Multi-level Validation of the ISOmetrics Questionnaire Based on Qualitative and Quantitative Data Obtained from a Conventional Usability Test

Jan-Paul Leuteritz[1], Harald Widlroither[1], and Michael Klüh[2]

[1] Fraunhofer IAO / Universität Stuttgart IAT, Nobelstr. 12, 70569 Stuttgart, Germany
[2] Hansgrohe AG, Auestr. 5, 77761 Schiltach, Germany
{Jan-paul.leuteritz,Harald.widlroither}@iao.fraunhofer.de,
Michael.klueh@hansgrohe.com

Abstract. Qualitative and quantitative data, collected during a usability evaluation of two innovative prototypes of a small display touch screen device, have been used to perform a multi-level assessment of the questionnaires used within the trial. The use of different validation methods is depicted and discussed concerning their advantages and disadvantages. The conclusions from the validation study are depicted, revealing that the usage of the ISOmetrics for testing uncommon prototypes may result in insufficient validity of the instrument.

Keywords: Validity, questionnaire, ISOmetrics, AttrakDiff, small display devices, shower control.

1 Introduction

1.1 The Goal of the Study

Questionnaires are, at a first glance, a highly attractive instrument for the evaluation of new, innovative prototypes: First, they are easy to use. Moreover, they can be given to a high number of participants without resulting in a lot of extra work for the experimenter. Their use is easy to argue because they are standardised, therefore they are rather objective in comparison to other usability evaluation methods. Last but not least, questionnaires give out a numeric measure of clearly defined dimensions of the users' cognitions or emotions, which makes their results easy to interpret and to explain to clients or colleagues.

However, depending on the exact evaluation task, some usability questionnaires are more adequate than others. Some of them might even become invalid or practically useless if they are used in a certain context. In order to decide which questionnaire to use within their specific projects, usability professionals need empirically based information on the strengths and weaknesses of certain groups of questionnaires or even specific instruments. As this kind of information is usually not provided in the testing manuals, other ways must be encountered to retrieve it.

This article proposes a method to tackle this problem. The idea behind this method is based on the assumption that many usability professionals frequently do evaluations

J.A. Jacko (Ed.): Human-Computer Interaction, Part I, HCII 2009, LNCS 5610, pp. 304–313, 2009.

under repeating conditions; they work, for example, with the same user group or a similar test pattern or they usually evaluate prototypes from a certain line of products. Hence, they could use data from their own tests to cross-validate their survey instruments and see what kind of information they yield. This solution is fine, as long as the cross-validation procedure does not consume too much effort.

In order to find out if such an approach could be recommendable, Fraunhofer Institute of Industrial Engineering (Fraunhofer IAO) conducted the study described in this article. An evaluation project commissioned by the German shower technology manufacturer Hansgrohe AG served as the basis of the multi-level validation approach.

A usability test design was developed that would not just answer the respective evaluation questions but that would also provide data for multi-level validation procedures of the questionnaires used. It was paid attention to keep the additional efforts which only served the validation task as low as possible.

This article presents the outline of the evaluation study and the detailed results of the multi-level validation approach. It aims at inviting other usability professionals to use and/or refine this method.

Fig. 1. Prototype A (© Hansgrohe AG, Design by Phoenix Design, Stuttgart)

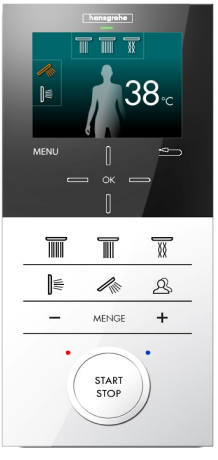

Fig. 2. Prototype B (© Hansgrohe AG, Design by Phoenix Design, Stuttgart)

1.2 The Evaluation Project

The devices to be tested were two prototypes of a wall-mounted device for controlling the different functions of a modern comfort-shower: hand showers, overhead-mounted shower plates offering various combinations of water rays, wall-mounted shower-heads, steam-bath functions, coloured lighting, and a music-player. The designs, including the interaction concept, had been created by Phoenix Design GmbH & Co.KG, Stuttgart.

Prototype A (Fig. 1) was a touch-screen device that featured two additional buttons and a pusher-and-rotator switch. Prototype B (Fig. 2) had a smaller screen that did not respond to touch input. It was instead controlled by a number of buttons, including a

set of four arrow-buttons, an "OK"-button, a "menu"-button, a back-button in form of a u-turn-arrow. Prototype B also featured the pusher-rotator switch.

The usability test was meant to identify the prototype with the better usability, which would then be finalised, while the other prototype would be discarded. Furthermore, the test had to provide information on how to improve the better prototype in the next design development phase.

2 Theoretical Background

2.1 Definition of Usability

The definition of usability on which this validation study is based had been taken from ISO 9241-11. The main advantage of ISO 9241 is that it is an international standard and therefore widely accepted. Furthermore, other definitions of usability (like Nielsen's definition, see Nielsen 1993) seemed to be less adequate for a validation study, as it was suspected that their subordinate constructs might not be independent factors and hence increase the preparatory efforts to be undertaken.

ISO 9241-11 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."

2.2 Measuring Usability

According to ISO 9241, efficacy and efficiency are best measured by so-called objective data, which means behaviour data such as error rates or the time needed to complete a task. This data can be collected during a standardised experiment. The measurement of satisfaction is more difficult, because satisfaction is the user's *subjective reaction* to the interaction with the product (ISO 9241). Hassenzahl (2004) states that user satisfaction is an emotion which results from the user comparing his expectations of the system to his actual experiences with it. Satisfaction is therefore only to be measured by asking the user about his feelings towards the system.

With regards to the above given argumentation, it was assumed that

- The most valid measure or criterion for the efficacy of use would be the number of tasks people were not able to finish by themselves.
- The most valid measure or criterion for the efficiency of use would be either the number of mistakes people made during the trial or the time they needed to complete all tasks.
- The most valid measure or criterion for the users satisfaction with the interface would be either the result of a questionnaire, most probably a semantic differential, or a quantified item on their preference or choice of prototype after the test.

2.3 Selection and Purpose of the Questionnaires

After collecting information about the available psychometric instruments, it was decided to use two questionnaires within the study:

1. The ***ISOmetrics*** (Gediga & Hamborg, 1999), which is supposed to measure usability, using the set of seven dimensions for the design of dialogue systems defined in ISO 9241-10. It's a five-point Likert-scale questionnaire. As the experiment focused on the dialogues of the shower system, the *ISOmetrics* seemed to be adequate. As the *ISOmetrics* is based on the ISO standard, it was expected to fit well into the theoretical approach chosen. According to the above given definitions of criterions, the *ISOmetrics* was in this study not the main source for usability measures but rather an additional instrument, the validity of which was to be examined. It was planned to compare the questionnaire results with the criterions for efficacy and efficiency and with the qualitative data collected during the test.
2. The second questionnaire, the ***AttrakDiff*** (Hassenzahl et al., 2003), is a seven-point-scaled semantic differential questionnaire, which is supposed to measure the *attractiveness* of a system to a user. Although Hassenzahl et al. (2003) do not directly state that the *AttrakDiff* questionnaire measured *satisfaction*, the construct of *attractiveness*, seems to reflect quite well the whole range of expectations a user can have. Hence, this was the instrument selected for the measurement of satisfaction. Validating the *AttrakDiff* in this context was more difficult because there is hardly a better criterion for the users emotions towards a technical system than their responses to an emotion-focused questionnaire. The only other criterion is the subsequent behaviour towards the system after the test – the motivation to carry on with the communication with the system. This is reflected in quantitative preference judgement, which was therefore selected as criterion for the *AttrakDiff*.

3 Method

3.1 Sample

22 users (12 women, 10 men) participated in the study, each providing both quantitative and qualitative data for the validation project. They had an arithmetic mean age of 39.1 years (SD = 14.5 years). The sample consisted of 10 potential customers, 4 elderly users (60+, selected for their lack of experience with information technology) and 8 additional users from Fraunhofer Institute.

3.2 Experimental Setting

The prototypes were simulated on a touch-screen monitor, mounted within the wall of a trade-fair mock-up of a shower cabin. The test was done without having water pour from the showers, the users wore normal clothing. Therefore, a video of the shower's functions was shown at the beginning of the test. Each participant tested both of the prototypes; the sequence was matched according to person characteristics. Each prototype test consisted in a set of tasks the participants had to complete and a questionnaire given after the completion of the task-set. The experiment ended with final questions, asking for a comparison between the tested devices.

It was assured that every participant completed all tasks. Whenever the participant was unable to complete a task by himself, the experimenter provided the information

for the next step and placed a marker in the log-file, indicating that help had been given. If the participant was able to continue by himself after receiving a hint, no further advice was given. Otherwise every assistance was rendered that was needed to complete the task. Participants were instructed to complete each task as fast as possible, without thinking aloud or giving comments. This should guarantee the reliability of the time-measures.

The test was conducted in German language, including instructions and questionnaires. Each test lasted between 90 and 120 minutes. All tests were conducted by the same instructor, using a written instruction. The first trials were supervised.

3.3 Variables Collected

For each participant's interaction with each prototype, the number of tasks was counted that he/she could not complete without the help of the test instructor (*number of hints*). For every task of each participant, the *number of errors*[1] they committed was counted and the *time to complete* the task was measured, using an automatic logging technique.

The questionnaire given to the participants after each of their two trials contained:

1. The *ISOmetric*s in a shortened version. Items that did not apply to shower controls had been deleted. The subscale "suitability for individualization had been removed entirely as none of the items fit. This shortened version is referred to as *ISOmetrics*$_{SDD}$ (*ISOmetrics for small display devices*) in the text below.
2. The *AttrakDiff* in its full version.
3. Additional items, including
   - one item to determine which of the two prototypes the user would prefer in the end and
   - one item that asked to quantify the superiority of the preferred prototype on a five-point-scale.

Qualitative data was taken from the participants' statements and comments during and after each task. All test sessions were videotaped in order to allow a thorough analysis of all the statements the users gave and all their actions, including errors that did not appear in the log-files (e.g., touching the screen of prototype B).

4 The Validation Procedure and Its Results

4.1 Reliability

As the instruments were not new but commonly used ones, there wasn't any attention paid to the factorial structure of the answers. The reliability of the results was calculated rather to exclude a reliability problem that would render all validation attempts useless. Cronbach's α was chosen as a correct factorial structure of the instruments had been assumed.

[1] "Errors" were all intended button-pushes that did not contribute to the solution of the task. Due to the specifications of the log-file, special exceptions were phrased to exclude, for example, unnecessary rotating of the pusher-and-rotator switch from the errors count.

Table 1. Reliability Estimation of the *ISOmetrics$_{SDD}$* subscales, using Cronbach's α

Scale	No. of items	prototype A	prototype B
Suitability for the task	7	.70	.90
Self-descriptiveness	4	.75	.87
Controllability	4	.68	.81
Conformity with user expectations	5	.76	.78
Error tolerance	3	.48	.74
Suitability for learning	4	.79	.90

4.2 Content Validity

A survey with three usability experts from Fraunhofer IAO, conducted before the usability evaluation of the shower prototypes, did not yield any majority vote calling for the deletion or the addition of a specific item or aspect to/from the *ISOmetrics$_{SDD}$*. The lowest mean-estimation of a subscale's validity was 82% (see table 3). Additionally, it has to be stated that the interviewed specialists did not even know the shower control prototypes and hence demanded the inclusion of items that would generally be useful but that had no application in this study.

Table 2. Consolidated ratings of the content validity of the ISOmetrics$_{SDD}$

	No. of evaluators requesting a change	Mean estimation of validity	Number of items to eliminate	Number of aspects missing
Suitability for the task	1	83 %	0	2
Self-descriptiveness	2	82 %	0	1
Controllability	1	90 %	2	2
Conformity with user expectations	0	88 %	0	1
Error tolerance	0	95 %	0	0
Suitability for learning	2	85 %	1	2

4.3 Criterion-Based Validity

Extreme-group-validation

The *ISOmetrics$_{SDD}$* questionnaire was clearly able to identify the "better" prototype, preferred by 20 of 22 participants. Prototype A yielded with 4.12 (SD = 0.50) a significantly higher sum-score than Prototype B with 3.29 (SD = 0.32) ($t(21) = 5.90$, $p = .00$). Hence, using the *ISOmetrics$_{SDD}$* would have led to the correct decision which prototype to discard.

Correlation of problem-counts and subscale-means
Another validation method applied here repeated a procedure that had already been used in a study which reported satisfying validity of the *ISOmetrics* questionnaire (Ollermann, 2004). A category system was created for all the usability problems encountered. Sources were the statements of the participants, the notes of the test instructor and the log-files. For each problem category, the number of occurrences was counted. Then, each problem category (40 for prototype A and 31 for prototype B) was assigned to one $ISOmetrics_{SDD}$ subscale. Afterwards, for each subscale the numbers of appearances of each assigned problem category were summed. This way, the whole sum of all usability problems encountered was split between the questionnaire subscales. Finally, the Pearson-correlation between the number of problems and the mean score of the subscale was calculated for both prototypes.

The application of Ollermann's method yielded less promising results: For prototype A the correlation between the usability problems encountered and the arithmetic mean scores of the $ISOmetrics_{SDD}$ subscales was $r = -0.259$ ($N = 6$; $p = .310$). For prototype B this correlation was $r = 0.020$ ($N = 6$, $p = .485$).[2]

Correlation of $ISOmetrics_{SDD}$ and metric criteria
The Pearson-correlation between the score-differences of the $ISOmetrics_{SDD}$ and the differences in errors committed was statistically not significant with $r = -.11$ ($N = 22$, $p = .66$). The Pearson-correlation between the score-differences of the $ISOmetrics_{SDD}$ and the differences in the time needed to complete all tasks was statistically not significant with $r = -.29$ ($N = 22$, $p = .19$). The Pearson-correlation between the (A-B) difference in *number of hints* (the number of tasks that could only be completed with the instructor's help) and the score-differences of the $ISOmetrics_{SDD}$ was $r = .386$ ($N = 22$, $p = .076$).

Correlation of *AttrakDiff* and the preference item
With a single item it was intended to create a criterion for the validity of the *AttrakDiff* questionnaire. The item requested the participant to describe the degree superiority of the better prototype to the weaker one using a 5-point Likert-Scale. The score was Pearson-correlated to the difference of the *AttrakDiff* sum scores (not-preferred prototype minus preferred prototype). The result was $r = -.44$, statistically significant with $p = .04$ ($N = 22$), which due to the value coding shows that indeed those participants who perceived their favourite to be to a great extent superior to the other prototype also yielded a higher difference in the *AttrakDiff* sum-scores, pointing in the same direction.

5 Discussion of the Results

The results of the survey among usability experts show that there are no severe problems concerning the content of the $ISOmetrics_{SDD}$ items. They apparently represent quite well the ISO definition of the different constructs describing the usability of dialogue systems. However, the correlation between the numbers of problems

[2] N in this case is not the number of participants but the number of subscales used.

assigned to each subscale and the mean scores of the subscales does not support these validity assumptions. Correlations with *N*=6 should not be over-interpreted and significances are not to be expected in any case. However, if one looks at the whole correlation matrix, he will find that the Pearson-correlation between the *ISOmetrics*$_{SDD}$ scores of prototype A and prototype B was $r = 0.416$ ($N = 6$; $p = .206$) and that the problem counts of A and B correlated by $r = 0.627$ ($N = 6$, $p = .091$). This indicates that the data are not totally random. There are coherences between the *ISOmetrics – scores* and between the problems found for both prototypes.

So the question is: Why do neither the sum-scores of the subscales correlate with the problem count, nor do the entire sum-scores correlate with the most objective measures of usability – *user mistakes* and *time-to-complete*?

There is just no match between the usability problems and the questionnaire results. In Ollermann's study, the first correlation coefficient found was $r = 0.277$. As there was one subscale that seemed to be responsible for this low result, this subscale was eliminated, resulting in the correlation jumping up to $r = 0.756$ ($p = .019$) (Ollermann, 2004). This procedure did not seem acceptable in this study because for the two prototypes, different subscales messed up the correlation.

Even more disappointing were the correlations of the *ISOmetrics*$_{SDD}$ scores with *number of errors*, as well as with *time to complete*.

The *AttrakDiff* questionnaire yielded promising results. Regarding the fact that it was validated using just one item, resulting in a possibly low reliability of the criterion, a correlation of $r = -.44$ can be considered sufficiently high to indicate that the results of the questionnaire do more or less reflect the constructs named in the respective theory (see Hassenzahl et al., 2003).

As a consequence of the described findings, it was assumed that the *ISOmetrics*$_{SDD}$ instrument had in this case not been measuring the system's usability. What did it measure instead?

It was presumed that the *ISOmetrics*$_{SDD}$ had failed because it tried to make usability-experts out of the users. Even for the authors of the study, the categorisation of the encountered usability problems to the questionnaire's subscales was a difficult task. Expecting a user to remember all the problems he had encountered and to correctly map them to questionnaire items seems impossible, especially if the user is asked to do so after testing an unknown system for 90 minutes. Most probably, the test participants will rather rely on their general perception of the system, on the emotional substrate of their recent experiences.

Two findings support this presumption:

1. The mean scores of the different subscales were quite similar. For prototype A the standard deviation of those subscale-means is $SD = 0.15$, for prototype B it is $SD = 0.32$, which seems small for a five-point Likert-scale. Ives, Olson and Baroudi (1983, as cited in Hartson et al., 2000) say that participants tend to fill in satisfaction questionnaires quite homogeneously. This might also apply to questionnaires like the *ISOmetrics*$_{SDD}$.
2. The correlation between the differences of the *AttrakDiff*-scores (A-B) and the differences of the *ISOmetrics*$_{SDD}$-scores (A-B) of the participants was $r = 0.81$ ($N = 22$; $p = .00$). This means that the *ISOmetrics*$_{SDD}$ perfectly measured the emotional value that the participants gave to the system, closely linked to what is called "satisfaction".

6 Conclusions

6.1 Concerning the Findings of the Study

When confronted with a system for the first time, users are probably unable to remember the usability problems they encountered and to cluster them correctly, producing a valid score in all the subscales of a questionnaire like the *ISOmetrics*. Participants rather seem to use the instrument to convey their overall satisfaction with the system to the test instructor. Therefore the use of questionnaires focusing on different categories of usability problems is not recommendable in certain test designs. According to the findings of this study, questionnaires like *SUMI*, *QUIS* and *ISOmetrics* need to be used carefully.

6.2 Concerning Multi-level Validations

The aim of this article and the work depicted here is to encourage usability experts to evaluate their measurement instruments with a method similar to this multi-level-approach. This approach of course has a downside, which is the small number of participants. In the above described case, only three usability experts have been interviewed, only two prototypes were used, only 22 participants have gone through the evaluation process, and only six subscales of the *ISOmetrics* were taken into account. Furthermore, aspects like the assignment of the encountered usability problems to certain scales could always be questioned. Finally, it could be argued that the changes done to the questionnaire (e.g. the deletion of items) had a bad effect on the validity of the whole instrument. The results of such a study may hence seem less apt for publication than the results of big validation studies, carried out with hundreds of participants.

The advantage of this method is that without going into unreasonable costs of money or time, it combines different forms of validations and collects information that is usually just lost.

Eventually, the question is if a usability practitioner's primary interest is to win a scientific argument and publishing results or if s/he just wants to get a hint on whether a certain tool is recommendable for the planned task or not. In the second case, the common perception of usability evaluation itself would then also apply to the evaluation of the assessment tools: Little and possibly unreliable information is better than none (see Nielsen, 1993). So if the assumption is true that not just literally the validity of a questionnaire but more generally the value gained from its results is possibly dependent on the product tested, the users, and other context parameters, then the here promoted method becomes recommendable.

References

1. ISO 9241, Ergonomics of human-system interaction. International Organization for Standardisation (1998)
2. Gediga, G., Hamborg, K.-C.: IsoMetrics: Ein Verfahren zur Evaluation von Software nach ISO 9241/10. In: Hollingm, H., Gediga, G. (eds.) Evaluationsforschung, pp. 195–234. Hogrefe, Göttingen (1999)

3. Hamborg, K.-C.: Gestaltungsunterstützende Evaluation von Software: Zur Effektivität und Effizienz des IsoMetrics^{L} Verfahrens. In: Herczeg, M., Prinz, W., Oberquelle, H. (eds.) Mensch & Computer 2002, pp. S.303–S.312. B.G. Teubner, Stuttgart (2002)
4. Hartson, H.R., Andre, T.S., Williges, R.C.: Criteria for Evaluating Usability Evaluation Methods. International Journal of Human-Computer Interaction, 2001 13(4), 343–349 (2000)
5. Hassenzahl, M., Burmester, M., Koller, F.: AttrakDiff.: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. In: Ziegler, J., Szwillus, G. (eds.) Mensch & Computer 2003, pp. 187–196. B.G. Teubner, Stuttgart (2003)
6. Hassenzahl, M.: Interaktive Produkte wahrnehmen, erleben, bewerten und gestalten. In: Thissen, F., Stephan, P.F. (eds.) Knowledge Media Design – Grundlagen und Perspektiven einer neuen Gestaltungsdisziplin, Oldenburg Verlag, München (2004)
7. Nielsen, J.: Usability Engineering. Morgan Kaufmann, Heidelberg (1993)
8. Ollermann, F.: Verhaltensbasierte Validierung von Usability-Fragebögen. In: Keil-Slawik, R., Selke, H., Szwillus, G. (eds.) Mensch & Computer 2004: Allgemeine Interaktion, pp. 55–64. Oldenburg Verlag, München (2004)