Katharina Sachse & Manfred Thüring
Technische Universität Berlin
Institute of Psychology and Ergonomics
Franklinstr. 28/29, FR 2-6
Berlin, 10587, Germany
Contact: Katharina.Sachse[at]tu-berlin.de

A Usability Evaluation Tool for Safety-Relevant Systems

TU Berlin, January 2012

Abstract

Software for safety-relevant areas of application (e.g. power plants, traffic control) has to satisfy particular requirements. To date, no instruments are available which measure usability with special respect to safety. We develop an approach to assess aspects of usability by considering safety-related priorities. Usability is measured with standardized user questionnaires (e.g. ISONORM 9241/110-S). By including a weighing procedure (AHP), aspects of special importance can be identified and taken into account for the usability evaluation. For this purpose, ratings for usability aspects are combined with importance assessments. As result, a usability score is calculated that can be used for benchmarking. The questionnaires and the weighing procedure will be provided as a computer application, which can be used for the direct collection, computation, and presentation of data.

Introduction

In more and more domains, software is used to support the monitoring and control of safety-relevant or safety-critical processes. In such domains, tasks must be accomplished which might entail severe negative consequences when performance is deficient. Some examples are flight control systems, control rooms of power plants, and electronic medical devices. While substantial measures are taken to ensure that these systems satisfy high safety standards, much less is done to warrant their usability. This is a severe shortcoming, since software-ergonomic deficits can result in serious problems or even accidents [6]. In safety critical domains, poor usability is no mere inconvenience, but a safety risk.

However, safety requirements might be in conflict with usability requirements. For instance, when inputs must be affirmed twice, error probability is reduced but efficiency is decreased as well. Methods for evaluating the usability of safety relevant systems must take such potential conflicts into account. We address this issue by developing a usability evaluation tool that measures usability aspects with special respect to safety requirements by applying a weighing procedure that allows prioritization.

Measuring usability of safety-relevant systems

The international standard IEC 61508-2: Functional safety of electrical/electronic/programmable electronic safety-related systems demands that the design of these systems “shall take into account human capabilities and limitations and be suitable for the actions assigned to operators and maintenance staff”. Further it is required that “the design of all interfaces shall follow good human-factor practice” [2]. To ensure the fulfillment of these requirements, usability evaluation is necessary. However, systems are often implemented without being adequately tested for usability. This has several reasons: Users of safety relevant systems are specialized operators who have little time for off-site

tests during the design process. Safety relevant and critical systems consist of various software and hardware components and testing a single component in the usability lab might neglect the working situation. For these reasons, lean methods implemented as computer-based tools are needed to measure usability of safety-critical systems efficiently in the field.

Usability Questionnaires

An efficient way to test the usability of human-machine-interfaces is the use of standardized questionnaires that allow the assessment of various evaluation aspects without requiring lab tests. Systems can be evaluated in their natural working context. Analysis of results is relatively easy and studies using the same questionnaire to assess different systems can be compared. Hence the questionnaire method seems particularly appropriate to verify the usability of safety-relevant systems.

Over recent years, a number of standardized user questionnaires have emerged (e.g., SUS [1], SUMI [5], Isonorm 9241-110/S [7]). Users rate their experiences with a system on several items which address different aspects of usability. The ratings are combined into a quantitative usability score. Some questionnaires also allow the computation of sub-scores which provide more detailed information. For instance, the German questionnaire Isonorm 9241-110/S addresses the seven dialogue principles defined in the ISO 9241-110: Ergonomics of human-system interaction - Part 110: Dialogue principles [3] (see table 1). Each principle is considered as a separate dimension and measured by three items. Table 2 gives an example of the items building the dimension error tolerance.

suitability for the task (TS)
suitability for learning (LS)
suitability for individualization (IS)
conformity with user expectations (EC)
self descriptiveness (SD)
controllability (C)
error tolerance (ET)

Table 1: Dialogue principles defined in ISO 9241-110 and measured with the questionnaire Isonorm 9241-110/S

The software…		
provides error messages which are difficult to understand.	○ ○ ○ ○ ○ ○ ○	provides error messages which are easy to understand.
error correction generally requires a lot of effort.	○ ○ ○ ○ ○ ○ ○	error correction generally requires little effort.
does not give concrete help for error correction.	○ ○ ○ ○ ○ ○ ○	gives concrete help for error correction.

Table 2: Items of the Isonorm 9241-110/S measuring ET [7]

Prioritizing Usability Aspects

As mentioned before, the compliance with usability requirements can be in conflict with the fulfillment of safety requirements. To give an example: Safety relevant systems might not provide the possibility to individualize the look of the user interface to assure that different operators use identical working environments. To consider such constrictions in the usability evaluation, a prioritization of aspects is necessary. Therefore, we combine the usability assessment with a weighing procedure to identify those usability aspects which are especially important for safety-relevant systems.

To prioritize aspects, it is inappropriate to ask for absolute importance (e.g. “How important is controllability?” not important – somewhat important – quite important – very important). The problem is that users might not differentiate between the various aspects and perceive each as very important. Therefore a weighing procedure is suggested that produces relative importance values by forcing users to weigh each aspect against all others.

One method to assess relative weights is the Analytic Hierarchy Process (AHP) [8]. Aspects are compared pair-wise and their relative weight is assessed by indicating on a rating scale ranging from 1 to 9 how much more important one aspect is in comparison to the other. If two aspects are of equal importance both receive the value 1. If one aspect is extremely more important than the other, it receives the value 9, while the reciprocal of this value is assigned to the other aspect. Values in between can be used to grade the importance assessments. The weights are computed by putting the importance ratings in a pair-wise comparison matrix, then summing each row and dividing each cell by the sum of their row. These normalized cell scores are summed up line by line and the sum is divided by the number of rows which results in the weights. All weights sum up to 1.

To apply AHP for the prioritization of usability aspects, the weighing procedure can be used to compare items of a usability questionnaire [9]. An example shall illustrate the calculation procedure. The priorities of aspects of error tolerance assessed by AHP are shown in table 3. In this example easy to understand error messages are the most important aspect followed by little correction effort. That the system gives concrete help is regarded as least important.

	Error message	Correction effort	Concrete help	Weight
Error message	1 (0,55)	2 (0,63)	3 (0,33)	0,50
Correction effort	$^1/_2$ (0,27)	1 (0,31)	5 (0,56)	0,38
Concrete help	$^1/_3$ (0,18)	$^1/_5$ (0,06)	1 (0,11)	0,12
Σ	1,83	3,2	9	1,0
Numbers in brackets are normalized cell scores (cell score divided by sum of their row)				

Table 3: Example for the calculation of weights using AHP

The resulting weights can be combined with usability ratings to receive a weighed usability score [9].

Usability Scoring Procedure

We illustrate the usability scoring procedure by using the questionnaire Isonorm 9241-110/S as an example. First, the user assesses the usability of a system by filling in the questionnaire, i.e. rating the several usability aspects on the items. This results in 21 item ratings R_i.

In a second step, the user has to prioritize the usability aspects by weighing the items. Two items of one dimension (e.g. error tolerance) are presented in parallel. The user has to indicate how much more important one aspect is compared to the other (e.g. easy to understand error messages versus little effort for error correction). To weigh all items within their respective dimensions, 21 pair-wise comparisons are necessary (3 comparisons of items for 7 dimensions). The item weights W_i are calculated as described above. To evaluate the overall importance of a single usability aspect the importance of the dimension to which this aspect belongs has to be considered. Therefore, the relative importance of the seven dimensions is also assessed by AHP which requires 21 pair-wise comparisons and results in 7 weights of dimensions W_d. To compute total weights of items the item weights W_i are multiplied with their respective weights of dimensions W_d, resulting in total weights of items Wt_i. To illustrate this calculation, we use the item weights W_i of the example above (see table 3) and set the weight of dimension W_d for error tolerance to 0.2. The resulting total weights of items Wt_i for

the aspects of error tolerance are 0.1 for understandable error messages, 0.076 for correction effort, and 0.024 for concrete help messages.

These total weights of items Wt_i are multiplied with their respective item ratings R_i. The resulting weighed item ratings RW_i are summed up within dimensions to dimension specific usability scores US_d. Table 4 shows the calculation using the example introduced above.

	Total weights of items Wt_i	Ratings R_i	Weighed ratings RW_i
Error message	0,100	2	0,200
Correction effort	0,076	3	0,228
Concrete help	0,024	6	0,144
		US_d	0,572

Table 4: Example for the calculation of weighed ratings RW_i and the dimension specific usability score US_d.

Finally, the sum of the seven dimension specific usability scores builds the total usability score US. Please see the appendix for an overview of the complete Usability Scoring Procedure.

The usability score US does not only reflect the compliance with usability standards but addresses usability with respect to particular requirements of the task by allowing for prioritization. Safety requirements can be considered by assigning high weights to those usability aspects that are relevant for safety (e.g. error tolerance). Therefore this procedure seems particularly useful for usability evaluations of systems that have to fulfill special safety standards which might restrict usability. The usability score can be used to compare several systems or to evaluate a single system against a usability standard.

ReMUS - A software tool for the usability scoring procedure

To evaluate software with the usability scoring procedure is rather complex. In particular, the AHP weighing procedure is quite demanding. Therefore, we develop a computer-based tool that combines usability questionnaires with weighing procedures and automatically calculates usability scores. The tool is called *ReMUS*, which is an acronym of the German term for Computer-Based Multiattributive Usability Scoring (***Re****chnerbasiertes* ***M****ultiattributives* ***U****sability* ***S****coring*).

In ReMUS, various standardized usability questionnaires are provided as computer-based versions (e.g. Isonorm 9241-110/S, SUS). The usability evaluator can select a questionnaire and combine it with a weighing procedure of her choice (e.g. AHP). All instruments are available in German and in English versions. Additionally, texts, pictures or self-created questions can be included to customize the survey. The tool automatically generates a ready-to-use computer-based survey that can be presented to users who are requested to assess a system. The survey can be executed online or offline. Results of the usability scoring procedure are directly calculated after a user has filled in the survey. The data are saved as excel-file which can be used for further analyses.

Currently, an executable prototype of ReMUS is under review. The tool will be applied to evaluate the usability of safety-relevant software in nuclear power plants. In doing so, we will address questions regarding two different user groups:

- Software evaluators: Are they able to configure surveys and interpret the results, i.e. is the usability evaluation tool usable?
- Participants of evaluations (i.e. operators of safety-relevant systems): Do they accept the complex scoring procedure? Are they able to fill out the survey with a reasonable effort?

Results will be used to improve the usability of ReMUS and to optimize the scoring procedure. The efficiency of the procedure can be increased by implementing standard weights for particular systems or applications. In these cases, the weighing procedure would be no longer necessary. However, the definition of standardized weights requires extensive analyses with datasets of many users and systems to test if weights are stable over time.

Further, it is planned to extend ReMUS by integrating domain-specific expert checklists to address specific requirements. These checklists will be based on norms and standards, for instance on the ISO 11064: Ergonomic design of control centres [4]. So it is possible to additionally address particular safety requirements of the human-machine-interface for the domain in question

Acknowledgements

The project is funded by the German Federal Ministry of Economics and Technology.

Citations

[1] Brooke, J. SUS: a "quick and dirty" usability scale. In Jordan, P. W., Thomas, B., Weerdmeester, B. A. & McClelland, A. L. (Eds.): *Usability Evaluation in Industry*. Taylor and Francis, London, GB, 1996.

[2] International Electrotechnical Commission. IEC 61508-2: Functional safety of electrical/ electronic/ programmable electronic safety-related systems. 2000.

[3] International Organization for Standardization. *ISO 9241-110: Ergonomics of human-system interaction -Part 110: Dialogue principles*. 2006.

[4] International Organization for Standardization. *ISO 11064: Ergonomic design of control centres*. 2008

[5] Kirakowski, J. *The Use of Questionnaire Methods for Usability Assessment*. http://www.ucc.ie/hfrg/questionnaires/sumi/sumipapp.html

[6] Leveson, N.G. *Safeware, system safety and computers*. Addison-Wesley, Amsterdam, Netherlands, 1995.

[7] Prümper, J. *ISONORM 9241/110-S: Evaluation of software based upon International Standard ISO 9241, Part 110.* http://www.f3.htw-berlin.de/Professoren/Pruemper/instrumente/ISONORM_9241_110-S_2010.pdf

[8] Saaty, L. How to make a decision: The Analytic Hierarchy Process. *European Journal of Operational Research* (1990), 9-26.

[9] Sachse, K. & Thüring, M. (2011). Usability sicherheitskritischer Software. In M. Eibl (Ed.) *Mensch und Computer 2011* (339-342). München: Oldenbourg Wissenschaftsverlag.

Appendix

The Usability Scoring Procedure

Examples

Results

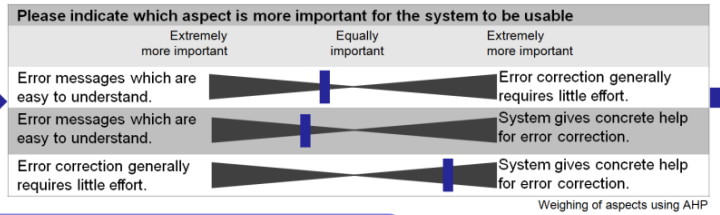
1 The user rates the usability of a system by filling out a standardized usability questionnaire.

Please rate: The software…		
provides error messages which are difficult to understand.	○X○○○○○	provides error messages which are easy to understand.
error correction generally requires a lot of effort.	○○X○○○○	error correction generally requires little effort.
does not give concrete help for error correction.	○○○○○X○	gives concrete help for error correction.

Items of the Isonorm 9241-110/S measuring error tolerance

→ Usability-Ratings R_i

2 The user prioritizes the usability aspects addressed in the questionnaire by weighing the items. Two aspects are presented in parallel and the user has to indicate how much more important one aspect is in comparison to the other.



Weighing of aspects using AHP

→ Item-specific Priorities P_i

In case of a multi-dimensional questionnaire, the user has to assess the importance of the dimensions using the same weighing procedure.

→ Dimension-specific Priorities P_d

3 Transformation of priorities into weights using the AHP-procedure.

The weights are computed by putting the priority ratings in a pair-wise comparison matrix, then summing each row and dividing each cell by the sum of their row. These normalized cell scores are summed up line by line and the sum is divided by the number of rows which results in the weights.

	Error message	Correction effort	Concrete help	Weight
Error message	1 (0,55)	2 (0,63)	3 (0,33)	0,50
Correction effort	1/2 (0,27)	1 (0,31)	5 (0,56)	0,38
Concrete help	1/3 (0,18)	1/5 (0,06)	1 (0,11)	0,12
Σ	1,83	3,2	9	1,0

Numbers in brackets are normalized cell scores.

Example for the computation of weights in AHP

→ Item-specific Weights W_i

→ Dimension-specific Weights W_d

In case of a multi-dimensional questionnaire, multiplication of item-specific weights W_i with dimension-specific weights W_d. For one-dimensional questionnaires $Wt = W_i$.

→ Total Weights of items Wt_i

4 Multiplication of total weights of items Wt_i with item ratings R_i.

	Rating R_i	Total weight of item Wt_i	Weighed Rating RW_i
Error message	2	0,100	0,200
Correction effort	3	0,076	0,228
Concrete help	6	0,024	0,144

Example for the computation of weighed ratings. The dimension-specific weight was set to 0,2

→ Weighed Ratings RW_i

In case of a multi-dimensional questionnaire, the resulting weighed ratings RW_i are summed up within dimensions to dimension specific usability scores.

→ Dimension-specific Usability Scores US_d

5 Finally, building the sum of all weighed ratings Rw_i (or all dimension specific usability scores US_d).

The resulting usability score does not only reflect the compliance with usability standards but addresses usability with respect to particular requirements of the task by allowing for prioritization. Therefore this procedure seems particularly useful for usability evaluations of systems that have to fulfill special safety standards which might restrict usability. The usability score can be used to compare several systems or to evaluate a single system against a usability standard.

→ Usability Score US