators that cover the entire range of the dimension. One way of assessing the content validity comes from the area of Personnel Psychology. Charles Lawshe (1975) suggested a quantitative method to select job criteria by involving a panel of judges. Since its introduction researchers have used this method to assess the content validity of questionnaire items. Take for example the work of my PhD student, Marije Kanis. She is studying the effect of her PosiPost mobile phone application on people’s social wellbeing. To do this, she needed a questionnaire that would measure this construct. After going through the literature she established a list of questions. To examine the content validity, she sent the questionnaire to a panel of experts, and asked them to indicate for each item whether it was ‘Essential’, ‘Useful, but not essential’, or ‘Not necessary’ to measure the underlying construct. Table 7 presents the feedback of 12 experts. For example, for the Happy item, all experts rated it as essential for measuring the construct wellbeing.

**Table 7: Frequency of panel members that gave an item a specific rating and the associated CVR value.**

Item	Essential	Useful, but not essential	Not necessary	CVR <sub>strict</sub>	CVR <sub>relaxed</sub>
Happy	12	0	0	1.00	1.00
Good natured	3	2	5	-.17	.17
Tense	9	3	0	.50	1.00
Helpless	7	4	1	.17	.83
Depressed	11	1	0	.83	1.00
CVI				.47	.80

To decide whether there is sufficient support, Charles Lawshe argued that at least half of the experts should rate an item as 'Essential', and to express this he devised the following Content Validity Ratio (CVR):

$$CVR = \frac{n_e - N/2}{N/2}$$

in which  $n_e$  is the number of experts that rated the item as "Essential", and  $N$  is the panel size. CVR ranges from  $-1$  to  $1$ . A zero value means that half the panel rated it as essential and the other half did not. A value less than zero means fewer than half of the panel rated the items as essential, and a value of more than zero means more than half of the panel rated the items as essential. In Table 7 this is labelled this as  $CVR_{strict}$ . The Content Validity Index (CVI) is also presented in the table. This is the mean CVR value. The next step is to decide which items will be included and which will be removed. To do this, we need to realise that the panel members are just a random sample of the entire population of experts, and thus findings include an element of random error. To have an acceptable level of confidence in a majority Table 8 put forward a critical number of experts and the associated CVR value for each panel size. Only items equal or above the critical value should be included. For a panel size of 12, the critical number of experts is 10, which means a critical CVR value of 0.67. Only the 'happy' and the 'depressed' item should therefore be included, and all the other items should be removed. This strict strategy might work well when we have a very long list of items and we need to reduce it to only the essential items. However, when the original list is not long, this strategy might result in a very short list. In which case you might consider also including the 'useful, but not essential' ratings. Table 7 shows the  $CVR_{relaxed}$  value, which looks at the combined frequency rating of 'essential' and the 'useful, but not essential' categories. Comparing these values with a critical CVR value of 0.67 suggests that only the 'Good natured' item should be removed.

**Table 8: Panel size, critical number of experts, and critical CVR value for One-Tailed test,  $p < .05$ .**

Panel Size	Critical Size	Critical CVR
5	5	1.00
6	6	1.00
7	7	1.00
8	7	0.75
9	8	0.78
10	9	0.80
11	9	0.64
12	10	0.67
13	10	0.54
14	11	0.57
15	12	0.60
20	15	0.50
25	18	0.44
30	20	0.33
35	23	0.31
40	26	0.30
45	29	0.29



50	32	0.28
60	37	0.23
70	43	0.23
80	48	0.20
90	54	0.20
100	59	0.18

Note: Calculation of the critical number of experts is based on a Cumulative Binomial Distribution with  $1 - P(\text{Bin}(\text{Panel size}, 0.5) < \text{Critical number of experts}) < 0.05$

## Criterion validity

With criterion validity we compare the results of a questionnaire with other trusted sources or standards that are known to measure the construct. This could be another questionnaire (*concurrent validity*) or a prediction (*predictive validity*), for example, two different user groups or mobile phones designs that you know should give different questionnaire results. To examine the concurrent validity of your questionnaire you have to have another questionnaire that measures the same thing. However, if you find one, people would argue that your questionnaire is redundant. Catch-22, you can never win, you might think. The way out is that you have to justify why your new questionnaire is an improvement, for example easier or more cost-effective to apply, or to analyse. Take for example the component-based usability questionnaire (Brinkman *et al.*, in press) that we created to measure the usability of different interaction components of a device. As this questionnaire could measure the usability of any interaction component, we examined its concurrent validity by applying it on the usability of a mobile phone's menu and at the same time ask people to fill out a standard usability questionnaire for menus (Norman, 1991). Analysis of the results showed a significant correlation. Furthermore, we also found a significant negative correlation between our measure and the number of menu problems mentioned in the debriefing of the usability experiment. Both these finding allowed us to claim a degree of concurrent validity.

In the study of the component-based usability questionnaire we also examined the predicted validity by asking users to rate two versions of an interaction component. We designed one version to be less usable and another version to be highly usable. For example, for entering text, one version used the Modified-Model-Position method and another version used the Repeated-Key method. From the literature (Detweiler *et al.*, 1990), we knew that the first one is less usable than the second one. We did this for a series of components and demonstrated that the predicted usability differences were also observed in the data from our new questionnaire.

## Construct validity

Construct validity is an issue when your questionnaire includes multiple indicators to measure a dimension. You want to know whether the indicators produce results within the conceptual boundaries of the construct. This means that indicators of a dimension should produce data that correlate with each other (*convergent validity*) and should divert from data produce from indicators from other dimensions (*divergent validity*). One way to examine this

is to apply a factor analysis. As I mentioned before, a factor analysis is a statistical procedure whereby you try to reduce the data of multiple items (i.e. indicators) into a reduced set of factors, or sometimes called components. If things work well, all indicators of a dimension should be associated with, and only with, a single unique factor. In other words each dimension should have its own factor. Table 9 shows part of the results of a factor analysis that Steve Love and I (Brinkman and Love, 2006) did on the data of our mobile phone design questionnaire that 156 people completed. In the columns you can see how the items load (i.e. correlated) on each component. For example, the first 3 items load only well on the first component and not on the other 3 components. This suggests both convergent and divergent validity for the gender dimension. Its items connected closely with each other, and are not entangled with the other constructs. Examining the last component, we decided to drop the last two items from the instruments. We regarded two items as too few to measure a dimension.

**Table 9: Example of rotated component loading matrix of the initial mobile phone design questionnaire.**

Scales	Components			
	1	2	3	4
Female - Male	<b>0.89</b>	-0.01	0.07	0.01
Girl - Boy	<b>0.85</b>	0.05	0.21	0.17
Feminine - Masculine	<b>0.85</b>	0.00	-0.02	0.17
Hard to use - Easy to use	0.05	<b>0.80</b>	-0.03	-0.26
Complex - Simple	-0.02	<b>0.79</b>	-0.08	0.22
Complicated - Simple to use	0.01	<b>0.79</b>	0.23	0.07
Childish - Sophisticated	0.15	-0.12	<b>0.81</b>	0.03
Silly - Classy	-0.04	0.14	<b>0.75</b>	-0.10
Novelty - Business	0.14	0.09	<b>0.70</b>	0.30
Fun - Technical	0.09	-0.08	0.05	<b>0.85</b>
Playful - Serious	0.20	0.13	0.07	<b>0.83</b>

Source: Brinkman and Love (2006).

## ***Administering a Questionnaire***

Once you have the questions, it is now time to think about how to put them together in a single questionnaire and how to give it to people. You also have to decide if you want to include everyone in the target population or that you focus only on a sample, and if so, how large this sample should be. Likewise, you also have to decide how you will distribute the questionnaire, for example, by post, email, or face-to-face. You also might like to combine your questionnaire with, for example, an interview at the end of an experiment in which people used a mobile phone. In that case you have to decide which one you would do first, the questionnaire or the interview.

## **Order of the questions**

When people answer a question, there is a risk that they might have been influenced by the previous questions in the questionnaire. Imagine that a questionnaire has the following topics: beauty of the design, likelihood of obtaining the phone, and its usability. Because people were first asked to think about the beauty of the phone, they might be more primed to consider

this criterion in their attitude towards obtaining the phone rather than, for example, its usability. This is called an *order effect*, and I expect there is no way of stopping it. Just asking people politely to stop thinking about the previous questions will not help much I am afraid. Instead of attempting to stop it, it is more effective to control a potential systematic bias in your results. A simple way of doing this is by breaking up the topic or dimension grouping, and presenting the questions in a random order. On a question level people might still be influenced by the previous question. However at a topic level this effect is no longer systematic, as questions of a topic are unlikely to be preceded always by questions from the same topic. Instead of complete random order, you can also systematically control the order of the questions by maximizing the number of different topic transitions between questions. Another alternative is to use multiple versions of the questionnaire. For example, I have done this when I used a computer to administer a questionnaire. I programmed it in a way so that each person got another randomly ordered questionnaire. Of course randomising a questionnaire at a question level is only possible when questions are self-explaining. If they are not and you need to introduce the topic with some text, then consider making different versions of the questionnaire where the orders in which topics are presented vary. For example, in one version you presented the topics in an order A, B, C, D and in another version you present the topics in a reversed order like D, C, B, A.

### **Number of participants**

To analyse and draw conclusions from a questionnaire there is a fundamental difference between a census, whereby all people of the population fill out the questionnaire; and a sample, whereby only some members of the population fill out the questionnaire. With a census you can give exact numbers about the population. With a sample you only give approximations. Conducting a census can cost much time and effort. Take my university town Delft. To ask all people between 15 and 29 years old about their attitude toward the design of a specific mobile phone would require over 25 thousand people to fill out the questionnaire. An endeavour I think you should not take lightly. Clearly, more efficient would be to ask a represented sample of these citizens to fill out the questionnaire, and use this data to make, with some margin of error, generalisations about the attitude of the whole population. Now how large should this sample be? This depends on a number of issues such as the research question, the margin of error that would be acceptable, and the level of acceptable risk that the true margin of error exceeds the accepted margin of error. Take the Delft example, if the primary question of the questionnaire is to measure the attitude of this population on an interval-level Likert scale, then Table 10 shows that we need at least a sample size of 119 people. In this case we would have a 95% level of confidence that the approximation of the mean would have a 3% margin of error. In other words on a 7-points Likert scale we could say with a 95% certainty that the population mean would be within a range of 0.42 points ( $2(7 \times 3\%)$ ) on the Likert scale. If the central focus of the questionnaire would be a yes/no question whether a person would buy the phone, then the sample should at least include 385 people. It would result in a 95% confidence level that the estimated frequency of people that would answer yes is within  $\pm 5\%$  of the proportion calculated. The data in Table 10 is

based on William Cochran's formulas as explained by James Bartlett, Joe Kotrlik and Chadwick Higgins (2001). Please read their paper if you are interested in setups with other confidence levels or margin of errors.

**Table 10: Sample size for interval and nominal data.**

Population size	Sample size	
	Categorical*	Interval**
100	80	55
200	132	75
300	169	85
400	196	92
500	218	96
600	235	100
700	249	102
800	260	104
900	270	105
1,000	278	106
1,500	306	110
2,000	323	112
4,000	351	119
6,000	362	119
8,000	385	119
≥ 10,000	385	119

\*With a margin of error = 5%, alpha = 5% and estimate of variance = 0.25;\*\*With a margin of error = 3%, alpha = 5% and estimation SD in population = 1.167;

Source: Cochran's formulas as explained by Bartlett, Kotrlik and Higgins (2001).

Now imagine that the research question is not about estimating the mean or the frequency, but it is about finding a difference between groups or a relationship between constructs. The size of this difference or relationship (i.e. *effect size*) would then be the crucial issue. If you look for a very small effect size, you will need a larger sample size than when you look for a large effect size. So what effect size do you expect? Now I hear you think, if I knew that, I did not need the questionnaire in the first place. And of course you are right. The right question therefore should be: what effect size would still be relevant? For example, you might find the attitude towards two mobile phone designs to be significant different, but this difference might be so small that it has no practical relevance. Jacob Cohen (1992) picked up on this and has classified three levels of effect size: small, medium and large whereby a medium effect size represents an effect that approximates the average size of observed effects in various fields. Based on these effect sizes he also gave the required sample size to conduct specific statistical tests. Table 11 shows the sample size that I obtained from the statistical application G\*Power<sup>1</sup>. The table shows statistical methods that are often used. These numbers ensure that if there is a difference you will have an 80% chance of finding them. The level of accepted risk (alpha) is set for 5%. The table shows for example that we would need at least a total sample size of 128 people (62 in each group) if we want to find a difference that is classified as a medium size effect with a Two Independent Sample *t*-test.

<sup>1</sup> <http://www.psych.uni-duesseldorf.de/abteilungen/aap/gpower3/>

**Table 11: Statistical tests and their required total sample size for small, medium and large effect size at power =.80 and alpha = .05.**

Aim	Data level	Tests	Effect size		
			Small	Medium	Large
Find differences from a norm or between groups or between questions	Categorical	$\chi^2$ with 1 df	785	88	32
		$\chi^2$ with 2 df	964	108	39
		Sign Test (Binominal with p =.5)	786	90	30
	Interval and Ratio	One-Sample <i>t</i> -test	199	34	15
		Paired-Sample <i>t</i> -test	199	34	15
		Two independent sample <i>t</i> -test	788	128	52
		One-way ANOVA 3 levels factor	969	159	66
	One-way ANOVA 4 levels factor	1096	180	76	
Find relationships	Interval and Ratio	Production Moment Correlation	782	84	29
		Multiple Regression 2 predictors	485	68	31
		Multiple Regression 3 predictors	550	77	36

Source: G\*Power application

### Distribution of questionnaire

There are different approaches you can follow to distribute a questionnaire, each with their own specific advantages and disadvantages. Without being exhaustive, this is a list of some approaches with issues you might want to consider:

- *Direct approach*; Some one has to walk up to a person, or go door to door and ask people if they are willing to fill out the questionnaire. This is a very time-consuming approach. Still, it had the advantage of completeness and accuracy, as people can get help if they do not understand a question. This approach is very appropriate for pilot testing the questionnaire.
- *Telephone survey*; It is less expensive than the direct approach as you remove the need to travel, and sometimes it is possible to enter the response directly into a computer. If you consider this approach you might first send people a letter to explain the research in advance. At the start of the call you can refer to this letter, avoiding in some cases the need to explain you research in detail. A problem of a telephone survey is that people might confuse it with telemarketing, and therefore they might be less willing to participate.
- *Postal survey*; This is a less expensive approach, and avoids embarrassment of personal questions. The questionnaire should have very clear questions and instructions. Still, this approach tends to have a low response rate.
- *Email survey or Internet Survey*; This is an even less expensive approach than postal survey. However, it could result in a potential bias, as not everyone might have access to this medium.

A questionnaire could also be part of a series of data collection activities. For example, you might observe how people use a mobile phone application, and interview them afterwards. The interview process might however influence people's attitude or memories. You should therefore consider asking people to fill out the questionnaire before you start the interview. If you aim for keeping people's experience with the mobile phone fresh for the interview, consider taking out the factual questions from the questionnaire, such as their

biographical data, and ask people to complete those questions before the experiment or after the interview.

### ***Preparing the data for analysis***

Once you have the data, it is time to analyse it. Especially when I have a large data set I use a computer. Often questionnaire data can not directly be entered in statistical applications, such as SPSS or SAS. It first has to be transformed or coded, before statistical methods can be applied on it. It would take too far to discuss these statistical methods in this chapter, however, I will give you some suggestions on how to prepare your data, and give you directions on which statistical methods might be appropriate for your data.

### **Coding and reversing scores**

Part of your data might already be coded in a number, for example a person's age, salary, or the answers on a Likert scale from 1, 2, 3, to 7. However for categorical (ordinal and nominal) data, such as a yes/no answer, nationality and gender, this is not the case. In these cases a coding, e.g. a number, has to be assigned to each answer category. For example, yes = 0 and no = 1, or for a question about the type of Internet access, none = 0, dial up = 1, wifi = 2, satellite = 3, broadband = 4, mobile = 5, other = 6. Be aware of checklist questions where people can give multiple answers. For example, which mobile phone functionalities do you often use? The response should be treated as multiple yes/no answers. Record for each option whether a person checked it (checked = 1) or not (not checked = 0). For ordinal data the coding should also reflect the order in the data. For example, educational degree could be coded as BSc = 1, MSc = 2, and PhD = 3. You also have to record that data was missing. You could do this with a special code for example – 9999. However, when you calculate the mean, you do not want this value to be included. In SPSS for example, you can therefore specify the value of missing data, which it will ignore when calculating for example the mean.

Once you have coded the data, you need to reverse the score of any scale that has an opposite direction to that of the dimension. Take, for example, the design of a mobile phone questionnaire. Although in this chapter I presented all the semantic differentials in the same direction, e.g. from less to more easy to use, in the actual questionnaire almost half of the scale were presented in the opposite direction, for example Complex - Simple. With these reverse scales Steve Love and I tried to avoid that people can rush through the questionnaire without reading it or try to give socially desirable answers by always 'agreeing' with the right side of a scale. If you have a 5-point scale ranging from 1 to 5, you reverse it by subtracting the value from the range plus one. For example, a value 4 becomes  $5+1-4 = 2$ . If you have a balanced scale around zero with positive and negative number, such as -2, -1, 0, 1, 2 then you simply reverse the data by multiply it with -1.

If you have a dimension with multiple indicators that all use the same scale, then you can calculate how a person scored on that dimension. First check the Cronbach's alpha and see if there is sufficient consistency across the indicators. If this is the case, you can add up all the scores of the indicators and create a total score for that person. Instead of a summation, you can also

take the average. This has two advantages. First, the values of this overall score are easier to understand as they are expressed in the original scale. Second, it is easier to deal with missing data from some of the indicators by simply taking the average of the remaining indicators.

## Statistical tests

Once the data has been prepared, the next step is to actually analyse it. The statistical tests that can be applied on the data depend on a number of factors such as the research question, the relation between the samples, and the level of measurement. Table 12 gives an overview of some statistical tests that might be appropriate in some situations. Tests for interval and ratio data (*parametric tests*) often have additional assumptions about the distribution of the data, e.g. normal distribution. If these assumptions can not be met, it is always an option to look at their ordinal or even nominal counter part (*non-parametric tests*).

**Table 12: Statistical tests.**

Aim	Samples	Data Level	Tests
Find difference	Single sample	Nominal	Binominal test, $\chi^2$ goodness-of-fit
		Ordinal	Wilcoxon signed-ranks test
		Interval / Ratio	z-test, One-Sample <i>t</i> -test
	Independent	Nominal	Fisher-Exact test, $\chi^2$
		Ordinal	Mann-Whitney <i>U</i> test
		Interval / Ratio	z-test, two independent sample <i>t</i> -test, ANOVA, MANOVA
Dependent	Nominal	McNemar	
	Ordinal	Sign Test, Wilcoxon matched-pairs signed-rank test, Friedman Test	
	Interval / Ratio	Paired-Sample <i>t</i> -test, repeated measures ANOVA, MANOVA	
Find relation		Nominal	Cramér's V, phi
		Ordinal	Kendall, Spearman correlation
		Interval / Ratio	Pearson product moment correlation, regression analysis

When you have only a single sample it is not always straightforward how to interpret the data. Take, for example, a mean usability score of 4.7 for a mobile phone's menu on 7-points Likert scale ranging from 1 = 'very low' to 7 = 'very high'. What can you conclude? In the absence of any norm you might test whether this is significant above the middle value of the scale 4. If this is the case, you could argue that people were on average positive about the usability. Still, there are some limitations. For example, did responders also see 4 as the boundary between a positive and negative attitude towards the usability? Next, how does a score of 4.7 compare to other menus? To overcome these problems, some questionnaire designers (e.g. Kirakowski and Corbett, 1993) provide norm data for their questionnaire. They based their norms on data that they collected with their questionnaire. Other people can use the norm to compare it with their own data set. For example, we validated the component-based usability questionnaire (Brinkman *et al.*, in press) in several studies where people had to use easy or difficult to use components. This allowed us to put forward a total score of 5.29 as a breakeven point between difficult or easy to use components. When people

use the questionnaire and get a significant total score above 5.29 they can claim that the usability of their component is more comparable to the easy-to-use components than with the difficult-to-use components in the norm set. In short, if you develop a new questionnaire think about the possibilities of creating a norm that people can apply when they use your questionnaire in the future.

### **Final remarks**

After reading this chapter you might have been left with one of two different feelings. You might now have some confidence in creating your own questionnaire. As author this would of course please me. On the other hand, you might also have come to the realisation that you can not set aside the time or make the effort needed to create a good questionnaire. As author this would still please me, as you would see the value of using existing standard questionnaires. Throughout the chapter I have already shamelessly promoted questionnaires that I helped to develop. Of course, you also might want to look at other questionnaires. Gary Perlman has a website<sup>2</sup> that allows you to administer and collect online a number of standard HCI questionnaires.

### **Recommended reading list**

Coolican, H., (2004). *Research methods and statistics in psychology*. London, UK: Hugh Coolican.

Kurpius, S.E., and Stafford, M.E., (2006). *Testing and measurement: A user-friendly guide*. Thousand Oaks, CA: Sage.

Robson, C., (2002). *Real world research: A resource for social scientists and practitioner-researchers*. Oxford, UK: Blackwell.

### **Reference list**

Bark, I., Følstad, A, and Gulliksen, J., (2005). Use and usefulness of HCI methods: Results from an exploratory study among Nordic HCI practitioners. In *Proceedings of HCI*. London: Springer-Verlag, pp. 201-217.

Bartlett, J.E., Kotrlik, J.W., and Higgins, C.C., (2001). Organizational research: Determining appropriate sample size in survey research. *Information technology, learning and performance journal*, 19(1), 43-50.

Brinkman, W.-P. and Love, S. (2006). Developing an instrument to assess the impact of attitude and social norms on user selection of an interface design: A repertory grid approach. *Proceedings of ECCE-13*, pp. 129-136.

Brinkman, W.-P., Haakma, R., and Bouwhuis, D.G. (in press). The theoretical foundation and validity of a component-based usability questionnaire. *Behaviour and Information Technology*.

---

<sup>2</sup> <http://hcibib.org/perlman/question.html>



Cohen, J., (1992). A power primer. *Psychological bulletin*, 112(1), 155-159.

Davis, F.D., (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13, 319-340.

Detweiler, M.C., Schumacher, M.C. and Gattuso, N.L., (1990). Alphabetic Input on a Telephone Keypad. In *Proceedings of the Human Factors Society – 34th Annual Meeting*, vol. 1. Santa Monica, CA: Human Factors Society, pp. 212-216.

Eagly, A.H., and Chaiken, S., (1993). *The psychology of attitudes*. Belmont, CA: Wadsworth.

Fransella, F., and Bell, R., (2004). *A manual for repertory grid technique*. Chichester, UK: Wiley.

Gulliksen, J., Boivie, I., Persson, J., Hektor, A., and Herulf, L., (2004). Making a difference – A survey of usability profession in Sweden. In *Proceeding of NordiCHI*. New York, NY: ACM Press, pp. 207-215.

ISO, (1998), *Ergonomic requirements for office work with visual display terminals (VDTs) Part 11. Guidance on usability (ISO no 9241-11)*. Geneva: Switzerland, International Organization for Standardization.

Kelly, Y. (1963). *A theory of personality: The psychology of personal constructs*. New York, NY: W.W. Norton & Company.

Kirakowski, J., and Corbett M., (1993). SUMI: The software usability measurement inventory. *British journal of educational technology*, 24, 210-212.

Lawshe, C.H., (1975). A quantitative approach to content validity. *Personnel psychology*, 28, 563-575.

Loewenthal, K. M., (2001). *An introduction to psychological tests and scales*. New York, NY: Psychology Press.

Mao, J.-Y., Vredenburg, K., Smith, P.W., and Carey, T. (2005). The state of user-centered design practice. *Communications of the ACM*, 48(3), 105-109.

Neuman, W.L., (1997). *Social research methods: Qualitative and quantitative approaches*. Boston, MA: Allyn and Bacon.

Norman, K.L., (1991). *The psychology of menu selection: Designing cognitive control of the human/computer interface*. Norwood, NJ: Ablex Publishing corporation.

Rust, J., and Golombok, S., (1999). *Modern psychometrics: The science of psychological assessment*. London, UK: Routledge.

## **Self Test**

What do the following words/terms mean:

- Attitude
- Closed and open-question
- Coding
- Concurrent validity
- Construct
- Construct validity
- Content validity
- Content Validity Ratio (CVR)
- Convergent validity
- Criterion validity
- Cronbach's alpha
- Dimension
- Divergent validity
- Effect size
- Face validity
- Indicator
- Likert scale
- Nominal, ordinal, interval, and ratio data level
- Order effect
- Predictive validity
- Reliability
- Repertory Grid technique
- Reverse scale
- Sample size
- Semantic differential scale
- Split-half method
- Test-retest reliability
- Thurstone's method of equal-appearing intervals

## **Exercises**

### *Exercise 1*

Make an operationalisation of the construct: attitude towards text messaging. Include at least three dimensions each with their own indicators.

### *Exercise 2*

The attitude towards mobile phones scale presented in Table 2 has not been examined on its reliability or validity. Therefore, plan and conduct this examination. Consider using an alternative scale presented in Table 13. I created this scale from the remaining items in the candidate pool by considering this time the median score of the judges and the Inter Quartile Range (IOR) of the scores.

Table 13: Items and scale value from an alternative attitude towards mobile phones scale.

<b>Item</b>	<b>Median</b>	<b>IOR</b>	<b>Scale Value</b>
Mobile phones are a fact of life	6	2	47
Mobile phones are a novelty	6	2	47
Mobile phones create reckless drivers	2.5	2.3	1
Mobile phones make people more accessible	9	2	87
Mobile phones are for communication, nothing more, nothing less	5	2	34
Mobile phones invade my privacy	3	1	8
Mobile phones facilitate my friendships	9	1	87
For me mobile phones are lifesavers	10	2	100
Mobile phones make me superficial	4	2	21
Mobile phones are a useful part of life	8	1	74
Mobile phones are the source of much social discourtesy	3	2	8
Mobile phones are convenient for me	8	2	74

Note: Scale values were obtained after a linear transformed to a 100-point scale.