

Effectiveness of Participatory Heuristic Evaluation: Case study with a Guideline Based Health Information System

L.W. Peute^a , M.M. van Engen-Verheul^a, E. Kilsdonk^a, N. Peek^a, M.W.M. Jaspers^a

^a Department of Medical Informatics, Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands

Abstract

This study reports on a Participatory approach to Heuristic Evaluation (PHE). In PHE Human Computer Interface experts are combined with (clinical) domain experts to achieve broader detection coverage of usability issues in the context of Health Information Systems (HIS). PHE effectiveness is validated by Think Aloud User Testing on a prototype guideline based HIS. The results of this study show that overall PHE effectiveness was low though comparable to effectiveness measurements of standard HE. The surplus value of PHE was visible in the PHE detection scope of clinical domain specific flaws and usability issues supplementary classified as customization issues.

1. Introduction

Guideline-based Health Information Systems (HIS) have the potential to improve patient care by guiding healthcare practitioners in diagnosis and therapy decisions [1]. However, disruption of/by? the system on the flow of clinical activities and discomfort in system interaction is known to limit their use [2]. Usability evaluation studies on these types of systems can therefore be considered exceptionally complex; requiring coverage of both general interface design flaws as well usability problems that concern the fit between the system and the clinical working practices of end-users.

Think Aloud (TA) usability testing is generally considered the ‘gold standard’ in usability evaluation by providing insight into these genuine user system interaction problems [3]. Nonetheless, accurate performance of the method is time and cost expensive. Usability interface inspection methods such as Heuristic Evaluation (HE) are considered discount techniques; requiring a selection of 3 to 5 Human Computer Interaction (HCI) experts to detect system interface violations to heuristic principles of good system design [4]. A deficit of this approach is that traditional HE is limited to revealing ‘general interface design’ usability problems [5]. Zhang at al. modified and extended a generally accepted heuristic set to make it more applicable in evaluating the usability of medical devices [6]. However, their extended focus was limited to usability issues related to patient safety. With the aim to advance the performance of HE in detecting clinical domain specific usability problems Scandurra et al. introduced clinicians as clinical domain experts to perform HE themselves on prototype health information systems [7]. In their study clinicians were able to emphasize the clinical domain specific usability problems residing in the system design but failed to notice the HCI specific usability problems.

By combining HCI experts with clinical domain experts in a Participatory approach to HE (PHE) potentially broader coverage of usability issues in the context of HIS might be achieved. In this study the effectiveness of PHE in predicting genuine user interaction

problems of a prototype guideline based information system is explored. We first validate the results of the PHE in TA sessions with 7 naïve potential system end-users. Subsequently, all usability problems revealed by PHE and TA detected and their corresponding heuristic classification in the system are analyzed in-depth. The use of TA to determine the effectiveness of PHE and the potential surplus value of combining clinical domain experts with HCI experts in a participatory HE approach are furthermore discussed.

2. System background

The MediScore CARDSS 2.0 system is based on the Cardiac Rehabilitation (CR) guidelines in the Netherlands, and provides an electronic patient record system with computerized decision support facilities. MediScore CARDSS 2.0, was recently developed by ItéMedical BV, a Dutch commercial vendor in healthcare IT. CR is a multidisciplinary therapy that is provided after cardiac events (e.g. myocardial infarctions) and cardiac interventions (e.g. heart surgery). Via a structured dialogue the system actively guides its users through the CR guideline to enter data for a CR needs assessment, consisting of: registration of the patient, entering data items concerning the patient's physical condition, psychological condition, social condition, cardiovascular risk profile and lifestyle, and finally identifying goals and interventions to formulate a preliminary rehabilitation programme containing the recommended therapies [8]. The PHE and the TA sessions were performed on a beta version of MediScore CARDSS 2.0 which was specifically made available for the purpose of the usability evaluation.

3. Methods

3.1 Participatory Heuristic Evaluation

For the PHE three pairs of evaluators were composed, each including a CR domain expert and a HCI expert. A mobile usability laptop with the beta version of the MediScore CARDSS 2.0 was available to each pair on which to perform the PHE study. Pairs were instructed to explore and inspect the interface of the system on potential usability problems and assess whether the revealed problems violated heuristic(s) described by Zhang et al [4]. As is common in standard heuristic evaluation the evaluators were not limited to the applied heuristics in their inspection of the interface but rather were instructed to inspect all usability issues they regarded relevant. To support matching of detected usability problems to the heuristics an instructional document was provided. This document contained a description and example of each of the 14 heuristics defined by Zhang. In addition, a structured problem-report form was given on paper to each pair with the following items: 1) description of the problem, 2) severity rating (1 - minor to 4 - catastrophe) defined by Nielsen et al. [4], 3) place interaction moment at which the problem was encountered and 4) the heuristic(s) violated.

The resulting lists of reported usability problems were subsequently examined by a usability specialist (LP) to aggregate the three resulting problem report forms in one master usability problem list. Within evaluator duplicates and overlapping usability problems were disregarded. Additional matching of all resulting problem descriptions to the heuristics was performed by LP for validation. Problem descriptions not classified by the Zhang heuristics were analyzed in terms of their content and system type specificity. The work of Kaplan on

evaluation of medical informatics applications [9] supported the classification of ..provided insights to classify these problems into domain-specific supplementary heuristics.

3.2 Think Aloud user testing

TA usability testing was performed with seven representative naïve end-users of the system coming from large and small hospital organizations and specialized rehabilitation clinics. Participants were requested to perform a CR needs assessment in the system by entering data from 1) a fictitious patient case and 2) a real patient from their own clinic. In both cases basic system functionalities of the system were covered by providing the participants with 7 system tasks to be performed. For the TA testing a mobile usability laptop with Morae™ software was used to record all verbal protocols and capture screen recordings including facial expressions of participants. All verbal protocols of participants were transcribed and verbal utterances and related video analysis were coded semi-bottom up by two usability experts (MvE, EK, LP) independently to reveal usability problems. Usability problems detected per participant and per CR assessment case were listed and each list was assessed by LP to discard within and between participant duplicates and merge all lists to one master usability problem TA list. Each usability problem description in the master TA list was subsequently given a heuristic classification. The same heuristic classification scheme used in the PHE was applied as reference model. This supported the matching between problems detected in the PHE and the TA usability testing.

3.3 Method Effectiveness in Predicting Usability Problems

To perform valid and accurate matching of usability problems predicted by PHE and observed in the TA sessions a two-way mapping procedure (forward and backward matching) was performed. First, usability problems and their corresponding heuristic classification in the HE master problem list were mapped on the usability problems in the TA master problem list. Then the usability problems and their heuristic classification in the TA master list were mapped on the PHE master list.

To assess the effectiveness of PHE we first assessed the performance of the method by studying the match between the usability problems predicted by PHE and those detected by the TA. The matching results were assessed in terms of *hits* (issues predicted by PHE and detected in TA), *False alarms* (Issues predicted but not encountered in TA) and *Misses* (issues not predicted but encountered in TA). Subsequently, the overall effectiveness of the PHE was computed based on the formulae by Hartson et al [10]: Effectiveness = Validity * Thoroughness, where Validity = Hits/(Hits + False alarms) and Thoroughness = Hits / (Hits + Misses). Weighted effectiveness of PHE for usability problems with 3 (moderate) and 4 (catastrophic) severity was additionally measured.

4. Results

In the working prototype of the Mediscore CARDSS 2.0 system PHE detected 60 usability problems with a mean severity of 2.4 and TA detected 91 usability problems with a mean severity of 2.8. The two way mapping procedure of the methods' master lists showed that the PHE predicted 35 (38%) of the 91 usability problems revealed in the TA sessions. Based on the applied definition of validity, the 25 problems predicted by the PHE but not encountered

in the TA sessions are considered *false alarms*. These 25 problems had a mean severity of 2.0. The validity of the method computed was 0.58.

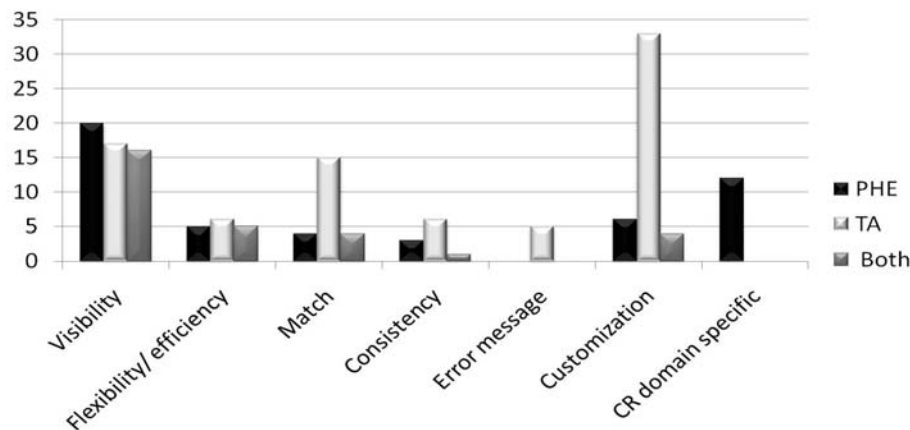
Fifty-six of the 91 problems detected by TA were not predicted by PHE (61%). These 56 problems had a mean severity of 2.9. Based on the definition of thoroughness these problems were considered as *misses*. The computed thoroughness of the evaluators applying the PHE was 0.38. Overall effectiveness of the application of the PHE with 3 pairs of evaluators, computed as the product of validity and thoroughness was 0.22. Seventeen of the 35 problems predicted by PHE and verified by TA had a severity of 3, and 5 of the 35 had a severity 4. Of the 25 predicted but not verified by TA: 5 were considered false alarms for severity 3. For severity 4 no false alarms were predicted. In the not predicted 56 usability problem revealed by the TA, 19 had a severity of 3 and 2 a severity of 4 and were as such considered '*misses*'. Effectiveness of PHE weighted for severity 3 and 4 (respectively 0.37 and 0.71) was high compared to its computed overall effectiveness. Table 1 provides an overview of the computations of Validity, Thoroughness and Effectiveness.

Table 1. Overall and Weighted validity, thoroughness and effectiveness of the PHE

	Overall	Severity 3	Severity 4
Validity (<i>hits/hits+ false alarms</i>)	35/(35+25)=0.58	17/(17+5)=0.78	5/(5+0)=1
Thoroughness (<i>hits/(hits + misses)</i>)	35/(35+56)=0.38	17/(17+19)=0.47	5/(5+2)=0.71
Effectiveness	0.58 * 0.38=0.22	0.78*0.47=0.37	1 * 0.71=0.71

Of the 35 problems predicted by PHE and verified in the TA sessions most (16 problems) concerned unclear visibility of system states, 5 problems concerned the (in)flexibility and (in)efficiency of the system and 4 related to the (mis)match between the system model and the user model of the system. Figure 1 provides an overview of the heuristic categories detected. The PHE also predicted 4 problems which were validated by TA but based on the work of Kaplan were supplementary classified as 'customization' issues, These problems concerned inaccurate entry fields and incorrect orders of clinical items on the screen and were as such subcategorized as problems in (mis)fit of the system to the work practices of end-users and (mis)fit of end-users with the Guideline Based HIS attributable to unfamiliarity with the CR guideline.

Analysis of the 25 problems, considered as 'false alarms', showed that 14 of these problems were classified supplementary to the used heuristic scheme of Zhang. Twelve problems were CR-domain-specific of which most concerned actual errors in CR related information contents and errors in measurement values in the system. Two problems predicted concerned 'customization' issues that highlighted concrete missing information contents in the system, such as missing contact details of a patient's registered general practitioner. Analysis of the 56 problems not predicted by PHE but detected in the TA sessions showed that 27 of the 56 problems for the most part concerned the mismatch between the system and the user's (cognitive) model, interface consistency problems, unclear error messages and issues of system feedback. The remaining 29 of the 56 problems were also supplementary classified as customization issues. However, final classification of these issues led to one more subcategory to the customization classification; fit to institutional setting.



*PHE= Total of detected usability problems by Participatory Heuristic Evaluation, TA = Total of problems detected by Think Aloud, Both = Detected problems by PHE and verified in the TA sessions. CR = Cardiac Rehabilitation. Note that only those heuristic categories are depicted for which 4 or more problems were detected.

Figure 1: Frequencies of heuristic violations by category and method.

5. Discussion and Conclusion

The measured overall effectiveness of the PHE in this study was low (0.22), but the weighted effectiveness for usability problems with severity 3 (major) provided better results (0.37). For usability problems with severity 4 (catastrophic), though based on small numbers, high effectiveness of the method was measured (0.71). Yet, the low overall effectiveness of PHE measured in this study is comparable to the effectiveness of standard HE reported in other studies [11]. Though HE is commonly considered to be a cost-effective method, it is somewhat surprising to assess that its effectiveness in general is low and its claim as discount method is criticized [12].

The PHE in this study missed 62% of the usability problems experienced by end-users in the TA sessions. Part of these ‘unpredicted’ usability problems was related to unclear error messages and the match between the system model and the users’ model. In this study PHE evaluators were instructed to explore the system for usability problems. A general limitation of this approach is that not all aspects of a system are given equal attention, leading to misses in usability problem detection. However, most of the problems missed by PHE but detected by the TA had been supplementary classified as customization issues related to the clinical domain context of the CR guideline. Even so, PHE did find issues classified as customization of which four were indeed verified by TA. In addition, PHE found 12 CR domain specific problems which were not verified by the TA, but which were considered valid problems.

Consistent with the definition of usability (ISO 9241) the TA method was applied as ‘gold standard’ in this study to compute and benchmark the PHE effectiveness. Nevertheless, an acknowledged caveat of employing the TA to compute PHE effectiveness is that TA cannot uncover all problems existing in a system. The 12 CR domain specific problems missed by TA end-users may likewise be explained by users’ cautious approach to the new system and their primary focus on task completion. In contrast, the PHE domain experts critically inspected the systems’ interface in an explorative manner allowing them to identify precisely

this type of problem. However, these 12 CR domain specific problems as well as the 2 customization issues predicted but not verified cannot be considered as misclassifications of *false alarms*. If these issues were taken into account this would lead to a 15% higher effectiveness of the PHE in revealing usability problems existent in the MediScore CARDSS 2.0 system. Even so, 62% of the problems actually experienced by end-users in the TA were still missed by PHE. In terms of the computed effectiveness the PHE did not outperform standard HE. Yet, the results of this study indicate that the potential surplus value of introducing domain experts in a participatory approach to HE might be in the detection of CR domain specific flaws in the system which are not in the scope of TA user task performance. Also, PHE appeared to be effective in detecting usability problems of a catastrophic nature. Future research on improving HE might benefit from the results presented in this study. Furthermore, research is needed to assess the value of the domain specific classifications applied in this study to extending the HE knowledge to guideline based HIS.

Competing interests – None. **Acknowledgement** – The authors would like to thank Ité Medical B.V. for their collaboration during the study. **Funding** – The study was funded by ZonMw, the Netherlands, organization for health research and development.

References:

- [1]. Shiffman RN, Liaw Y, Brandt CA, Corb GJ: 'Computer-based guideline implementation systems: a systematic review of functionality and effectiveness', *J Am Med Inform Assoc*, 6(2),104-14, (1999).
- [2]. Goud R, de Keizer NF, ter Riet G, Wyatt JC, Hasman A, Hellemans IM, Peek N: 'Effect of guideline based computerised decision support on decision making of multidisciplinary teams: cluster randomised trial in cardiac rehabilitation', *BMJ*, 338:b1440 (2009).
- [3]. M.W. Jaspers: 'A comparison of usability methods for testing interactive health technologies: methodological aspects and empirical evidence', *Int J Med Inform*, 78 (5), 340-53, (2009) .
- [4]. Nielsen J: 'Heuristic evaluation', in J. Nielsen, R.L. Mack (eds), *Usability inspection methods*, Wiley, New York, 25-62(1994).
- [5]. Yen P, Bakken S: 'A comparison of Usability evaluation methods: Heuristic Evaluation versus end-user Think-Aloud Protocol – An example from a web-based communication tool for nurse scheduling', "AMIA Annual Symposium proceedings", American Medical Informatics Association, 714-718 (2009).
- [6]. Zhang J, Johnson TR, Patel VL, Paige DL, Kubose T: 'Using usability heuristics to evaluate patient safety of medical devices', *J Biomed Info*, 36,23- 30, (2003).
- [7]. Scandurra I, Hagglund M, Engstrom M, Koch M: 'Heuristic evaluation performed by clinicians: education and attitudes', *Inform Techn in Healthcare*,, 105-215 (2007).
- [8]. Netherlands Society for Cardiology (NVVC): 'Dutch Clinical algorithm for assessment of patient needs in cardiac rehabilitation and secondary prevention', Utrecht: NVVC (English version available at <http://kik.amc.uva.nl/KIK/reports/TR2011-03.pdf>. Last accessed 08-03-2011), (2010).
- [9]. Kaplan B: 'Evaluating informatics applications – some alternative approaches: theory, social interactionism, and call for methodological pluralism', *Int J Med Inf* ,; 64, 39-56, (2001).
- [10]. Hartson HR, Andre TS, Williges RC: 'Criteria for evaluating usability evaluation methods', *Int J Human-computer Interaction*, 13 (4), 241-444, (2001).
- [11]. Hvannberg, E, Law E L-C, Lárusdóttir M: 'Heuristic Evaluation: Comparing Ways of Finding and Reporting Usability Problems', *Interacting with Computers*, 19(2), 225-240, (2007).
- [12]. Law EC, Hvannberg E: 'Analysis of strategies for improving and estimating the effectiveness of heuristic evaluation', *Proceedings of the third Nordic conference on Human computer interaction, "NordiCHI 04"*, 241-250, ACM Press, (2004).