Comparing Heuristic Walkthrough and User Studies in Evaluating Digital Appliances

Eva-Maria Meier, Patricia Böhm, Christian Wolff

Media Informatics Group, University of Regensburg, Germany
maria1.meier@stud.uni-regensburg.de, {patricia.boehm,christian.wolff}@ur.de

Abstract

In this paper we present an empirical study comparing user studies and expert evaluations based on a specific set of heuristics for evaluating information appliances with a heuristic walkthrough. The study looks at an e-book reader as well as a digital music player. In the user study, question-answer protocols are used as means of intervention during the experiments. To gain insight into performance of the evaluation methods the identified problem sets were analyzed. Results for the thoroughness, validity and effectiveness are presented and compared with prior studies.

Keywords: intergenerational; knowledge sharing; generation; organizations; information and communication technology; tacit knowledge

1 Introduction

In the context of the internet of things (IoT), interactive technology is set to become more diverse than ever: People will interact on different platforms, using different devices and non-standard interface design. The promise of the invisible computer (Norman, 1998), rather than indicating less interaction, actually points towards ubiquitous media interaction. In this context, adequate methods for evaluating interactive systems are needed. In this paper,

In: M. Gäde/V. Trkulja/V. Petras (Eds.): Everything Changes, Everything Stays the Same? Understanding Information Spaces. Proceedings of the 15th International Symposium of Information Science (ISI 2017), Berlin, 13th—15th March 2017. Glückstadt: Verlag Werner Hülsbusch, pp. 146—157.

we present a comparative study of different approaches to evaluation information appliances: Starting from earlier research on heuristics of information appliances (Böhm, Schneidermeier & Wolff, 2014), two devices are evaluated using a heuristic walkthrough as well as user studies. The effectiveness of both methods is compared using thoroughness and validity as major criteria. In this study, we want to examine

- how good heuristic walkthroughs can be adopted using our set of heuristics for information appliances,
- gain information on thoroughness and validity of findings for the heuristic walkthrough using user studies as the methodological point of reference (comparison of evaluation methods) and
- compare the performance of the two methods for different appliances being examined.

The rest of this paper is organized as follows: In chapter 2, we give a short overview of the state of the art. In chapter 3, heuristics for evaluating information appliances are introduced. In chapter 4 we present metrics for comparing the output of different usability evaluation methods (UEMs). Chapter 5 gives an overview of the design of our study. Results are presented and discussed in chapter 6, and chapter 7 draws conclusions and gives a short outlook.

2 State of the art

In the last decades, starting from early human factors research and later becoming a major field of inquiry within computer and information science, human-computer interaction has developed a broad variety of methods for evaluation usability, defined in ISO DIN EN 9241-11:1998 (1998) as

> "[t]he 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."

The heterogeneity of evaluation methods is illustrated in handbooks such as Martin & Hannington (2012). While many methods have been well-established for a long time and can be organized along major oppositions like expert/user study, formative/summative evaluation, quantitative vs. qualitative

results, less is known about the relative effectiveness and efficiency of the methods themselves (cf. Law & Hvannberg, 2004).

For gaining insight into the performance characteristics of a usability evaluation method (UEM), Hartson et al. (2001) suggest using user studies as the reference point for comparison: Using real users for an application for device can generate a set of actual interaction problems against which other evaluation methods – in our case: heuristic walkthroughs based on our set of information appliance evaluation heuristics – can be matched.

3 Heuristics for evaluating information appliances

Picking up well-known sets of heuristics as put forward by Molich & Nielsen (1990) or Shneiderman ("eight golden rules", Shneiderman, 1987) as well as relevant standards (ISO 9241-110:1998(en), 1998) and literature on heuristic evaluation, we have developed a set of heuristics for the evaluation of information appliances (Böhm, Schneidermeier & Wolff, 2014). This set consists of eight heuristic principles at the top layer, with additional sub-heuristics defined for each main category. The eight heuristics are defined as follows (translated from German):

1. *Consistency*: The appliance is designed consistently and conforms to applicable standards.
2. *Feedback*: Each interaction step should have immediate, appropriate, and recognizable feedback.
3. *Easy handling*: Handling should be as efficient as possible but at the same time give the user a feeling of being in charge.
4. *Error avoidance*: The design should take precaution that interaction errors concerning not-supported interaction, inadvertent interaction, or mix-up of function in interaction do not occur.
5. *Suitability for the task:* The device should provide the functions expected and needed by the user; the user interface should be designed to fit the tasks.
6. *Help and documentation:* In case of interaction problems and for helping to learn new functions, the device should provide adequate information.

7. *Self-descriptiveness:* Interaction for basic functions should be understandable without instructions or handbook usage.
8. *Flexibility*: Users with different competencies cans use the device under different circumstances in everyday situations.

4 Problem comparison

The comparison of different UEM should be based on *common problem descriptions* (Lavery et al., 1997: 257 f.), consisting of a description of *context*, *cause*, *outcomes*, *breakdown in User's interaction*, *outcomes of the breakdown, outcome, solution*. This template was used for describing all problems found in this study. Problem descriptions from different evaluation methods can be compared and mapped if they are *consistently described in a common format* as suggested by Lavery et al. (1997).

In our study, we want to know how many actual or existing user problems can be found with a heuristic method. Thus, we need an initial set of "existing problems". This can be generated by different methods (cf. Hartson et al. 2001: 390), we decided to generate a reference problem set by performing a user study. Given this set, the following metrics can be calculated (Sears et al., 2001: 388; Sears, 1997):

$$Thoroughness = \frac{number\ of\ real\ problems\ found}{number\ of\ real\ problems\ that\ exist}$$

$$Validity = \frac{number\ of\ real\ problems\ found}{number\ of\ issues\ identified\ as\ problems}$$

$$Effectiveness = Thoroughness \times Validity$$

For expert-based heuristic evaluations, the problem of false positives is well-known: Expert may "discover" usability problems that do not occur in users' actual interaction with a system or appliance. Therefore, *effectiveness* of a method means setting *thoroughness* in relation with *validity* as a rate for false positives (Sears, 1997: 213): The more problems described by experts turn out to be no "real" problems, the smaller the value for validity will be.

5 Study design

In Böhm, Schneidermeier and Wolff (2014) the following selection criteria for (information) appliances are defined: Small display, hard key as well as soft key interaction controls, and mobile context of usage.

5.1 Device selection

In our study, we have selected two devices which fulfil these elementary criteria (fig. 1):

- *Kobo Glo*
- *Apple iPod Nano*

Fig. 1 Kobo Glo e-book reader (left), *Apple iPod Nano* music player (right)

While *Kobo Glo* is a digital e-book reader, the Apple *iPod Nano* serves as a (MP3) music player. Selection criteria for the devices were as follows:

- In our previous study, we had evaluated a digital camera as well as a copying machine; in this study other/additional device types should be studied.
- At the same time, we intended to have devices with different assumed design / UX quality: Going along with the well-established UX design often found in Apple device we assume that the *iPod Nano* would have high usability ratings. At the same time, the opposite might be observed for the *Kobo Glo*, produced by a little known manufacturer with less experience in UX matters.

5.2 Task design

For developing adequate tasks for the user study that was performed to generate the reference set of "real" problems, a preliminary survey was conducted among possible users of both device types, e-book readers (20 questionnaires) as well as digital music players (27 questionnaires). For both device types, core tasks (e.g. reading; page navigation; searching / listening to music; shuffle function) as well as supporting tasks (setting markers; adjusting type size / video watching, surfing the internet) were identified. From this collection the actual tasks for the evaluation were generated.

6 Evaluation

The evaluation comprised two parts, the user study as well as the heuristic evaluation by experts. Both are briefly described below; we will not go into the details of particular interactions problems found in the study (cf. Meier, 2015, for an in-depth discussion of identified problems) as we want to focus on the aspect of method comparison here.

6.1 User study

In the user study, 20 test persons were presented with the tasks for the two devices. Following Nielsen, 2000, we assume that with 20 test persons, the actual amount of "real" problems can be approximated quite well. In the user study, test persons were recorded using a webcam. A moderator was present for conducting the experiment. In addition to the *thinking aloud* method, a *question-answer-protocol* was used for documenting problem situations (Grossman et al., 2009). Pre- as well as post-task questionnaires were used to document demographics and users' experience with the device types as well as post-test ratings of the devices and their respective functionality and usability. All usability problems were documented using the template suggested by Lavery et al. (1997) along with a severity rating of the usability problem.

6.2 Heuristic evaluation by experts

Two experts were selected for the heuristics walkthrough based on the heuristics as introduced in chapter 3 above. Both experts have a background in usability engineering (information science / media informatics), with one expert having a junior level of experience while the other was already at a senior level (5+ years of UX experience). Using the test scenarios and heuristics as described above, both experts performed heuristic walkthroughs and documented their results. An explicit guideline for the heuristic evaluation was used in order to make sure that both experts followed a similar process in the heuristic walkthrough. A camera was used for documenting the usability problems found. For the problem documentation, the same template as in the user study was used.

7 Results and interpretation

The participating test persons in the user study consisted of 2 pupils, 7 undergraduate as well as graduate students, and 11 adult employed persons (age range 17–49 years). 20% of them possess an e-book reader, 65% a music player (digital media device). The following figures 2 and 3 show the amount of problems identified per user and device:

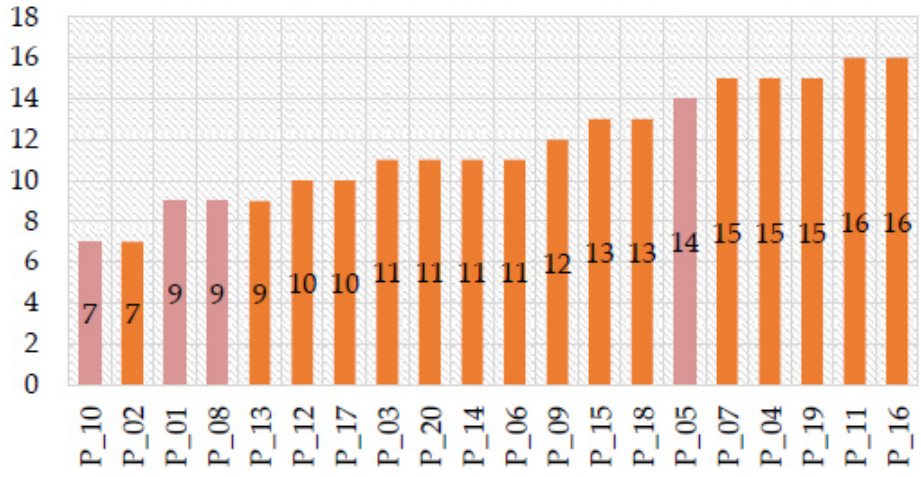

Fig. 2 Number of identified problems per User – *Kobo Glo* e-book reader

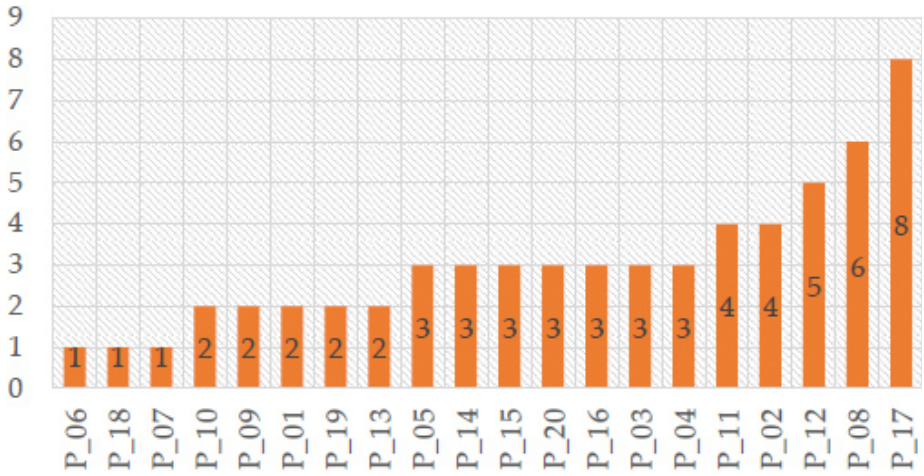

Fig. 3 Number of identified problems per User – *Apple iPod Nano* music player

In the post-task questionnaire, only few users showed willingness to buy the e-book reader (15%), while 70% would be willing to do so for the music player. The e-book reader was rated with 37.25 on average on the system usability scale (SUS) (Brooke, 1996, 2013), a fairly low value, while the music player reached a quite impressive value of 84.75 on average. The initial assumption of a broad difference in design and UX quality could be confirmed.

In the heuristic walkthrough, the experts identified the following number or problems (cf. tables 1 and 2, items with no problems found left out):

Table 1:
Number of identified problems per Expert – Kobo Glo e-book reader

Criterion	# problems for expert 1	# problems for expert 2
Consistency	4	13
Feedback	1	2
Easy handling	7	3
Error avoidance	8	13
Suitability for the task	3	2
Sum	23	38

Table 2:
Number of identified problems per Expert – Apple iPod Nano music player

Criterion	# problems for expert 1	# problems for expert 2
Consistency	0	10
Feedback	0	0
Easy handling	1	0
Error avoidance	3	2
Suitability for the task	1	0
Help and documentation	1	2
Self descriptiveness	1	0
Sum	7	14

In a next step, the problems identified by both experts were mapped onto each other using the common problem description format (Meier, 2015: 107 ff.). After this identification of overlaps, the overall expert problem set was constructed. Next, the problems sets for both UEMs were compared:

- *Kobo Glo*: 30 "real" user problems out of 38 overall problems could be mapped to 30 expert problems (out of 43 problems).
- *Apple iPod Nano*: 8 "real" user problems out of 13 overall problems could be mapped to 8 expert problems (out of 17 problems). For details of the problems and their mapping as well as a discussion of all problems identified by one type of UEM only, see Meier (2015: 113 ff.).

Finally, we have calculated the quantitative metrics as introduced in chapter 4 above:

$$Thoroughness_{Kobo\ Glo} = 30 \div 38 = 0.79$$

$$Thoroughness_{Apple\ iPod\ Nano} = 8 \div 13 = 0.62$$

$$Validity_{Kobo\ Glo} = 30 \div 43 = 0.70$$

$$Validity_{Apple\ iPod\ Nano} = 8 \div 17 = 0.47$$

Finally, effectiveness was calculated as the product of thoroughness and validity:

$$Effectiveness_{Kobo\ Glo} = Thoroughness_{Kobo\ Glo} \times Validity_{Kobo\ Glo} = 0.79 \times 0.70 = 0.55$$

$$Effectiveness_{Apple\ iPod\ Nano} = Thoroughness_{Apple\ iPod\ Nano} \times Validity_{Apple\ iPod\ Nano} = 0.62 \times 0.47 = 0.29$$

8 Discussion and outlook

Table 3 shows a comparison of these results with results from a previous study (Böhm, Schneidermeier & Wolff, 2014):

Table 3: Comparison of results with a prior study

	Camera	Copying Machine	*Kobo Glo* e-book reader	*Apple iPod Nano*
Thoroughness	0.77	0.79	0.79	0.62
Validity	0.91	0.79	0.70	0.47
Effectiveness	0.70	0.62	0.55	0.29

It becomes clear that results for the music player a much worse than for all other three devices. One might assume that the higher design quality – or the assumed higher design quality *as perceived by the experts* – plays a role in this outcome. The differences are more or less the same for both experts in the study.

Regarding the method, the combination of information appliance heuristics and heuristic walkthrough using a guideline worked quite well as the heuristics are more precise (only the top level is presented in chapter 3 above) than the more general heuristics discussed in the literature.

For the user study, using a question-answer-protocol as a means of documenting interventions by the moderator was successful in the sense that more tasks could be completed by the users. Finally, using precise templates for problem description had a steep learning curve in the beginning, but proved to be very helpful for problem identification and mapping.

The effects of (perceived) good interaction design quality for expert-based UEMs should be studied in more detail in the future.

References

Böhm, P., T. Schneidermeier, and C. Wolff (2014): Heuristiken für Information Appliances. In: A. Butz, M. Koch, & J. Schlichter (Eds.): *Mensch & Computer 2014 – Tagungsband*. Berlin: De Gruyter Oldenbourg (pp. 275–284).

Brooke, J. (1996): SUS: A “quick and dirty” usability scale. In: P. W. Jordan, B. Thomas, B. A. Weerdmeester, & A. L. McClelland (Eds.): *Usability Evaluation in Industry*. London: Taylor and Francis (pp. 4–7).

Brooke, J. (2013): SUS: a retrospective. In: *J. Usability Studies*, 8 (2), 29–40.

Grossman, T., G. Fitzmaurice, and Attar, R. (2009): A survey of software learnability: metrics, methodologies and guidelines. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM (pp. 649–658).

Hartson, H. R., T. S. Andre, and Williges, R. C. (2001): Criteria for evaluating usability evaluation methods. In: I*nternational journal of human-computer interaction*, 13 (4), 373–410.

ISO 9241-11:1998(en) (1998): Ergonomic requirements for office work with visual display terminals (VDTs) – Part 11: Guidance on usability. https://www.iso.org/obp/ui/#iso:std:iso:9241:-11:ed-1:v1:en

Law, E. L. C., and E. T. Hvannberg (2004): Analysis of strategies for improving and estimating the effectiveness of heuristic evaluation. In: *Proceedings of the third Nordic conference on Human-computer interaction*. ACM (pp. 241–250).

Lavery, D., G. Cockton, and M. P. Atkinson (1997): Comparison of evaluation methods using structured usability problem reports. In: Behaviour *& Information Technology*, 16 (4–5), 246–266.

Lazar, J., J. H. Feng and H. Hochheiser (2010): *Research methods in human-computer interaction*. Chichester: Wiley.

Martin, B., and B. M. Hanington (2012): *Universal methods of design: 100 ways to research complex problems, develop innovative ideas, and design effective solutions*. Beverly, MA: Rockport Publishers.

Meier, Eva-Maria (2015): *Heuristische Evaluation versus Nutzerstudie – Vergleich der beiden Methoden für die Evaluation von Geräten* [Heuristic Evaluation vs User Studies – A Comparison of Both Methods for Evaluating Information Appliances]. B. A. Thesis, Media Informatics Group, University of Regensburg, September 2015.

Nielsen, J. (2000): Why You Only Need to Test with 5 Users. http://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/.

Nielsen, J., and R. Molich (1990): Heuristic evaluation of user interfaces. In: *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM (pp. 249–256).

Norman, D. A. (1998): *The invisible computer: Why good products can fail, the personal computer is so complex, and information appliances are the solution*. Cambridge, Mass.: MIT Press.

Sears, A. (1997): Heuristic walkthroughs: Finding the problems without the noise. In: *International Journal of Human-Computer Interaction*, 9 (3), 213–234.

Shneiderman, B. (1987): *Designing the user interface: Strategies for effective human-computer interactions*. Reading, Mass.: Addison-Wesley.