Expert vs Novice Evaluators
Comparison of Heuristic Evaluation Assessment

Magdalena Borys and Maciej Laskowski
Institute of Computer Science, Lublin University of Technology, Nadbystrzycka 38D street, Lublin, Poland

Keywords: Usability, Website Usability Evaluation, Heuristic Evaluation.

Abstract: In this paper authors propose the comparison between results of website heuristic evaluation performed by small group of experts and large group of novice evaluators. Normally, heuristic evaluation is performed by few experts and requires their knowledge and experience to apply heuristics effectively. However, research involving usability experts is usually very costly. Therefore, authors propose the experiment in order to contrast the results of evaluation performed by novice evaluators that are familiar with assessed website to the results obtained from expert evaluators in order to verify if they are comparable. The usability of website was evaluated using the authors' heuristics with extended list of control questions.

1 INTRODUCTION

The notion of usability gained its popularity in the early 1980s as one of key factors of software quality. Since then, the different methods and approaches have been proposed to evaluate the usability of an user interface and interactions with system at each stage of software development process.

Taking into consideration the multitude of usability evaluation methods, the main focus is laid on the relative effectiveness of empirical usability studies in opposition to other, less costly, methods. The expert based methods hold the promise of usability results that are keeping costs low by relying on expert review or analysis of interfaces rather than observing actual user behaviour (Hollingsed and Novick, 2007). Despite the fact that expert based methods usually do not require special equipment and lab facilities, it should be taken into account, that experts with appropriate skills and level of experience are hard to find and still cannot be included in all software project budgets.

The experiment presented in the paper shows that the results of evaluation performed by novice evaluators who are familiar with assessed website and have been introduced to some basic concepts of usability can be compared to the results obtained by experienced expert evaluators. And therefor to answer research question*: it is possible to substitute few skilled and experienced evaluator with large number of novice evaluators familiar with evaluated subject*?

The problem of novices vs experts is not a new one, but the experiment described in this paper was based on authors' own heuristics with extended list of control questions.

2 RELATED WORKS

While analysing the problem of novice and experienced evaluators, several existing papers, such as the work of Hertzum and Jacobsen (Hertzum and Jacobsen, 2001) should be taken into account, where the evaluator effect for both novice and experienced evaluators in usability evaluation methods (UEMs) such as heuristic evaluation, cognitive workthrough and thinking-aloud study was investigated. They stated that despite used UEM the average agreement between any 2 evaluators who have evaluated the same interface using the same evaluation method will ranges from 5% to 65%. What is more, none of the investigated UEMs was considered as consistently better than the others in revealing evaluator effect As the simplest solution for coping with evaluators is to involve multiple evaluators in usability evaluation.

The evaluation performance of two types of cognitive styles was explored by Ling and Salvendy (Ling and Salvendy, 2009). In the proposed experiment the results indicate that the field independent individuals, who tend to be less

Borys M. and Laskowski M..
Expert vs Novice Evaluators - Comparison of Heuristic Evaluation Assessment.
DOI: 10.5220/0004970901440149
In *Proceedings of the 16th International Conference on Enterprise Information Systems* (ICEIS-2014), pages 144-149
ISBN: 978-989-758-029-1

influenced by the information from the visual fields and consider all the other information from senses, performed evaluation with significantly higher thoroughness, validity, effectiveness and sensitivity than the field dependent individuals, who tend to be greatly influenced by the dominant visual field.

The set of experiments comparing effectiveness of MOT technique (inspection by metaphors of human thinking) with heuristic evaluation, cognitive walkthrough and "think aloud" testing for novice evaluators was introduced by Frøkjær and Hornbæk (Frøkjær and Hornbæk, 2008). In experiments they demonstrated that MOT was more useful as an inspection technique for novice evaluators that heuristic evaluation - the evaluators found an equal number of problems with the two methods, but problems found with MOT are more serious and complex to repair, and more likely to persist for expert users. However, understanding MOT as a technique for evaluating interfaces appeared to be difficult.

Another comparative study was proposed in (Lanzilotti et al., 2011). The study involves novice evaluators and end users, who evaluated an e-learning application using one of three techniques: pattern-based inspection, heuristic evaluation and user testing. In the study authors show that pattern-based inspection reduces reliance on individual skills and permits the discovery of a larger set of different problems and decrease evaluation cost. Moreover, the results of study indicated that evaluation in general is strongly dependent on the methodological approach, judgement bias and individual preferences of evaluators. Authors also state that patterns help to share and transfer knowledge between inspectors and thus simplify the evaluation process for novice evaluators.

More experiments on how to improve heuristic evaluation done by novice evaluators was performed by Botella, Alracon and Penalver (Botella, Alracon and Penalver, 2013). They proposed the framework for improving usability reports for novice evaluators by combining the classical usability report with the interaction pattern design. However, they did not provide any evidence of method usage or its comparison to other techniques.

3 HEURISTIC EVALUATION METHOD

Heuristic evaluation is one of the most widely used methods for application evaluation. While using the application, an expert checks and marks the predefined areas in order to note the compliance with interface design guidelines called also heuristics and look for potential problems.

3.1 General Description

In heuristic evaluation method, each of those predefined areas can be divided into several more detailed sub-areas and be assigned with questions for the expert to answer while working with an application.

The main advantage is the method cost, which does not require representative samples of users, special equipment or laboratory settings. Moreover, experts can detect wide range of system problems in a limited period of time. As the main drawback studies list the dependency on experts' skill and the fact that experience as heuristics are often generic (Lanzilotti et al., 2011). Other studies list that heuristic evaluation can lead to situation when many small and minor usability problems are detected and improved whereas major usability problems remain unnoticed (Koyani, Bailey and Nall, 2004).

Contrary to the widespread assumption, experts usually do not acquire better results in performing specific tasks in the tested system, as they usually do not know that system before the testing. But their expert status is based on their own experience with different kinds of software. This, as proven by studies, allows them to perform faster than the novices (Dillon and Song, 1997) and to spend less time handling the errors despite making number of errors comparable to the novice users (Jochen et al., 1991).

The most known guidelines concerning user interface are:

- Nielsen's heuristics (Nielsen and Molich, 1990);
- Gerhardt-Powals' cognitive engineering principles (Gerhardt-Powals, 1996);
- Weinschenk and Barker classification (Weinschenk and Barker, 2000);
- Connell's Full Principles Set (Connell, 2000).

3.2 Applied Heuristics

Authors decided to used heuristics that they created and applied in previous research (Borys, Laskowski and Milosz, 2013). The proposed heuristics covers the following areas:

- Application interface.
- Navigation and data structure.
- Feedback, system messages, user help.
- Content.
- Data input.

Table 1 shows the detailed list of areas and subareas ("LUT list") with questions assigned to each point. Accordingly, Table 2 presents the grading scale used to assess each tested area.

Table 1: LUT list of predefined testing areas with questions assigned.

Web application interface	
Layout	Is the layout readable?
	Is it adjusted to different resolutions?
	Is it adjusted to mobile devices?
	Is it consistent?
	Does it support task implementation?
Colour scheme	Is there proper contrast between text and background ?
	Is the colour scheme readable for people with colour vision disorders?
	Is the colour scheme readable on various kinds of displays?
Navigation and data structure	
Ease of use	Is the access to all sections of a web application easy and intuitive?
	Is the access to all functions of a web application easy and intuitive?
Information hierarchy	Isn't the information hierarchy too complicated?
Information structure	Is the information structure understandable for users?
	Is it consistent?
	Is it well planned?
Screen elements	Do they support the navigation?
Feedback, system messages, user help	
System messages (general)	Do they provide enough information on the status of actions performed by user?
System messages (errors)	Do they contain hints on problem solution?
Feedback and user help	Does the information appear in places, where it may be needed?
	Is the provided information understandable for an average user?
	Is the provided information accessible for an average user?
	Is it possible for an average user to perform actions suggested by system help in order to solve the encountered problem?
Content	
Labels	Do the labels used in the interface provide enough information?
	Do all the interface elements have necessary labels?
Naming	Is the interface naming understandable for its users?
	Is the interface naming consistent?
Page text	Is it understandable for users?
Data input	
Data	Is the data validated by the form elements?
	Do the forms have elements acting as hints for the input data (e.g. on format or data range)?
	Can average user fill in the form easily?
Forms	Are they designed in a readable way?
	Are they adjusted to the mobile devices?
	Do they allow user to input all of the necessary data?

Table 2: Grading scale applied to LUT list.

Grade	Description
1	Critical GUI errors were observed, preventing normal usage or discouraging user from using the web application.
2	Serious GUI issues were encountered, which may prevent most users from task realization.
3	Minor usability GUI – ere observed, which if accumulated may have negative impact on user performance.
4	Single minor GUI issues were observed, which may have negative impact on user work quality (e.g. poor readability).
5	No GUI issues influencing either user performance or work quality were identified.

The results of proposed evaluation approach can be used to calculate Web Usability Points as a complex factor (rate) of the usability of websites GUI. WUP metric uses grades (Table 2) granted by experts to each question from the LUT list (Table 1). WUP for websites can be calculated using following formula:

$$WUP = \frac{1}{n_a}\sum_{i=1}^{n_a}\frac{1}{s_i}\sum_{j=1}^{s_i}\frac{1}{q_{ij}}\sum_{k}^{q_{ij}} p_{ijk} \quad (1)$$

where:

n_a - number of areas,

s_i - number of subareas in *i*-area,

q_{ij} - number of questions in *i*-area and *j*-subarea,

p_{ijk} - grade value (points) granted to *k*-question in *j*-subarea in *i*-area.

The value of WUP varies from 1 to 5. The higher value, the better usability of the interface.

4 EXPERIMENT AND RESULTS

4.1 Research Question

The goal of our experiment was to examine two set of results of web usability assessment based on heuristic evaluation method from the point of view of the following research question: "*Whether the website usability evaluation results based on well-defined heuristic from large group of novices are comparable to results gained from small group of experts in the domain of web usability?*"

4.2 Research Hypotheses

To examine the research question following research

hypothesis was formulated: "*Results of evaluations provided by large group of novices and results of evaluations provided by small group of experts are comparable.*"

4.3 Research Methodology

The research hypotheses were verified by experimental works. The experiments were conducted on university websites. The heuristic evaluation were performed by 26 computer science students who had no or very little knowledge in usability domain and by 4 experts in web usability. All students were familiar with the websites before the evaluation which facilitate theirs performance in evaluation. In order to make sure that all the notions and ideas of website usability and heuristics are clear and understood, all the students were involved in a hour training before the experiment was conducted.

Both groups, novices and experts, used heuristics described in section 3.2 to guide their evaluations.

Despite the fact that all evaluations consisted of numerical assessment of each section and a list of detected usability problems, only numerical assessments were taken into consideration during statistical analysis. Statistical analysis were performed in program *Statistica 10*.

4.4 Results

The means and standard deviations for each testing areas of LUT list for whole population, novice and expert evaluators group are shown in Table 3. What is interesting, all plots shows greater divergence in experts' assessment than in novices' one.

Distributions for each testing areas (except *Application interface)* area in novices and experts groups are normal (based on Kolmogorov-Smirnov normality test with $p > 0.95$). Therefor the t-tests for two independent group for each testing areas (except *Application interface*) were performed and indicated no statistical significant difference between groups.

The box-and-whisker plot for two independent groups – novices and experts – for each testing area of applied usability heuristics are presented on Figure 1-4.

Since distribution for Application interface testing area does not follow normal distributed data, nonparametric test for two independent group were performed. Both Mann-Whitney U test and Wald-Wolfowitz runs test do not indicated any significant differences (for $p < 0.05$), therefor there is no reason to reject the hypothesis that results from novice and expert evaluators are comparable.

Table 3: Group means and std. deviations for testing areas.

Testing area	N	Mean	Std. deviation
Population			
1	30	3.91	0.77
2	30	3.22	1.16
3	30	3.53	0.69
4	30	3.78	0.82
5	30	3.29	0.88
Group: Novice evaluators			
1	26	3.99	0.69
2	26	3.24	1.16
3	26	3.45	0.67
4	26	3.75	0.83
5	26	3.25	0.84
Group: Expert evaluators			
1	4	3.39	1.13
2	4	3.13	1.31
3	4	4.06	0.72
4	4	4.00	0.82
5	4	3.54	1.23

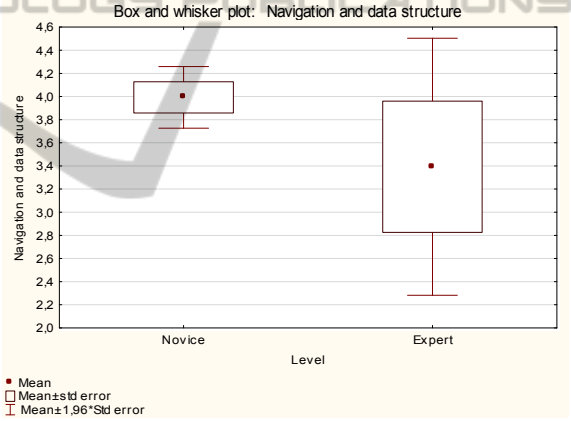

Figure 1: Box and whisker plot for two independent group: Navigation and data structure.

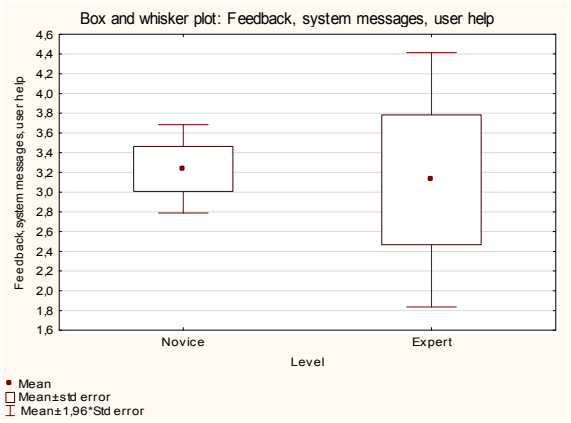

Figure 2: Box and whisker plot for two independent group: Feedback, system messages, user help.

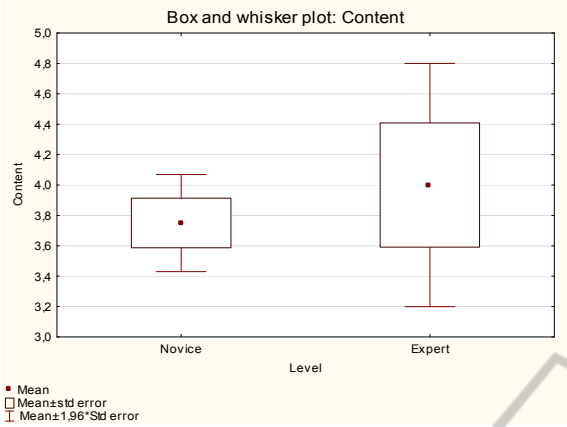

Figure 3: Box and whisker plot for two independent group: Content.

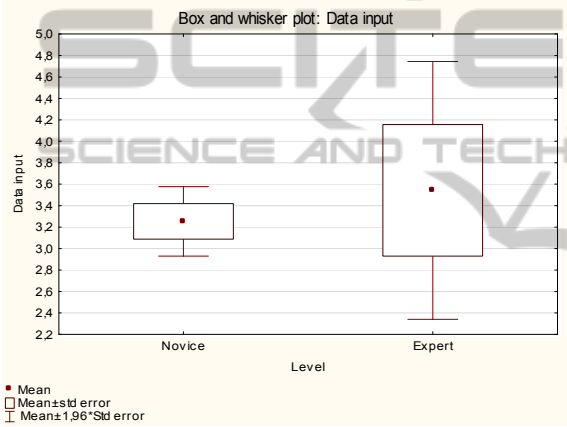

Figure 4: Box and whisker plot for two independent group: Data input.

5 CONCLUSIONS

The main purpose of this research was to investigate if the usability assessment performed by novices provides comparable results in the web usability evaluation using heuristics to the results obtained by experts. Hypothesis has been fully confirmed.

Moreover, the LUT list of predefined testing areas helped both experts and novices to perform their evaluation in a systematic and methodical manner.

Authors came across a few interesting research problems to be solved in the future, such as:

- what is the lower limit of the number of novices in order to get the same (or comparable) results as the group of experts?
- what are the limitations of the described method in a number of tested applications or websites?
- what is the influence of length of the experiment on both experts' and novices' performance?

The aforementioned questions are just a contribution to further studies and experiments, which may result in interesting outcomes.

ACKNOWLEDGEMENTS

Authors would like to thank dr Malgorzata Plechawska-Wojcik, Marcin Badurowicz and Kamil Żyła and to all students who took part in the experiment for their contribution to this paper.

Authors are the participants of the project: "Qualifications for the labour market - employer friendly university", co-financed by European Union from European Social Fund.

REFERENCES

Borys, M., Laskowski, M., Milosz, M., 2013. Memorability experiment vs. expert method in websites usability evaluation. In *ICEIS 2013 - Proceedings of the 15th International Conference on Enterprise Information Systems*, 3, SciTePress.

Botella, F., Alarcon, E., Penalver, A., 2013. A new proposal for improving heuristic evaluation reports performed by novice evaluators. In *Proceeding ChileCHI '13 Proceedings of the 2013 Chilean Conference on Human - Computer Interaction.* ACM, NY, USA, pp. 72-75.

Connell, I. W., 2000. Full Principles Set. Set of 30 usability evaluation principles compiled by the author from the HCI literature. In http://www0.cs.ucl.ac.uk/staff/i.connell/DocsPDF/PrinciplesSet.pdf.

Dillon, A., Song, M., 1997. An empirical comparison of the usability for novice and expert searchers of a textual and a graphic interface to an art-resource database. Digital Information.

Frøkjær, E. and Hornbæk, K. 2008. Metaphors of human thinking for usability inspection and design. In *ACM Transaction on Computer-Human Interaction,* 14 (4), pp. 1-33.

Gerhardt-Powals, J., 1996. Cognitive engineering principles for enhancing human - computer performance, In *International Journal of Human-Computer Interaction*, 8(2), pp. 189-211.

Hertzum, M., Jacobsen, N. E. 2001. The evaluator effect: A chilling fact about usability evaluation methods. In *International Journal of Human-Computer Interaction*, 13(4), pp. 421-443.

Hollingsed T., Novick D. G., 2007. Usability Inspection Methods after 15 Years of Research and Practice. In *SIGDOC '07 Proceedings of the 25th annual ACM international conference on Design of communication*, NY, pp. 249-255.

Jochen P. et al., 1991. Errors in computerized office work: Difference between novice and expert users. ACM SIGCHI Bulletin.

Koyani S., Bailey R. W., Nall J. R., 2004. Research-Based Web Design & Usability Guidelines. In Computer Psychology.

Landauer Th. K., 1996. *The Trouble with Computers: Usefulness, Usability, and Productivity*. MIT Press.

Lanzilotti, R., Ardito, C., Costabile, M. F., De Angeli, A., 2011. *Do patterns help novice evaluators? A comparative study*. In International Journal of Human-Computer Studies, 69(1–2), pp. 52-69.

Ling, Ch., Salvendy, G., 2009. Effect of evaluators' cognitive style on heuristic evaluation: Field dependent and field independent evaluators. In *International Journal of Human-Computer Studies*, 67(4), pp. 382-393.

Nielsen, J., and Molich, R., 1990. Heuristic evaluation of user interfaces. In *Proceedings ACM CHI'90 Conference*, pp. 249-256.

Weinschenk, S., Barker, D. T., 2000. Designing Effective Speech Interfaces, Wiley, 1 edition.