e-Health – For Continuity of Care
C. Lovis et al. (Eds.)
© 2014 European Federation for Medical Informatics and IOS Press.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License.
doi:10.3233/978-1-61499-432-9-930

Revisiting Heuristic Evaluation Methods to Improve the Reliability of Findings

Mattias GEORGSSON[ab1], Charlene R. WEIR[b,c] and
Nancy STAGGERS[bd]
[a] *Blekinge Institute of Technology School of Computing, Sweden*
[b] *Department of Biomedical Informatics, University of Utah, USA*
[c] *Veterans Health Administration, USA*
[d] *School of Nursing, University of Maryland, USA*

Abstract. The heuristic evaluation (HE) method is one of the most common in the suite of tools for usability evaluations because it is a fast, inexpensive and resource-efficient process in relation to the many usability issues it generates. The method emphasizes completely independent initial expert evaluations. Inter-rater reliability and agreement coefficients are not calculated. The variability across evaluators, even dual domain experts, can be significant as is seen in the case study here. The implications of this wide variability mean that results are unique to each HE, results are not readily reproducible and HE research on usability is not yet creating a uniform body of knowledge. We offer recommendations to improve the science by incorporating selected techniques from qualitative research: calculating inter-rater reliability and agreement scores, creating a codebook to define concepts/categories and offering crucial information about raters' backgrounds, agreement techniques and the evaluation setting.

Keywords. Heuristic Evaluation, eHealth, mHealth, Usability Research Methods

Introduction

The hallmarks of building science include the reliability and reproducibility of findings. Thus, commonly accepted research methods include techniques such as a priori definitions, initial consensus discussions among evaluators and calculations for inter-rater reliability (IRR) before independent evaluations begin. On the other hand, usability methods such as heuristic evaluation (HE) run counter to these time-honored techniques. HE emphasizes independent, individual evaluations and the generation of unique usability problems that are subsequently compiled and assigned to general heuristic categories. Expert evaluators are specifically cautioned not to cooperate during initial product assessments. While these methods allow for independent thinking and maximizing initial usability problem identification, they also make the results unique to each evaluation and are not, therefore, readily reproducible. Building science using this technique is difficult due to the large variability across evaluators and the lack of consensus on terms and definitions of usability problems. In this paper we use a current HE as a case study and offer suggestions to modify the technique to improve the reliability and reproducibility of heuristic evaluation findings. These modifications

[1] Corresponding Author.

would result in important additions to the HE method and building a body of knowledge for usability evaluations employing heuristic evaluations.

1. Methods

The HE method is one of the most common in the suite of tools for user-centered design because it is fast, inexpensive and resource-efficient while yielding many usability issues [1, 2]. It is an informal expert usability evaluation first designed by Nielsen and Molich, [3, 4] published in 1994 [5] and is still used today [6]. The classic method has 10 heuristic categories assigned to a usability problem: H1 - Visibility of system status, H2-Match between the system and real world, H3 - User control and freedom, H4-Consistency and standards, H5 - Error prevention, H6 - Recognition rather than recall, H7 - Flexibility and efficiency of use, H8 - Aesthetic and minimalist design, H9 - Help users recognize, diagnose and recover from errors, H10 - Help and documentation [5]. Severity ratings are then determined according to impacts using this scale: 0 Not a usability problem at all, 1 Cosmetic problem only, 2 Minor usability problem, 3 Major usability problem, 4 Usability catastrophe [5].

Identified usability experts independently use typical user tasks to interact with the product's interface, determine existing usability problems and assign specific heuristic(s) violation(s) to each identified problem. Nielsen specifically indicates that communication among evaluators occurs only after the independent evaluations. Only then are the problems combined and discussed among the experts. After having reached consensus on the initial ascribed problems, any duplicates are removed, and the problems are compiled into a master list. The master list is sent out anew to each evaluator to be rated independently for severity and then averaged for each issue.

Nielsen and Molich stated their method is improved by having several evaluators because they provide collective expertise to the process [3]. Nielsen [6] recommended three to five evaluators identify usability issues and provide the associated severity ratings [1]. Three to five expert evaluators find on average between 74% and 87% of the extant usability problems [6]. Results are even higher if the expert evaluators have dual domain backgrounds such as expertise in both usability and the topic/domain at hand. As few as two to three dual domain experts typically find significantly more problems, e.g., between 81% and 90% of existing usability problems [6].

1.1. Sample Project to Illustrate HE Methodological Issues

To illustrate typical issues with the current heuristic evaluation method, we use a recent HE process as an example. The data were part of a larger project for a mHealth self-management system for diabetes patients where patients were able to monitor online parameters for their disease such as glucose levels, blood pressure, exercise and weight Three double domain experts (usability and PhD-prepared registered nurses) evaluated the diabetes mHealth application using Nielsen's HE guidelines described above [5, 6].

To begin, each evaluator viewed a short video demonstrating the application. The evaluators had a guide with common scenarios for interacting with the system and tasks representing patients' typical use such as entering and modifying values, interpreting measurements, and set goals. These preliminary steps were done to increase the consistency of the evaluation process across experts and to assure all evaluators had the same familiarity with the application. In congruence with Nielsen's guidelines, the evaluators performed the same eight tasks during their interactions with the application and independently identified usability problems to determine their compliance with Nielsen's published heuristics.

The evaluators discussed identified usability problems after completing the evaluations. They compiled a master list of problems by consensus. The master list was then sent out to evaluators for their independent severity ratings; these were averaged to determine an overall severity rating for each usability problem.

Nielsen is silent about the need for inter-rater reliability (IRR) in the HE process. However, for the purposes here and to illustrate our points about lack of reliability across all evaluators, we used Krippendorff's α (alpha), an IRR method recommended for use with more than two raters. This method accommodates both larger and smaller sets of data as well as missing data. It is particularly recommended for fully-crossed designs (using the same set of evaluators) and ordinal data [7, 8]. We used SPSS statistical software program (version 22.0) with the KALPHA macro [8].

2. Results

Initial independent ratings resulted in 141 usability problems and 289 assigned heuristic violations. Evaluator 1 detected 86 usability problems, evaluator 2 found 33 usability problems and evaluator 3 discovered 22 usability problems. The initial problems were at various levels of conceptualization, levels of granularity and the issues overlapped in some cases. IRR on the problems and heuristic categories was not able to be calculated due to the varying conceptualizations, although the variability across evaluators was clear. Subsequently, the evaluators examined the usability problems in detail, discussed and consolidated them where feasible. Only nine initial problems were the same across all evaluators and these were consolidated into three overall usability problems. Two evaluators found 12 usability problems that were similar and these were consolidated into six issues. The remaining 120 were unique usability issues. This process resulted in a final master list of 129 usability issues.

Severity ratings had similar variability. Evaluators rated 17 of the 129 problems with the same severity score. Two of the three evaluators had the same rating for 90 problems, but for the remaining 22 usability issues, no similarities existed. At this point in the process, Nielsen recommends calculating mean severity scores, but we were curious about the extent of variability. The percentage of agreement between the three evaluators on all problems was 13.18%. When Krippendorff's α was calculated with the KALPHA macro for inter-rater reliability the result was very low - 0.0815 (Table 1).

Table 1. Krippendorff's α Reliability Estimate

A	Units	Observers	Decisions
0.0815	129	3	387

3. Discussion

The wide variability across even dual-domain evaluators is evident. For initial usability problem identification, the numbers, levels and types of usability problems varied as did the assigned heuristics. Only nine common usability problems were identified across all evaluators and 120 issues were unique. For the severity scores, the inter-rater reliability (IRR) was a mere 13.18% (KALPHA less than α 0.10) across the three evaluators. This variability existed despite the fact that the three evaluators were prepared at the PhD level, were informatics experts and dual domain experts in both nursing and usability methods. The reasons for the differences include variation backgrounds, expertise and human judgment that is common across most domains. Another possible for the reason for this variability is that the underlying dimensions of usability are, in themselves, not valid. These differences were the genesis for this article.

This variability is not atypical of HE results. Past usability experts have not quantified the extent of the existing variability or called attention to its implications. After reviewing numerous publications, we conclude that HE evaluators have not in the past and do not currently calculate IRR. From the perspective of our ability to build science, extensive variability is a concern and additions to the HE method from qualitative research are recommended.

3.1. Recommendations for Improving Heuristic Evaluation Techniques

The scientific goals of assessing reliability of measurement in heuristic evaluation (HE) are two-fold. First, establishing reliability of judgment supports the validity of the categories. Further work should address questions of validity, e.g., two disease states may have the same signs and symptoms, but we need characteristics to distinguish between the two or we have to conclude that they are the same disease. A second perspective is to assume that the categories exist and can be measured the same by different raters (IRR). These speak to the generalizability of findings.

Improving HE techniques requires distinguishing between these two goals. Establishing the first goal requires assessing what is traditionally known as IRR (inter-rater reliability). IRR metrics vary depending on the type of data, but Kappa, Kendall coefficient of concordance and Krippendorff's α are recommended for ordinal and categorical data, e.g. HE methods, as they correct for chance agreement. Although a high IRR supports that the individual raters can reliably discriminate categories and therefore, the categories have validity, it doesn't say much about the degree of agreement between raters, which would relate to the generalizability of results. To be able to make the claim that different raters would "see" the same categories, a measure of inter-rater agreement (IRA) is also required. The simplest of these is the percent agreement. Some metrics include both, such as the Bland-Altman plots and the Intra-class correlation metric usually used for interval or ratio data [9-11].

In the field of usability, the next step in building a scientific foundation is to both report these two statistics and to conduct the necessary work to reach an acceptable level. To more fully understand what the usability problems, HE categories and severity assessments mean in terms of future research and development, information is also necessary about the background of the raters, specific procedures used to build IRA and the evaluation settings. While creating consensus on a master list is helpful, a priori IRR/IRA processes provide better reliability.

We strongly advise the creation of a codebook, an iterative process of establishing IRA/IRR. The codebook is the information needed by other researchers to establish reproducibility of results and should be provided in the appendices of published HE assessments. Establishing IRA is a rich experience that allows researchers to more fully comprehend the nature of the phenomena being investigated; science will be significantly improved as a result of adding this level of depth to the HE process [12].

The benefits for modifying HE techniques are clear. A priori definitions, common understanding of usability problems/categories derived from initial consensus discussions, and obtaining adequate IRR/IRA before independent expert evaluations would allow evaluators to create reliable and reproducible results for HE assessments in the future. A challenge is also obvious. A priori work will mean it will take more time to create the products that will improve the consistency of results across and among evaluators. This is in contrast to the published HE benefits, that HE is a discount usability technique cited as fast and resource efficient [1, 3]. However, this challenge is off-set with the knowledge that others could use these products and obtain consistent results.

4. Conclusions

Current methods in heuristic evaluation promote significant variability in usability problem generation and severity ratings even across dual domain expert evaluators. In contrast, building science requires techniques that emphasize reliability and the reproducibility of results. In this paper, we used a case study to illustrate the wide variability of findings across dual domain experts. We recommend that future usability researchers incorporate several crucial qualitative research techniques into heuristic evaluations. Specifically, we recommend developing a coding manual, calculating inter-rater reliability and inter-rater agreement among evaluators and providing information about evaluators and settings. These modifications will allow evaluators to be more consistent, results to be reproducible and a more uniform body of knowledge in usability evaluations will be available.

References

[1] Nielsen J. Usability engineering. Boston: Academic Press; 1993. xiv. p. 358.

[2] Jeffries R, et al., User interface evaluation in the real world: A comparison of four techniques, in Proceedings of the SIGCHI Conference on Human Factors in Computing System; New Orleans, Louisiana, USA; 1991. p. 119-124.

[3] Nielsen J, Molich R. Heuristic evaluation of user interfaces. In: Proceedings of the SIGCHI conference on Human factors in Computing Systems: Empowering People;Seattle, Washington, United States; 1990. p. 249-256.

[4] Molich R, Nielsen J. Improving a human-computer dialogue. Commun ACM. 1990; 33(3): 338-348.

[5] Nielsen J. Heuristic evaluation. In: Mack RL, Nielsen J, editors. Usability inspection methods. New York: John Wiley & Sons inc.; 1994. p. 25-62.

[6] Nielsen J. Ten usability heuristics. Available from: http://www.nngroup.com/articles/ten-usability-heuristics. Accessed April 23, 2014.

[7] Nielsen J. Finding usability problems through heuristic evaluation. Proceedings of the SIGCHI conference on Human factors in computing systems; Monterey, California, United States. ACM; 1992. p. 373-80.

[8] Hayes AF, Krippendorff K. Answering the Call for a Standard Reliability Measure for Coding Data. Commun Meth Meas. 2007;1(1):77-89.

[9] deVet H, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability measures. J Clin Epidemiol. 2006;59:1033–39.

[10] Gisev N, Bell JS, Chen TF. Inter-rater agreement and inter-rater reliability: Key concepts, approaches and applications. Res Soc Admin Pharm. 2013; 9(3), 330-8.

[11] Kottner JL, Audige L, Brorson S., et al. Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. J Clin Epidemiol. 2011; (64): 96–106.

[12] Streiner DL, Norman GR. Health Measurement Scales: A Practical Guide to Their Development and Use, 4th ed. New York, NY: Oxford University Press; 2008.