User Experience Questionnaire Handbook

All you need to know to apply the UEQ successfully in your projects

Author: Dr. Martin Schrepp

21.09.2015

Introduction

The knowledge required to apply the User Experience Questionnaire (UEQ) is currently split into several independent publications. The goal of this handbook is to bring all these pieces of knowledge together into one document. This will make it easier for practitioners to apply the UEQ in their evaluation projects.

We focus on the most important facts to keep the document short (since each additional page will reduce the number of people who read it significantly). We cite some publications for those who want to dig deeper into the subject. A much bigger list of publications concerning the construction and methodological development of the UEQ is available on the web site www.ueq-online.org. This site contains also a big list of papers that used the UEQ to evaluate products or to clarify scientific questions.

Construction of the UEQ

The original German version of the UEQ was created in 2005. A data analytical approach was used in order to ensure a practical relevance of the constructed scales, i.e. the scales were derived from data concerning a bigger pool of items. Each scale describes a distinct quality aspect of an interactive product.

In brainstorming sessions with usability experts, an initial item set of 229 potential items related to user experience was created. This item set was then reduced to an 80 items raw version of the questionnaire by an expert evaluation.

In several studies focusing on the quality of interactive products, including e.g. a statistics software package, cell phone address books, online-collaboration software or business software, data were collected with this 80 items raw version. In total 153 participants answered the 80 items of the raw version. Finally, the 6 UEQ scales and the items representing each scale were extracted from this data set by principal component analysis.

The items have the form of a semantic differential, i.e. each items is represented by two terms with opposite meanings. The order of the terms is randomized per item, i.e. half of the items of a scale start with the positive term and the other half of the items start with the negative term. We use a seven-stage scale to reduce the well-known central tendency bias for such types of items.

An example of an item is:

attractive o o o o o o o unattractive

The items are scaled from -3 to +3. Thus, -3 represents the most negative answer, 0 a neutral answer, and +3 the most positive answer.

The consistency of the UEQ scales and their validity (i.e. the scales really measure what they intend to measure) was investigated in 11 usability tests with a total number of 144 participants and in an online survey with 722 participants. The results of these studies showed a sufficiently high scale consistency (measured by Cronbach's Alpha). In addition, a number of studies showed a good construct validity of the scales.

If you want to know more details about construction and validation of the UEQ, see:

Laugwitz, B., Schrepp, M. & Held, T. (2008). Construction and evaluation of a user experience questionnaire. In: Holzinger, A. (Ed.): USAB 2008, LNCS 5298, 63-76.

Scale structure

The UEQ contains 6 scales with 26 items:

- *Attractiveness:* Overall impression of the product. Do users like or dislike the product?
- *Perspicuity:* Is it easy to get familiar with the product? Is it easy to learn how to use the product?
- *Efficiency:* Can users solve their tasks without unnecessary effort?
- *Dependability:* Does the user feel in control of the interaction?
- *Stimulation:* Is it exciting and motivating to use the product?
- *Novelty:* Is the product innovative and creative? Does the product catch the interest of users?

Attractiveness is a pure valence dimension. *Perspicuity*, *Efficiency* and *Dependability* are pragmatic quality aspects (goal-directed), while *Stimulation* and *Novelty* are hedonic quality aspects (not goal-directed).

The *Attractiveness* scale has 6 items, all other scales have 4 items. The Figure 1 shows the assumed scale structure of the UEQ and the English items per scale.

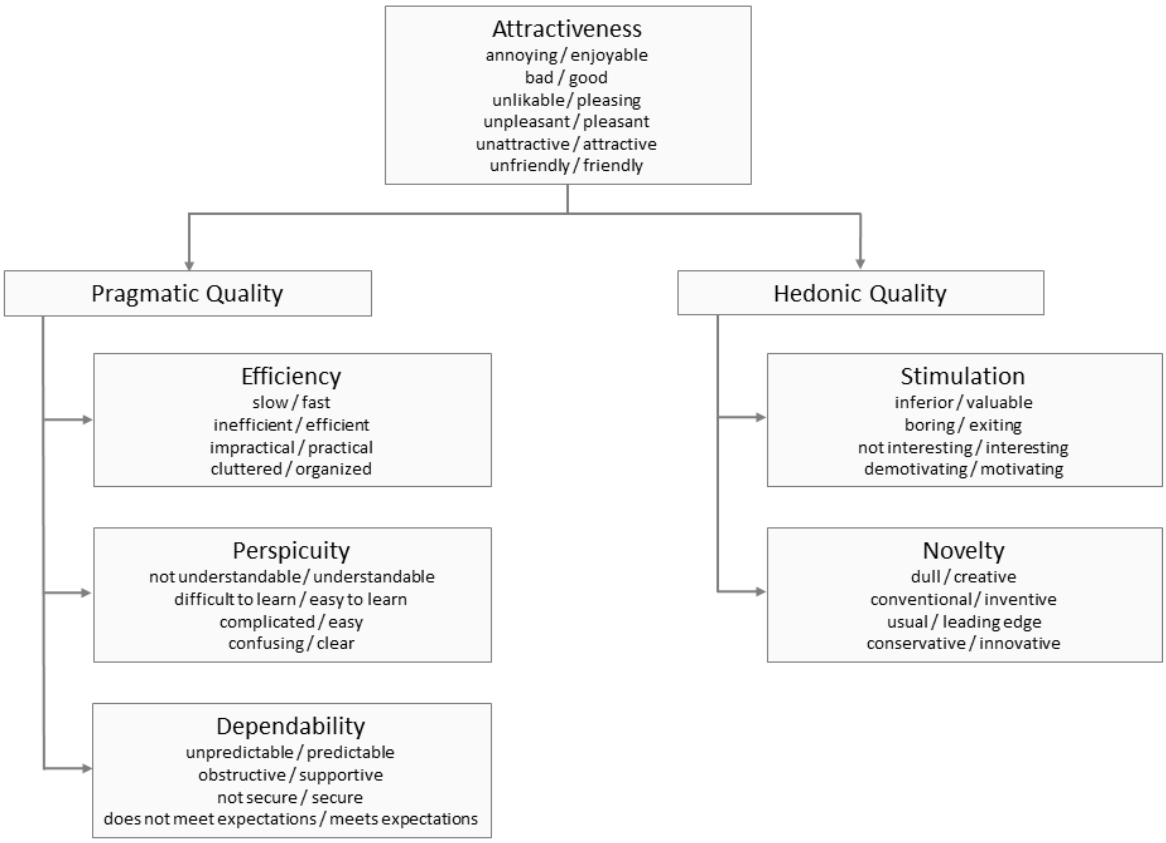

Figure 1: Assumed scale structure of the UEQ.

Typical application scenarios

Several different questions can be the reason behind the wish to measure the user experience of a product quantitatively with the UEQ.

Compare the user experience of two products

There can be several reasons to compare two products using the UEQ. A typical scenario is to compare an established product version with a new redesigned version to check if the new version has better user experience. Another scenario is to compare a product to direct competitors in the market.

Products can be compared relatively easy by a statistical comparison of two UEQ measurements. Thus, the UEQ evaluation of both products or both product versions are compared on the basis of the scale means for each UEQ scale.

Lets look at an example. Figure 2 shows a comparison of two hypothetical product versions A (new) and B (old).

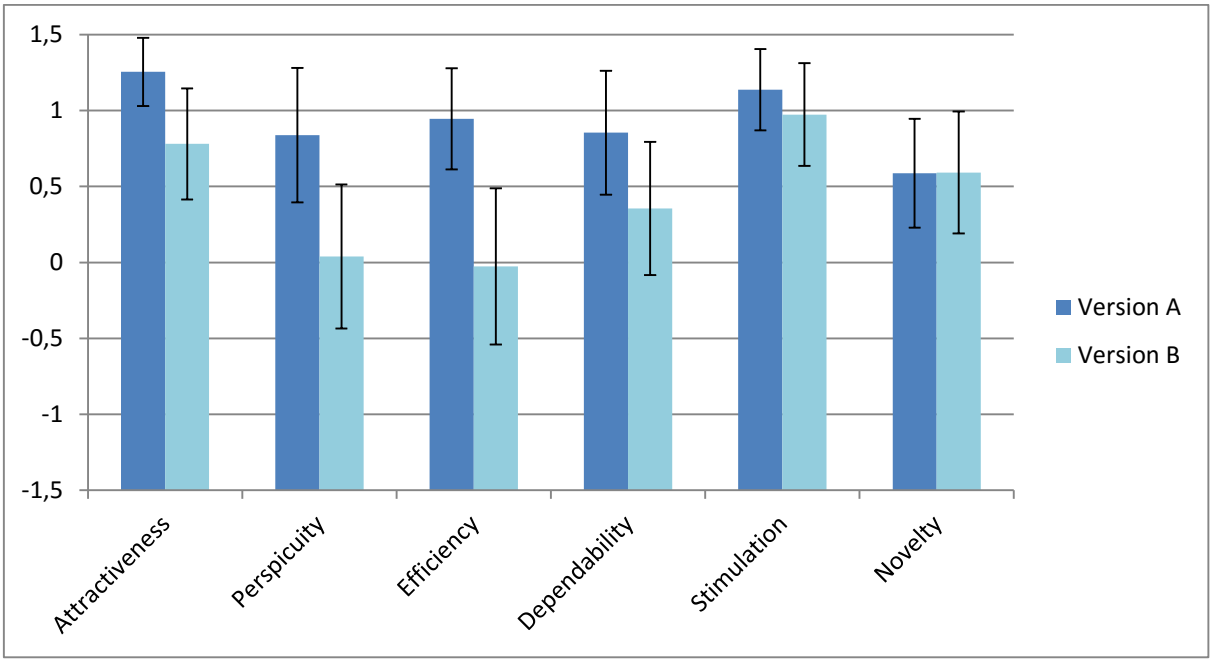

Figure 2: Comparison of two hypothetical product versions.

As we can see, the new version A shows for all scales, with the exception of *Novelty* where the values are identical, better values than the old version B. However, if you want to draw conclusions on this (especially if your sample is small) you have to check if the differences are significant.

The error bars represent the 95% confidence intervals of the scale mean. What do they mean? Assume you could repeat an evaluation infinitely often under the same conditions. Then of course due to some random influences you would not measure the exactly same scale mean in each repetition. The 95% confidence interval is the interval in which 95% of the scale means of these hypothetical repetitions are located. Thus, it shows how accurate your measurement is.

If the confidence intervals of the two measurements do not overlap, then the difference is significant on the 5% level. In our example above this is only true for the scale *Efficiency*. But the opposite conclusion is not true, i.e. *if the confidence intervals overlap the differences can still be significant*. Thus, it make sense to do a significance test (a simple two sample t-test assuming unequal variances can be done with the Excel *UEQ_Compare_Products.xlsx* available in the language packs on www.ueq-online.org or can easily be done with each statistics package).

If you compare a new version with an already used version, you should try to collect the data after the users have made themselves familiar with the new version. If you start data collection directly after the new version is launched, problems occurring from the change itself (i.e. things work in the new version differently than in the old version and users are irritated or angry about that, even if the new version is better) may influence your results heavily.

Test if a product has sufficient user experience

Does the product fulfil the general expectations concerning user experience? Such expectations of users are formed by products they frequently use.

Sometimes the answer to this question is clear directly from the results, as in the following example:

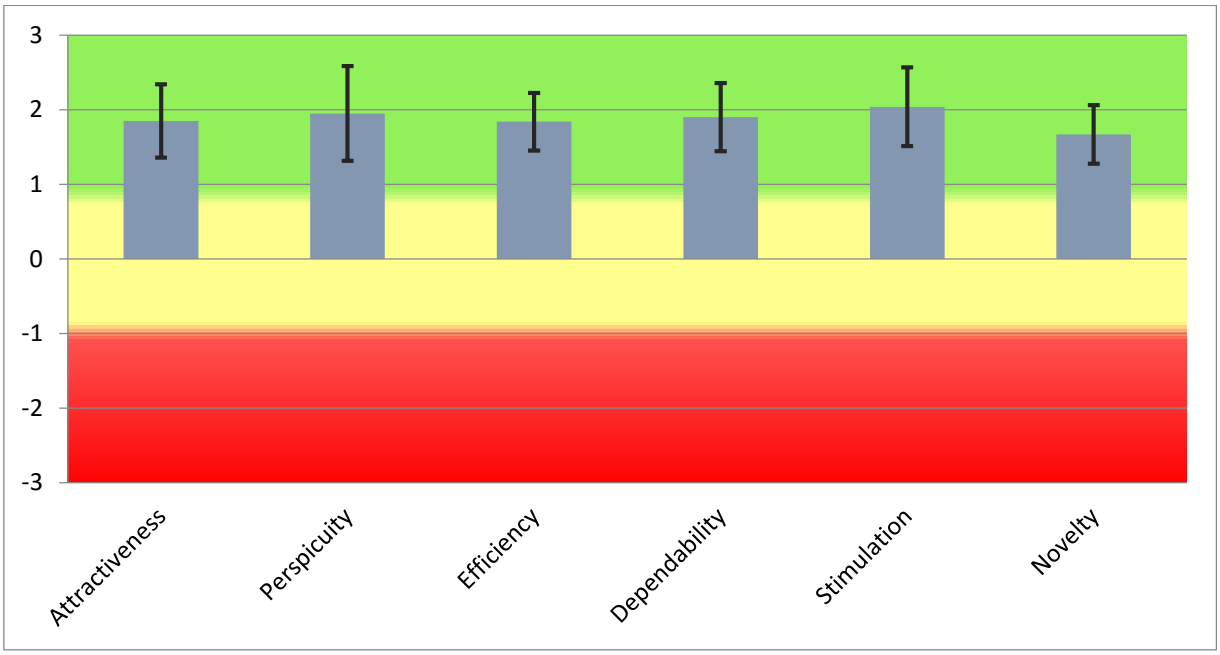

Figure 3: Example of a product with excellent results.

Here obviously all scales show an extremely positive evaluation. The standard interpretation of the scale means is that values between -0.8and 0.8 represent a neural evaluation of the corresponding scale, values > 0,8 represent a positive evaluation and values < -0,8 represent a negative evaluation.

The range of the scales is between -3 (horribly bad) and +3 (extremely good). But in real applications in general only values in a restricted range will be observed. It is due to the calculation of means over a range of different persons with different opinions and answer tendencies, for example the avoidance of extreme answer categories, extremely unlikely to observe values above +2 or below -2.

But in typical evaluations things are not so obvious. To get a better picture on the quality of a product it is thus necessary to compare the measured user experience of the product to results of other established products, for example from a benchmark data set containing quite different typical products.

The UEQ offers such a benchmark, which contains the data of 246 product evaluations with the UEQ (with a total of 9905 participants in all evaluations). The benchmark classifies a product into 5 categories (per scale):

- Excellent: In the range of the 10% best results.
- Good: 10% of the results in the benchmark data set are better and 75% of the results are worse.

- Above average: 25% of the results in the benchmark are better than the result for the evaluated product, 50% of the results are worse.
- Below average: 50% of the results in the benchmark are better than the result for the evaluated product, 25% of the results are worse.
- Bad: In the range of the 25% worst results.

The benchmark graph from the Excel-Tool shows how the UX quality of your evaluated product is.

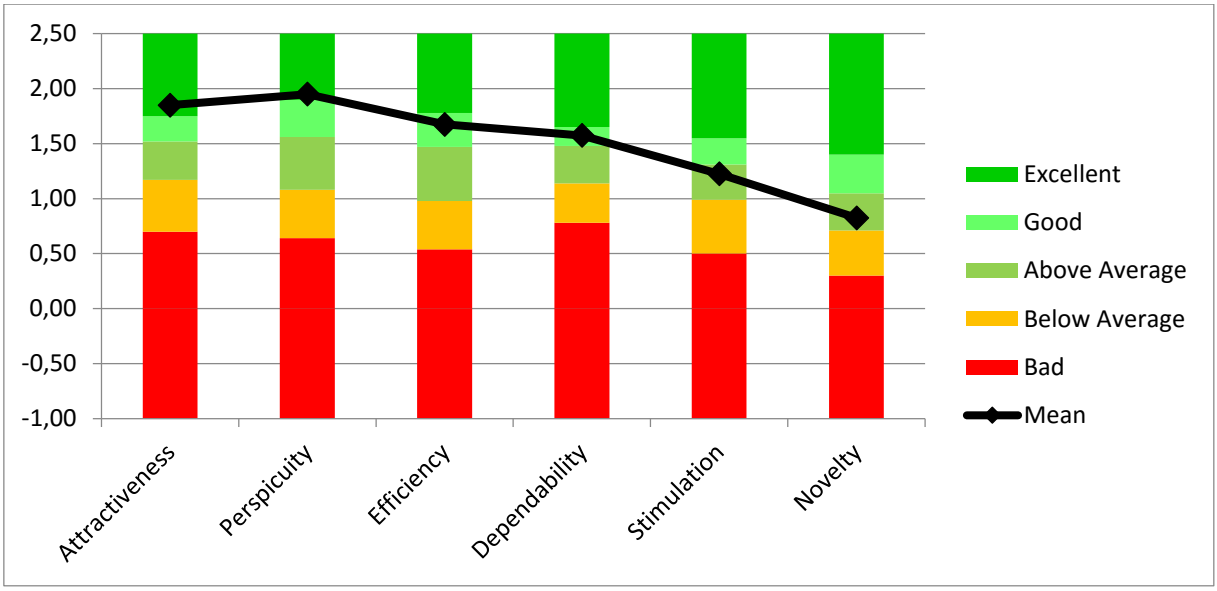

Figure 4: Benchmark graph for a hypothetical product.

Determine areas of improvement

What should be changed in order to improve the user experience of the product? This question cannot be answered directly by a quantitative measurement of user experience. To answer this question, a connection of product features to the measurement is required.

However, with a questionnaire like the UEQ it is possible to make at least educated guesses about the areas where improvements will have the highest impact. For an evaluated product, the UEQ shows a pattern of 6 measured user experience qualities. From this pattern it is possible to make at least some assumptions where to look for improvements.

More details about the application scenarios of the UEQ and some examples are contained in:

Schrepp, M.; Hinderks, A. & Thomaschewski, J. (2014). Applying the User Experience Questionnaire (UEQ) in Different Evaluation Scenarios. In: Marcus, A. (Ed.): Design, User Experience, and Usability. Theories, Methods, and Tools for Designing the User Experience. Lecture Notes in Computer Science, Volume 8517, S. 383-392, Springer International Publishing.

Details about the creation of the first benchmark can be found in (paper is only available in German):

Schrepp, M.; Olschner, S. & Schubert, U. (2013). User Experience Questionnaire Benchmark - Praxiserfahrungen zum Einsatz im Business-Umfeld. In: Brau, H.; Lehmann, A.; Petrovic, K.; Schroeder, M. (Eds.); Usability Professionals 2013 S. 348 – 353, 2013.

Translation to other languages

The UEQ is already translated into a number of languages. The complete and actual list of all available language versions can be found on www.ueq-online.org. If there is no version available for your language, then simply create one!

The following process was used in nearly all current translations and usually helps to get a good result out of the translation process.

1. Choose one of the existing language versions as a basis for the translation. We call the language of this version in the following the source language.
2. Translate the instructions and the terms of the items into your language (target language). Remember that each item consists of two terms that represent the opposites of a semantic dimension. Make sure that this property is also true after the translation.
3. Find somebody else who translates the items of the target language back to the source language. It is important that the translation into the target language and the translation back are done by two independent persons.
4. Analyse the deviations between the UEQ version you based the translation on and the version that result from translation back into the source language.
5. If possible, collect some data (preferably from several products) with the new translation and check if the Cronbach alpha values are sufficiently high. If this is not possible publish your translation on www.ueq-online.org, others may use it and contribute their data.
6. Do not forget to send us your translation! We will publish it with your name as author on the UEQ website. Even if it is not perfect, that is a good start and others may use it and share the data.

More details about the translation process and methods to check the quality of the language version are given in:

Rauschenberger, M., Schrepp, M., Cota, M.P., Olschner, S. & Thomaschewski, J. (2013). Efficient measurement of the user experience of interactive products - How to use the User Experence Questionnaire (UEQ). Example: Spanish Language Version. International Journal of Interactive Multimedia and Artificial Intelligence, Vol. 2, Nr. 1, S. 39- 45.

Online application

The UEQ is short enough to be applied online. This saves usually a lot of effort in collecting the data. However, please consider that in online studies you may have a higher percentage of persons who do not fill out the questions seriously. This is especially true if the participants get a reward (for example participation in a lottery) for filling out the questionnaire.

A simple strategy to filter out suspicious responses is based on the fact that all items in a scale more or less measure the same quality aspect. Thus, the responses to these items should be at least not too different.

As an example look at the following responses to the items of the scale *Perspicuity*:

not understandable	o o o o o x o	understandable
easy to learn	o o o o o o x	difficult to learn
complicated	o o o o x o o	easy
clear	o o o o o x o	confusing

Obviously, these answers are not very consistent. If they are transfered to the order negative (1) to positive (7), then we can see that the ratings vary from 1 to 6, i.e. the distance between the best and worst answer is 5. Thus, a high distance between the best and the worst answer is an indicator for an inconsistent or random answer behaviour.

If such a high distance occur only for a single scale this is not really a reason to exclude the answers of a participant, since such situations can also result from response errors or a simple misunderstanding of a single item. If this occurs for several scales, then it is likely that the participant has answered at least a part of the questionnaire not seriously.

Thus, a simple heuristic is to consider a response as suspicious if for 2 or 3 scales (it is open to your decision how strict you will apply this rule) the distance between best and worst response to an item in the scale exceeds 3.

This heuristic is also implemented in the Excel-Tool in one of the worksheets.

Applying the UEQ as part of a Usability Test

Often the UEQ is used as part of a classical usability test to collect some quantitative data about the impression of the participants concerning user experience. The best point in time to handle the questionnaire to the participants is directly after they finished working on the test tasks. If participants fill the questionnaire only after they had a lengthy discussion about the product with the person conducting the test, this will influence the results. The goal of the UEQ is to catch the immediate impression of a user towards a product. Thus, try to get the answers to the UEQ before you discuss with the participants.

Some of the participants may be unfamiliar with the special item format of the questionnaire. Mention that this is a scientifically evaluated questionnaire to measure user experience when you hand it over to the participant. This will improve the quality and consistency of the answers.

How to use the Excel-Tool

The goal of the Excel-Tool is to make the analysis of UEQ data as easy as possible for you. You just need to enter the data in the corresponding work sheet and then all relevant computations (with the exception of significance tests if you want to compare two products) are done automatically. The Excel-Tool contains comments that explain the different calculations, thus we need not to go into detail here. Make sure that you always use the most actual version of the Excel-Tool that is available on www.ueq-online.org in each language pack (File UEQ_Data_Analysis_Tool.xlsx), since the UEQ team tries to continuously improve this tool based on user feedback. A test to compare two products is possible in the Excel UEQ_Compare_Products.xlsx that is also available on www.ueq-online.org.

How to interpret the data

What do the error bars mean?

The error bar of a scale shows the 95% confidence interval of the scale mean. The error bars are displayed together with the scale means in the corresponding diagram of the Excel-Tool.

Assume you could repeat an evaluation often under the same conditions. Then of course due to some random influences you would not measure the exactly same scale mean in each repetition. The error bar describes the interval in which 95% of the scale means of these repetitions will be located. Thus, it shows how accurate your measurement is. The size of the error bar depends on the sample size (the more participants you have the smaller typically the error bar) and on how much the different participants agree (the higher the level of agreement, i.e. the more similar the answers are, the smaller is the error bar).

Thus, if the confidence interval is relatively large you should interpret your results carefully. In this case your measurement may not be very accurate. Typically this is due to a too small sample size, i.e. if possible you should collect some more data.

What does the Cronbach-Alpha values mean?

One of the worksheets in the Excel-Tool shows the Cronbach-Alpha coefficients ($\alpha = n * r/1 + (n-1) * r$, where *r* is the mean correlation of the items in a scale and *n* is the number of items in a scale) for the six scales of the UEQ. The Alpha-Coefficient is a measure for the consistency of a scale, i.e. it indicates that all items in a scale measure a similar construct.

There are no clear rules that describe how big the Alpha-Coefficient should be. Some rules of thumb consider values >0.6 or >0.7 as a sufficient level. If the Alpha-Value of one of the scales is too small, then you should interpret this scale carefully.

There are two main reasons for small Alpha-values. First, a higher number of participants may misinterpret some items in the scale. For example, the item *unsecure / secure* is associated with the scale *Dependability*. It usually is interpreted in the sense that the interaction is save and controllable by the user. In the context of social networks this item can be misinterpreted as *"Are my data secure?"*, i.e. the item gets a non-intended meaning in this context. This lowers the correlations to the other items in the scale and therefore the Alpha-value. Second, a scale can be irrelevant for a certain product. In this case the answers of the persons will be not very consistent, since participants will have problems to judge a UX quality aspect that is for the product under investigation not important. This can also lead to low Alpha-Values. It is clear from these examples that in such cases the scale mean should be interpreted with care.

If the Alpha-value for a scale is small, it makes sense to look at the means of the single items. There you can sometimes directly see if some item is not interpreted in the usual way. The next figure shows such a case as an example.

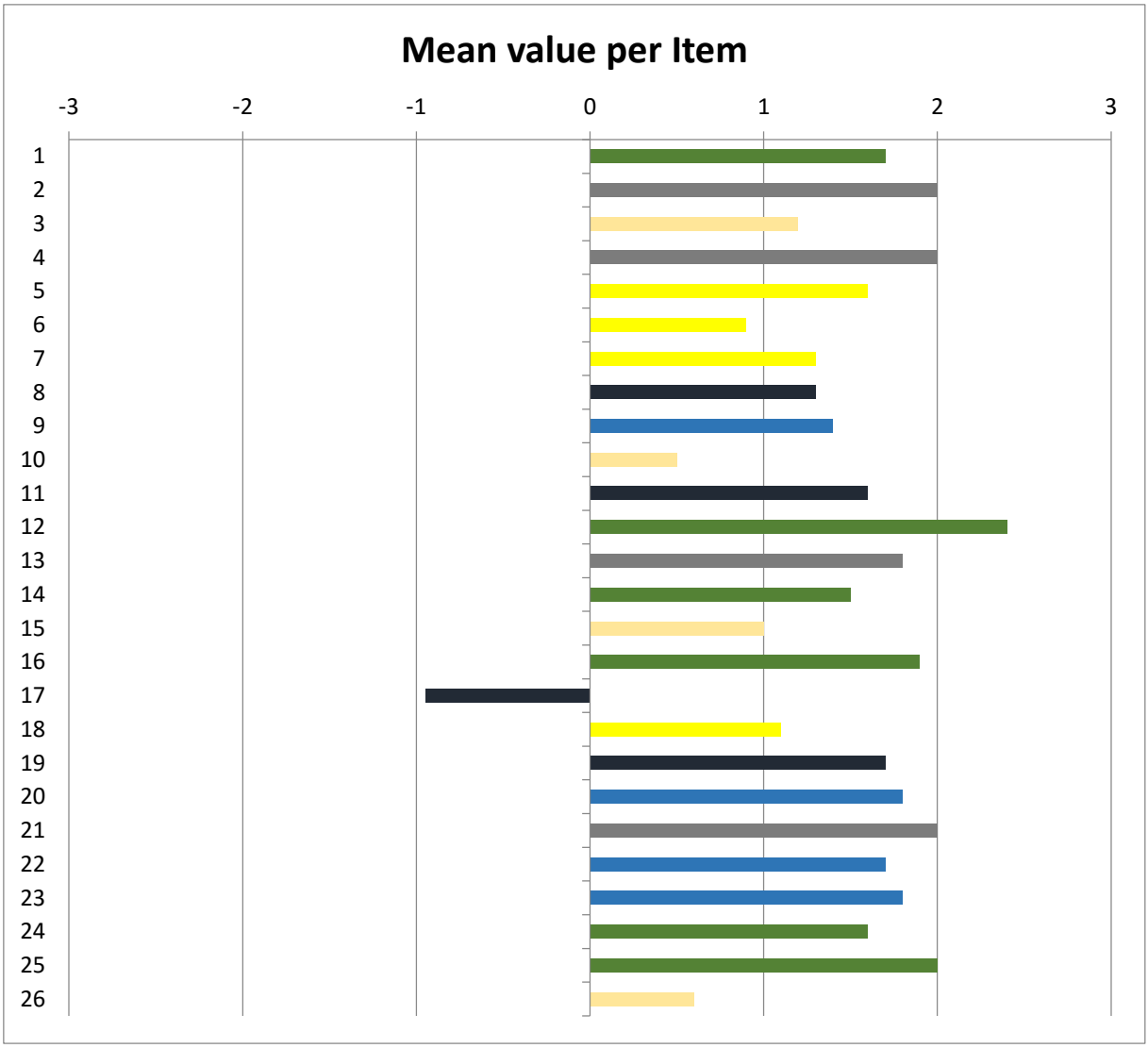

Figure 5: Typical example of an item that is misinterpreted in the given context.

This example shows the result of an application of the UEQ in the context of a social network. It is obvious that the item 17 (not secure / secure) has a negative mean, while all other items of this scale (black bars) have a highly positive mean. This shows that there is maybe a problem with this item in this context.

Typical questions

Can I change some items?

You should NOT change single items or leave out some items in a scale! If you do this it is very difficult to interpret your results, for example the benchmark values that are calculated based on the original items should not be used, simply because the answers are not comparable.

Do I need all scales?

You can leave out complete scales, i.e. delete all items of a certain scale from the questionnaire. This can make sense to shorten the questionnaire in cases where it is clear that a certain scale is not of interest.

How long do participants need to fill out the questionnaire?

Applying the UEQ does not require much effort. Usually 3-5 minutes are sufficient for a participant to read the instruction and to complete the questionnaire.

How many data do I need?

The more data you have collected the better and more stable will be the scale means and thus the saver will be the conclusions you draw from these data. However, it is not possible to give a minimum number of data you need to collect to get reliable results. How many data you need does depend also on the level of agreement of the users that participate in the questionnaire (the standard deviation per scale). The more they agree, i.e. the lower the standard deviation of the answers to the items is, the less data you need for reliable results. For typical products evaluated so far around 20-30 persons already give quite stable results.

The Excel-Tool for data analysis contains a worksheet named Sample_Size. Here the standard deviation per scale is used to estimate how many data you need to reach a certain precision (measured by the width of the confidence interval) in your measurement. Obviously, the precision depends on the conclusions you want to draw from the data. For typical product evaluations a precision of 0.5 seems to be adequate (see the detailed explanations in the worksheet).

How to sell it to customers and management

If you report to your management or other stakeholders, it is important to communicate the meaning of the UEQ scales. If you evaluate a financial software or another business tool it may look strange to your stakeholders if you report on *Stimulation* or *Originality*. Do they want an *original* accounting system? Probably not.

If necessary, change the scale names and explain clearly what each scale means. In the example above you may want to change *Stimulation* to *Fun of use* and *Originality* into *Interest*. Use terms that fit the language of your stakeholders. Important is the semantic meaning of the scales, i.e. if you change a scale name make sure the new name still covers the meaning of the scale.

Can I calculate a UX KPI from the results

Managers love KPI's. They reduce complex constructs to a single numerical value and thus give the impression that things can be controlled and improved by measuring and optimizing this single KPI. However, for a construct like UX that combines a high number of quite distinct quality aspects, this is not easily possible.

The UEQ gives back 6 scale values that can be interpreted. It is from the design of the questionnaire not possible to combine these into a single KPI. What can maybe theoretically justified (but even this is a little bit questionable) is to compute a value for *Pragmatic Quality* from the three scales *Efficiency*, *Perspicuity* and *Dependability* and a value for *Hedonic Quality* from *Stimulation* and *Novelty*.

To compute a single UX KPI from the UEQ results requires knowledge about the relative importance of the UEQ scales to the overall impression concerning UX. One method to do this is to add some questions that describe the content of the 6 scales in plain language and to ask the participants how important these scales are for their impression on UX. These ratings

about the importance can then be combined with the scale means to compute the desired UX KPI. How important a quality aspect is for the overall UX impression does obviously depend on the concrete product evaluated. Thus, it is not possible to provide any general data about importance of scales.