Construction and Evaluation of a User Experience Questionnaire

Bettina Laugwitz, Theo Held, and Martin Schrepp

SAP AG
Dietmar-Hopp-Allee 16, 69190 Walldorf, Germany
bettina.laugwitz@sap.com,
theo.held@sap.com,
martin.schrepp@sap.com

Abstract. An end-user questionnaire to measure user experience quickly in a simple and immediate way while covering a preferably comprehensive impression of the product user experience was the goal of the reported construction process. An empirical approach for the item selection was used to ensure practical relevance of items. Usability experts collected terms and statements on user experience and usability, including 'hard' as well as 'soft' aspects. These statements were consolidated and transformed into a first questionnaire version containing 80 bipolar items. It was used to measure the user experience of software products in several empirical studies. Data were subjected to a factor analysis which resulted in the construction of a 26 item questionnaire including the six factors Attractiveness, Perspicuity, Efficiency, Dependability, Stimulation, and Novelty. Studies conducted for the original German questionnaire and an English version indicate a satisfactory level of reliability and construct validity.

Keywords: User experience; Software evaluation; User satisfaction; Questionnaire; Usability assessment; Perceived usability.

1 Introduction

Questionnaires are a commonly used tool for the user-driven assessment of software quality and usability. They allow an efficient quantitative measurement of product features.

Some questionnaires can under certain circumstances be used as a stand-alone evaluation method, for example the IsoMetrics questionnaire [1]. But in general, user questionnaires have to be combined with other quality assessment methods to achieve interpretable results (see e. g. [2]). In such a context, some usability questionnaires provide rough indicators for certain product features [3], while others are designed to discover specific usability problems (e. g. SUMI, see [4]).

In any case, the results have to be interpreted by a trained usability expert, taking into account also the results from other assessment methods that have been used.

The quantitative data from an assessment done by the users of a product can be a useful addition to methods that allow a sophisticated assessment of the strengths and weaknesses of interactive products, like for example usability tests or heuristic evaluation methods [5].

A. Holzinger (Ed.): USAB 2008, LNCS 5298, pp. 63–76, 2008.

A very effective way to get helpful feedback by end-users is to allow them to assess what concerns them most immediately: How did the interaction with the product feel, how was the use experience? This does not only include usability aspects as they are described by ISO 9241-10 [6] or the criteria of effectiveness or efficiency according to ISO 9241-11 [7]. The more fuzzy criteria that are subsumed under the concept of *user experience goals* [8] are an even more promising subject to a questionnaire assessment done by the users themselves. These criteria are for example reflected in the concepts of hedonic quality [9] or user satisfaction according to ISO 9241-11 [7] (for a deeper discussion on user satisfaction see e. g. [10]).

The objective of the construction process described below was to develop a questionnaire that allows a *quick assessment* done by end users covering a preferably *comprehensive impression of user experience*. It should allow the users in a very *simple and immediate way* to express feelings, impressions, and attitudes that arise when experiencing the product under investigation.

The available questionnaires lay emphasis on one or two of the mentioned criteria but none meets all three requirements. This paper contains an overview over the objectives, theoretical assumptions, and procedure of the construction process as well as the results of some validation studies investigating the quality of the questionnaire.

2 Construction of the Questionnaire

2.1 Objectives

Quick assessment: Generally, questionnaires are a particularly efficient method to apply and analyze. The application of some questionnaires may nevertheless be rather time consuming when the absolute amount of time is considered. With the SUMI questionnaire [4] the users have to decide on their level of agreement with 50 statements on usability. The long version of IsoMetrics [1] requires ratings for 75 different items. In these cases, the goal is to achieve a comprehensive usability evaluation including detailed descriptions of particular usability problems, on the sole basis of the questionnaire data. This is not what our questionnaire aims at. Rather, it is supposed to be an efficient tool to enhance the results from expert evaluations or usability testings.

Comprehensive impression of user experience: Traditional methods often focus on usability criteria in a narrower sense, which correspond roughly to the concepts of usability goals [8] or pragmatic quality [9]. More recent approaches increasingly give attention to the subjective reactions, also including emotional aspects of the user's experience, which can be subsumed under the concept of user satisfaction as outlined in ISO 9241-11 [7]. These criteria are also referred to as user experience goals [8], or as hedonic quality aspects [9]. A discussion of relevant usability criteria for special user groups, for example elderly persons, can be found in [11].

According to Norman [12] product design affects users on three levels of information processing, namely on a visceral level, on a behavioral level, and on a reflective level. This implies that usability criteria do not cover all aspects relevant for the user experience. This is also supported by studies (for example [13]) which show that

there is a dependency between aesthetic impression of a user interface and its perceived usability.

It could be shown that semantic differentials for assessing the pragmatic and hedonic quality (e. g. [9]) are applicable not only to the evaluation of websites or games but also for business software [14]. However, this particular questionnaire (AttrakDiff2) lays a greater emphasis on the hedonic aspects of product quality than on the pragmatic aspects. This may not be perfectly appropriate for a comprehensive evaluation of professional software. A contrary perspective is represented by the SUMI questionnaire [4]. Here only one of six scales aims at the measurement of emotional aspects.

An overall picture has to include as many product aspects and features as possible that are of relevance for the user. For the new questionnaire no potential (hedonic or pragmatic) criteria should be excluded or favored a priori. The initial item pool should include a range of criteria as wide as possible, reduction and selection taking place on the basis of empirical data using an explorative factor analysis.

Simple and immediate: How does the interaction with the product feel? Which were the most striking features of the product and of the interaction? The user should be enabled to give his rating about the product as immediately and spontaneously as possible. A deeper rational analysis should be avoided.

The questionnaire should not force the user to make abstract statements about the interaction experience or remember details that are likely to be forgotten or had been overlooked in the first place. An explicit evaluation demanded by the user retrospectively is not always reliable (see e.g. [15]). This is supported by results [16] where differently colored UIs affected users' feelings differently (e. g. as measured with a mood questionnaire), while this difference was not reflected by users' answers on questions regarding the UI quality.

Experts are able to evaluate user interfaces in detail. Detailed data can also be gained from the observation of a user when interacting with the product.

Thus, a user questionnaire can lay its emphasis on criteria which are accessible immediately: the user's subjective perception of product features and their immediate impact on the user him/herself.

2.2 Theoretical Background

For the construction of our questionnaire we rely on a theoretical framework of user experience [3]. This research framework distinguishes between perceived ergonomic quality, perceived hedonic quality and perceived attractiveness of a product. The framework assumes that perceived ergonomic quality and perceived hedonic quality describe independent dimensions of the user experience.

Ergonomic quality and hedonic quality are categories that summarize different quality aspects. The focus of ergonomic quality is on the goal oriented or task oriented aspects of product design. High ergonomic quality enables the user to reach his or her goals with efficiency and effectiveness. The focus of hedonic quality is on the non-task oriented quality aspects of a software product, for example the originality of the design or the beauty of the user interface.

Thus, it is assumed that persons perceive several distinct aspects when they evaluate a software product. The perceived attractiveness of the product is then a result of an averaging process from the perceived quality of the software concerning the relevant aspects in a given usage scenario.

According to this assumption the constructed questionnaire should contain two classes of items:

- items, which measure the perceived attractiveness directly,
- items, which measure the quality of the product on the relevant aspects.

2.3 Generation of the Item Pool

Two brainstorming sessions (each lasting about one and a half hours) with fifteen SAP usability experts were conducted. The experts were asked to propose terms they suppose to be characteristic for the assessment of user experience. A moderator took down the proposed terms. The experts were asked the following questions:

- To which properties of products are users particularly responsive?
- Which feelings or attitudes of users are caused by products?
- What are the typical reactions of users during or after usability studies?

All redundant answers were removed from the list of the initial 229 expert proposals. All proposals that were not formulated as adjective were replaced by the corresponding best fitting adjective. The consolidated cleaned up list consisted of 221 adjectives.

Seven usability experts then individually extracted a "top 25" list out of the whole set of terms. In addition, they marked terms they considered to be inappropriate with a "veto" (unlimited number). Adjectives that received more than one veto or occurred less than twice in the top 25 lists were removed.

After this procedure a set of 80 adjectives remained. Since the target format of the questionnaire is a semantic differential, the best fitting antonym for each of the 80 adjectives had to be identified. The sequence of adjective pairs and the polarity of each pair was then determined randomly. In addition, a second version of the list with complementary order and polarities was prepared.

Both lists had the format of a seven stage semantic differential (another example of an application of semantic differentials in product design can be found in [17]).

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 ① ② ③ ④ ⑤ ⑥ ⑦ unattractive

2.4 Data Collection

In order to examine the specific properties of the adjective pairs concerning the assessment of software products, the eighty items raw-version of the questionnaire was used in six investigations. In the following, each of the six investigations is briefly explained.

- SYSTAT (number of participants N=27; location: University of Mannheim; paper-pencil version of the questionnaire): The participants of an introductory course for the statistics software package Systat were asked to perform a given task with the

software or to observe a person that works on the task. After that the participants completed the questionnaire in order to assess the software quality.

- Cell Phone (*N*=48; University of Mannheim; paper-pencil): The participants of a psychology class were asked to add an entry to the address book of their cell phone and then to delete this entry. This application should then be evaluated with the questionnaire.
- BSCW (*N*=14; University of Mannheim; paper-pencil): Students rated the online-collaboration software BSCW that had been used during a lecture. Each of the participants had worked actively with the software before completing the questionnaire.
- Selection (*N*=26; University of Mannheim; paper-pencil): The participants of a computer-science course had the choice to assess one of the following products: Eclipse Development Workbench, Borland JBuilder, Microsoft Visual Studio, Mozilla 1.7 Browser, Microsoft Internet Explorer 6, and Firefox 1.0. Ratings were provided for Firefox 1.0, Microsoft Internet Explorer 6, and the Eclipse Workbench.
- CRM Mobile (*N*=15; SAP AG, Walldorf; paper-pencil): During a regular meeting of SAP usability experts, a user interface variant of the SAP Customer Relationship Management (CRM) software was demonstrated. The experts filled out the questionnaire after the demonstration.
- CRM PC (*N*=23; SAP AG, Walldorf, online version of the questionnaire): An online investigation consisting of a short demonstration of a further variant of SAP CRM and the electronic version of the questionnaire was conducted with SAP usability experts.

All in all, 153 participants provided complete datasets. 76 of the participants had completed the first version of the questionnaire while 77 had completed the second version (see above). Those data were used for the process of item reduction as described in the following section.

2.5 Reduction of the Item Pool

As described above the questionnaire should contain items that measure the perceived attractiveness directly and items that measure the quality of the product on the relevant aspects.

For this reason the item set was split into two subsets. The first subset contains 14 items that represent an emotional reaction on a pure acceptance/rejection dimension. These items of valence do not provide any information concerning the reason for the acceptance or rejection of the product. Examples for items from the first subset are *good/bad* or *pleasant/unpleasant*. The second subset contains the remaining 66 items from the item pool.

A factor analysis (principal components, varimax rotation) of the first subset of items extracted one factor concerning the Kaiser-Guttman criterion[1]. This factor explained 60% of the observed variance in the data. This factor is called *Attractiveness*. To represent this factor in the questionnaire we picked the six items with the highest

[1] If we apply the scree test [18] as a decision criterion to determine the number of factors also only a single factor results from the analysis.

loading on the factor. The original German items and their English translations can be found in Appendix 1 (for details on the translation procedure see chapter 2.6).

A factor analysis (principal components, varimax rotation) of the second subset of items extracted five factors. The scree test was used to determine the number of factors[2]. These five extracted factors explain 53% of the observed variance in the data[3]. We named these factors according to the items that showed the highest factor loadings as *Perspicuity* (examples for items: easy to learn, easy to understand), *Dependability* (predictable, secure), *Efficiency* (fast, organized), *Novelty* (creative, innovative) and *Stimulation* (exiting, interesting).

Per factor, we chose four items to represent this factor in the questionnaire. Those items were selected that had high loadings on the respective factor and low loadings on all other factors. The original German items and their English translations can also be found in Appendix 1.

All items that were not selected to represent one of these five factors were eliminated from the data matrix. The reduced data set was now again analyzed by a factor analysis (principal components, varimax rotation).

This analysis extracted again five factors according to the scree test. These five factors explained 70% of the variance in the reduced data set. The table containing the loadings of the items of the second subset[4] on these factors can be found in Appendix 2.

For the final questionnaire we randomized the order of the remaining 26 items. In addition the polarity of the items (i.e. the order of the positive or negative term per item) was randomized.

The final questionnaire contains thus the scales Attractiveness (six items), Perspicuity, Dependability, Efficiency, Novelty and Stimulation (four items each). We call this questionnaire in the following *User Experience Questionnaire* (UEQ).

To guarantee an efficient handling of data a tool (based on Excel) was developed that calculates the scale means and basic statistics from collected questionnaires.

2.6 Creation of an English Version

The basic version of the questionnaire was prepared in German language. In order to develop an equivalent English version, the following procedure was applied.

In a first step, the German version was translated by a native English speaker. The results of this first translation were checked by a group of native English speakers.

According to this feedback, a reworked version was created. The new version was translated back to German language by a professional translator (native German speaker). The differences between the re-translated German version and the original German version were examined and discussed with the translator as well as the native English speakers. Based on this last consolidation, the final English version was created. For first empirical data on the quality of the English version see 3.3.

---

[2] We choose the scree test since the Kaiser-Guttman criterion tends to extract too many factors in item sets that contain a large number of items. For our data set the Kaiser-Guttman criterion would lead to a solution with 13 factors.

[3] The variance explained by each factor is 28.7% for the first, 11.1% for the second, 5.3% for the third, 4.5% for the fourth, and 3.3% for the fifth extracted factor.

[4] The items representing the factor Attractiveness are not contained in the table. These items show, as expected, high loadings on all factors.

3 Validity of the Questionnaire

Concerning the validity of the questionnaire we are currently able to report data from two usability studies.

3.1 Validation Study 1

As described above the design of the UEQ fits perfectly into an existing research framework on user experience [3]. Perspicuity, Efficiency and Dependability represent ergonomic quality aspects. Stimulation and Novelty represent hedonic quality aspects.

The task oriented aspects Perspicuity, Efficiency and Dependability should show a strongly negative correlation with task completion time. The faster a user can solve his or her tasks with a software product the higher should be his or her rating concerning these ergonomic quality aspects.

On the other hand we expect no substantial correlation of the non-task related aspects Stimulation and Novelty with task completion time. We tested these two hypotheses in a usability test.

Participants. The 13 participants were recruited during the 2005 annual conference of the German SAP User Group (DSAG). They were not paid for their participation. All had high experience using computers, and experience with SAP software.

Procedure. The participants had to walk through a scenario that contained typical tasks of a sales representative. The scenario for the test was described to the participants in a step-by-step instructional document. The scenario contained a number of typical tasks a sales representative has to perform frequently during his or her daily job (plan customer visits, search for contact persons, find the last customer interactions, etc.). Each task was motivated by a little story, which explained the context of the task and why the task is performed.

Each test session was conducted as follows:

1. The participant was greeted and guided to the test station.
2. The moderators introduced themselves and collected basic demographic data.
3. The participant was given an overview of the test session and about the intention of the test.
4. The participant was then asked to solve the described tasks. The tasks description was available on paper during the whole session. The participant was instructed to think aloud during his or her attempt to solve the tasks.

After the participant finished the last task, the screen was turned off and the participant filled out the User Experience Questionnaire.

The screen was turned on again and the participant had the chance to discuss usability problems of the software and to ask questions.

The moderators asked follow-up questions related to the usability problems they observed during the test .

The total time required by participants to solve all tasks varied between 33 and 65 minutes (M = 41.62 minutes, SD = 9.64 minutes).

Results. Table 1 shows the correlations of the observed task completion times and the observed values of the UEQ scales. As a measure of scale reliability we give in addition Cronbach's alpha coefficient per scale.

Table 1. Correlation of the UEQ scales with the observed task completion times and Cronbach's alpha per scale

UEQ Scale	Correlation with task completion time	Cronbach's Alpha
Attractiveness	-.54	.89
Perspicuity	-.66 *	.82
Efficiency	-.73 *	.73
Dependability	-.65 *	.65
Stimulation	.10	.76
Novelty	.29	.83

* Significant with $p < .05$.

The correlations show the expected pattern. Perspicuity, Efficiency and Dependability show a significant correlation ($p < .05$) with task completion time. Novelty and Stimulation show only a weak correlation with task completion time.

Thus, our hypotheses do not have to be rejected. This can be seen as a first indicator for the validity of the questionnaire. The values of Cronbach's Alpha coefficient are an indicator for a sufficient reliability, but here we have to consider that the number of test participants was only small.

3.2 Validation Study 2

In a second validation study we investigated the relation of the UEQ scales to the scales of the AttrakDiff2 questionnaire [9]. This questionnaire was developed inside the above mentioned research framework from Hassenzahl [3]. It contains the scales Pragmatic Quality, Hedonic Quality (which is here split into the two sub-aspects Identity and Stimulation) and Attractiveness.

The concept behind the Attractiveness scales is nearly identical in both questionnaires. These scales should thus show a high positive correlation. In addition we can expect that the UEQ scales Perspicuity, Efficiency and Dependability show a high positive correlation to the AttrakDiff2 scale Pragmatic Quality. The UEQ scales Novelty and Stimulation should show a high positive correlation with the AttrakDiff2 scale Stimulation.

The concept behind the AttrakDiff2 scale Identity is quite different to the concept of any of the UEQ scales. For this scale we can thus not formulate any hypothesis concerning its dependency to the UEQ scales. We tested our hypothesis again in a usability test.

Participants. 16 students of the University of Cooperative Education in Mannheim, Germany, participated in this test. All had sufficient experience using computers. The participants were not paid for their participation in the study.

Procedure. The participants had to walk through a scenario which contained typical tasks in a CRM system (create a new account, create activities with the account, search for data of already existing accounts, etc.). The scenario for the test was described to the participants in a step-by-step instructional document. Each task was motivated by a little story, which explained the context of the task and why the task is performed.

The procedure for the test sessions was identical to the one for validation study 1 including the task completion step (step 4, see 3.1). After that, the sessions proceeded as follows:

5. Immediately after the participant finished the last task, the screen was turned off. Eight of the participants filled the UEQ and eight of the participants filled out the AttrakDiff2 at this point in time. It was randomly determined per participant to which of these two groups he or she was assigned.
6. The screen was turned on again and for around 30 minutes the participant and the moderator discussed about usability problems which were observed during the test session.
7. The participants that had already filled out the UEQ were now asked to fill the AttrakDiff2 and vice versa. Thus, each participant evaluated the tested user interface with the UEQ and with the AttrakDiff2 questionnaire. Since some of the items in both questionnaires are similar the delay introduced by step 6 is intended to reduce dependencies between the two evaluations.

Results. Table 2 shows the correlations of the UEQ scales with the AttrakDiff2 scales. The results show the expected pattern. The UEQ scales Perspicuity, Efficiency and Dependability show a significant correlation with the AttrakDiff2 scale Pragmatic Quality. The AttrakDiff2 scale Stimulation shows a high correlation with the UEQ scales Novelty and Stimulation.

The AttrakDiff2 scale Identity shows a high positive correlation with the UEQ scale Dependability, but no significant correlation with the UEQ scales Novelty and Stimulation.

Table 2. Correlations of the single scales from the User Experience Questionnaire and the scales of the AttrakDiff2 questionnaire

		User Experience Questionnaire (UEQ)					
		Attractiveness	Perspicuity	Efficiency	Dependability	Stimulation	Novelty
AttrakDiff2	Attractiveness	.72 *	.56 *	.30	.51 *	.51 *	.40
	Pragmatic Quality	.33	.73 *	.59 *	.54 *	.31	.07
	Identity	.45	.45	.29	.62 *	.30	.32
	Stimulation	.42	-.17	-.40	-.14	.72 *	.64 *

* Significant with $p < .05$.

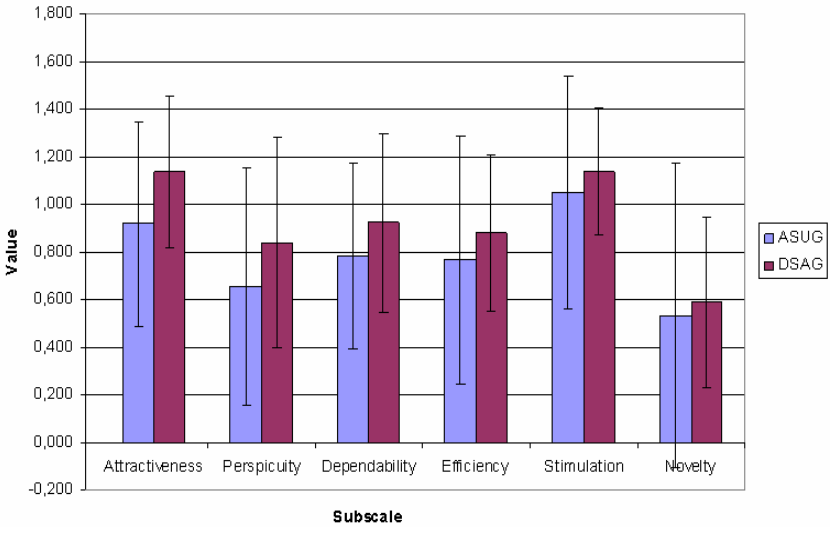

Fig. 1. Questionnaire scores from two parallel investigations. Investigation "ASUG" has been conducted at a conference of the American SAP User Group, while "DSAG" ran at a conference of the German SAP User Group. The raw data have been transformed so that the final data may range from -3 to +3. The error bars represent the 95% confidence interval for each arithmetic mean.

Thus, our hypothesis does not have to be rejected. This is again an indicator concerning the validity of the UEQ questionnaire. But again we have to mention that the number of participants in the study was small, so these results need to be confirmed in bigger validation studies.

3.3 First Data on an English Version

Though this has not yet been tested systematically, there are indicators that the language versions are sufficiently equivalent. For instance, two parallel investigations, one conducted in Germany and one in the US with the respective questionnaire versions delivered questionnaire scores as shown in Figure 1.

The one investigation was conducted at the 2005 fall conference of the American SAP User Group (ASUG), while the other investigation ran at the annual conference of the German SAP User Group (DSAG). The scenario and the SAP system were the same in both investigations; the only difference was the user interface language. The differences of the average scores on the different dimensions appear to be only marginal.

In another investigation, only the English version of the UEQ was used. This investigation was conducted as an online study with 21 participants who had tested a new software product for about one week. Each of the participants filled out the UEQ at the end of the testing period. In order to get an indicator for the reliability of the questionnaire, the Cronbach's Alpha coefficient was calculated for each of the subscales. Table 3 displays those values.

Table 3. Cronbach Alpha values for an investigation conducted with the English version of the UEQ

UEQ Scala	Cronbach's Alpha
Attractivity	.86
Perspicuity	.71
Efficiency	.79
Dependability	.69
Stimulation	.88
Novelty	.84

Except for the subscale Dependability, in each of the other cases the Alpha value exceeds the threshold of .7. According to this result, it may be assumed that the reliability of the English version of the questionnaire is sufficiently high.

4 Conclusions

For the construction of the user experience questionnaire UEQ the process should ensure that as many relevant product features as possible were taken into account. The factors revealed by the factor analysis support the assumption that ‚soft' (user experience) criteria and 'hard' (usability) criteria are of similar relevance for the end user (two scales and three scales, respectively). This fact is not reflected adequately by the structure of other user feedback questionnaires.

Studies reported here indicate a satisfactory level of reliability and construct validity. Data from the English and the German version of the questionnaire that have been collected in parallel studies confirm a good congruence of both language versions.

The user experience questionnaire UEQ in its current form appears to be an easy to apply, reliable and valid measure for user experience that can be used to complement data from other evaluation methods with subjective quality ratings. Nevertheless, further research will be done to provide a more detailed and extensive picture of UEQ's features from a methodical as well as from a practical point of view. In particular, the overall factor structure and the relative weakness of the "Dependability" scale will be in the focus of future studies.

References

[1] Gediga, G., Hamborg, K.-C., Düntsch, I.: The IsoMetrics Usability Inventory: An operationalisation of ISO 9241-10. Behaviour and Information Technology 18, 151–164 (1999)

[2] Dzida, W., Hofmann, B., Freitag, R., Redtenbacher, W., Baggen, R., Geis, T., Beimel, J., Zurheiden, C., Hampe-Neteler, W., Hartwig, R., Peters, H.: Gebrauchstauglichkeit von Software: ErgoNorm: Ein Verfahren zur Konformitätsprüfung von Software auf der Grundlage von DIN EN ISO 9241 Teile 10 und 11, Schriftenreihe der Bundesanstalt für Arbeitschutz und Arbeitsmedizin [Usability of Software: ErgoNorm: A method to check software conformity on the basis of DIN EN ISO 9241 parts 10 and 11, Institute Report Series of the BAuA]. Bundesanstalt für Arbeitschutz und Arbeitsmedizin, Dortmund, Germany (2000)

[3] Hassenzahl, M.: The effect of perceived hedonic quality on product appealingness. International Journal of Human-Computer Interaction 13, 481–499 (2001)
[4] Kirakowski, J., Corbett, M.: SUMI: The Software Usability Measurement Inventory. British Journal of Educational Technology 24, 210–212 (1993)
[5] Nielsen, J.: Heuristic Evaluation. In: Nielsen, J., Mack, R.L. (eds.) Usability Inspection Methods, pp. 25–62. Wiley, New York (1994)
[6] ISO 9241-10: Ergonomic requirements for office work with visual display terminals (VDTs) - Part 10: Dialogue principles. Beuth, Berlin, Germany (1996)
[7] ISO 9241-11: Ergonomic requirements for office work with visual display terminals (VDTs) - Part 11: Guidance on usability. Beuth, Berlin, Germany (1998)
[8] Preece, J., Rogers, Y., Sharpe, H.: Interaction design: Beyond human-computer interaction. Wiley, New York (2002)
[9] Hassenzahl, M., Burmester, M., Koller, F.: AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. [AttrakDiff: A questionnaire for the measurement of perceived hedonic and pragmatic quality]. In: Ziegler, J., Szwillus, G. (eds.) Mensch & Computer 2003: Interaktion in Bewegung, Teubner, Stuttgart, Germany, pp. 187–196 (2003)
[10] Lindgaard, G., Dudek, C.: What is this evasive beast we call user satisfaction? Interacting with Computers 15, 429–452 (2003)
[11] Holzinger, A., Searle, G., Kleinberger, T., Seffah, A., Javahery, H.: Investigating Usability Metrics for the Design and Development of Applications for the Elderly. In: Miesenberger, K., Klaus, J., Zagler, W., Karshmer, A. (eds.) ICCHP 2008. LNCS, vol. 5105, pp. 98–105. Springer, Heidelberg (2008)
[12] Norman, D.: Emotional Design: Why We Love (or Hate) Everyday Things. Basic Books, New York (2004)
[13] Tractinsky, N.: Aesthetics and Apparent Usability: Empirical Assessing Cultural and Methodological Issues. In: CHI 1997. Electronic Publications (1997), http://www.acm.org/sigchi/chi97/proceedings/paper/nt.htm
[14] Schrepp, M., Held, T., Laugwitz, B.: The influence of hedonic quality on the attractiveness of user interfaces of business management software. Interacting with Computers 18, 1055–1069 (2006)
[15] Nielsen, J.: Jakob Nielsen's Alertbox August 5, 2001: First rule of usability: Don't listen to users (2001), http://www.useit.com/alertbox/20010805.html
[16] Laugwitz, B.: Experimentelle Untersuchung von Regeln der Ästhetik von Farbkombinationen und von Effekten auf den Benutzer bei ihrer Anwendung im Benutzungsoberflächendesign. [Experimental investigation of the aesthetics of colour combinations and of its impact on users when applied to graphical user interface design]. dissertation.de-Verlag im Internet, Berlin (2001)
[17] Komine, K., Sawahata, Y., Uratani, N., Yoshida, Y., Inoue, T.: Evaluation of a prototype remote control for digital broadcasting receivers by using semantic differential method. IEEE Transactions on Consumer Electronics 53(2), 561–568 (2007)
[18] Catell, R.B.: The scree test for the number of factors. Multivariate Behavioural Research 1, 245–276 (1966)

Appendix 1: Original German Items and Their English Translation

Scale	Original German items		English translation	
Attractiveness	unerfreulich	erfreulich	annoying	enjoyable
Perspicuity	unverständlich	verständlich	not understandable	understandable
Novelty	kreativ	phantasielos	creative	dull
Perspicuity	leicht zu lernen	schwer zu lernen	easy to learn	difficult to learn
Stimulation	wertvoll	minderwertig	valuable	inferior
Stimulation	langweilig	spannend	boring	exiting
Stimulation	uninteressant	interessant	not interesting	interesting
Dependability	unberechenbar	voraussagbar	unpredictable	predictable
Efficiency	schnell	langsam	fast	slow
Novelty	originell	konventionell	inventive	conventional
Dependability	behindernd	unterstützend	obstructive	supportive
Attractiveness	gut	schlecht	good	bad
Perspicuity	kompliziert	einfach	complicated	easy
Attractiveness	abstoßend	anziehend	unlikable	pleasing
Novelty	herkömmlich	neuartig	usual	leading edge
Attractiveness	unangenehm	angenehm	unpleasant	pleasant
Dependability	sicher	unsicher	secure	not secure
Stimulation	aktivierend	einschläfernd	motivating	demotivating
Dependability	erwartungskonform	nicht erwartungskonform	meets expectations	does not meet expectations
Efficiency	ineffizient	effizient	inefficient	efficient
Perspicuity	übersichtlich	verwirrend	clear	confusing
Efficiency	unpragmatisch	pragmatisch	impractical	practical
Efficiency	aufgeräumt	überladen	organized	cluttered
Attractiveness	attraktiv	unattraktiv	attractive	unattractive
Attractiveness	sympathisch	unsympathisch	friendly	unfriendly
Novelty	konservativ	innovativ	conservative	innovative

Appendix 2: Loadings of the Final Questionnaire Items on the Extracted 5 Factors

Items	Factors				
	Perspi-cuity	Efficien-cy	Depen-dability	Stimula-tion	Novelty
confusing / clear	.661				
easy to learn / difficult to learn	.856				
complicated / easy	.851				
not understandable / understandable	.857				
usual / leading edge		.849			
dull / creative		.785			
conservative / innovative		.772			
conventional / inventive		.790			
demotivating / motivating			.601		
boring / exiting			.661		
inferior / valuable			.725	.422	
not interesting / interesting			.838		
obstructive / supportive				.505	
does not meet expectations / meets expectations	.438			.549	
unpredictable / predictable				.791	
not secure / secure				.740	
inefficient / efficient					.722
slow / fast					.723
cluttered / organized					.650
impractical / practical				.419	.635

Only loadings > .4 are shown in the table.