On the performance of novice evaluators in usability evaluations

Panayiotis Koutsabasis, Thomas Spyrou, Jenny S. Darzentas and John Darzentas
University of the Aegean
Department of Product and Systems Design engineering
Hermoupolis, Syros, Greece
Konstantinoupoleos 2, GR-84100
kgp {tsp, jennyd, idarz} @ aegean.gr

Abstract

The paper investigates the performance of novice evaluators in usability evaluations by presenting the results of a comparative usability evaluation that was conducted by nine novice evaluator teams. The evaluation teams performed considerably well in terms of the validity of their results, which counts for their participation in usability evaluation projects. The thoroughness of the results obtained was found to be in a relatively stable ratio of about 20-25% of the total number of problems found for eight of the nine teams, which gives a clear indication of the degree to which novice evaluators can identify usability problems. The consistency of the results was not satisfactory, although similar to results of other studies that involve professional evaluators. The paper suggests that when novice evaluators have to be employed for usability evaluations and it is important to find most usability problems then parallel usability evaluations can provide overall valid and thorough results.

Keywords: Novice evaluators, usability evaluation, heuristic evaluation, cognitive walkthrough, think-aloud, co-discovery learning, validity, thoroughness, consistency.

1. Introduction

In principle, all UEMs require the participation of expert evaluators that can either perform the evaluation themselves (according to usability inspection methods) or organise a particular method's requirements for employing users (according to user testing methods). In practice, many usability evaluations are conducted with the guided participation of novice evaluators for many practical reasons of convenience. For example, Hertzum and Jacobsen (2001) refer to 11 usability evaluation studies from which 4 were conducted solely from student evaluators and another 4 from teams that included both expert and novices. Although many researchers who have introduced UEMs have argued for specific skills that usability evaluators should possess (e.g. Nielsen 1994; Lewis and Wharton 1997), knowledge about the effectiveness of results when employing novice evaluators in pragmatic case studies is not readily available.

The paper investigates the performance of novice evaluators in usability evaluation by presenting the results of a comparative usability evaluation of an academic Web site (same evaluation target for all teams) that was conducted by nine teams of novice evaluators. The UEMs employed to conduct the evaluations were: heuristic evaluation, cognitive walkthroughs, the think-aloud protocol and co-discovery learning. The paper provides a methodical comparison of the results of the usability findings of these parallel usability evaluations regarding:

- Validity (or accuracy), i.e. the ratio of the number of real usability problems with respect to the total number of findings (whether problems or not) of each particular evaluation team
- Thoroughness (or completeness), i.e. the ratio of the number of real usability problems found by each evaluation team with respect to the total number of real usability problems that exist in the target system
- Consistency, i.e. the extent to which multiple applications of a particular usability inspection method produce similar results

The paper is structured as follows: the next section discusses related work mainly in terms of refining the criteria for assessing the results of a comparative usability evaluation; section 3 overviews the method used for the conduct of parallel evaluations and interpretation of results; section 4 presents the results of the comparative study in terms of validity, thoroughness and consistency detected; and finally section 5 presents the conclusions of this work discussing the implications in terms of the setup of usability evaluations with novice evaluator participation.

2. Related work

The assessment and comparison of the results of multiple applications of UEMs is a complex task. In the framework of comparing UEMs, Hartson et al (2003) remark that "*researchers find it difficult to reliably compare UEMs because of a lack of standard criteria for comparison; standard definitions, measures, and metrics on which to base the criteria; and stable, standard processes for UEMs evaluation and comparison."* Among the criteria that have been identified for the comparative study of UEMs, the most important for the purposes of this work are those of validity, thoroughness and consistency.

The **validity** of usability evaluation results can be identified (Sears, 1997 and Hartson et al, 2003) as the ratio of the number of 'real' (or relevant) usability problems found divided to the number of total usability findings obtained by the application of the UEM. Thus, high validity results in few 'false alarms' for each evaluation team. Clearly, the 'realness' (or relevance)[1] of usability findings is essentially subjective and needs to be determined in a way that should be explicitly described by any com-

[1] Realness and relevance are used interchangeably in related work.

parative usability evaluation. According to Hartson (2003) the realness of usability findings can be determined by comparing with a standard usability problem list; or expert review and judgment; or finally by end-user review and judgement. Any approach includes advantages and drawbacks regarding applicability, cost-effectiveness and trustworthiness and in this respect further research includes formalistic approaches to address these concerns like severity ratings (Nielsen, 1993) and combinations of severity and probability of occurrence (Rubin, 1994).

Provided the realness problem is reasonably addressed, the **thoroughness** of an application of a UEM can be defined as the ratio of the number of real usability problems identified by the evaluation team or method with respect to the total number of real usability problems that exist in the system (Sears, 1997, Hartson et al, 2003). The total number of usability problems is considered in comparative evaluation studies as the sum of the unique real usability problems identified by methods used. The thoroughness criterion can provide an overall picture about the effectiveness of the results obtained by the application of the UEM.

The **consistency** of results in comparative usability evaluation studies has not been extensively discussed in the literature, although it has been related to reliability (Hartson et al, 2003) and repeatability (Oorni, 2003). In our work, we use as a working definition of consistency the extent to which multiple usability evaluations (which are conducted by different evaluation teams and using different UEMs) produce 'reasonably similar' results. This working definition is similar with the approach followed by Molich et al (2004) in their comparative usability studies.

3. The method

3.1. Evaluation object

The web site evaluated is that of the Department of Product and Systems Design Engineering[2], University of the Aegean, which has been operating since September 2000. The web site was designed to address the emerging needs of the new department and has been extended since by the addition of web-based subsystems (both open source and in-house developments) for the support of administrative and teaching tasks.

3.2. Participants

The usability evaluations were conducted by MSc students of the department in terms of partial fulfilment of their obligations for the course on interaction design. Interaction design is offered at the first semester of the MSc course. The object of the MSc

[2] http://www.syros.aegean.gr

course is the holistic design of innovative interactive products and systems for the information society, with the use of new technologies and creative use of knowledge from a wide range of arts and sciences. The interaction design course aims at introducing students to core interaction design issues including creation of awareness on usability and accessibility, application of UEMs and user-centred, contextual methods for the design of interactive systems.

The students of the MSc course have a wide range of backgrounds about design having graduated from departments such as arts, graphic design, industrial engineering, civil engineering, naval engineering and information systems. All students had considerable knowledge about the web site since that they had used it before the usability evaluation exercise. According to Nielsen (1992) who reports in the context of heuristic evaluation: "*usability specialists are better than non-specialists at performing heuristic evaluation, and 'double experts'* with specific expertise in the kind of interface being evaluated perform even better". Thus the lack of previous experience of selected subjects on usability evaluations was partly compensated by their good knowledge of the target system.

3.3. Usability evaluation methods used

Students were split into three-person teams on the basis of their own preferences to maximise the result of their effort. The teams selected from the following four UEMs:

- Heuristic evaluation: described by Nielsen (1993) as the process of having a small set of expert evaluators examining the interface on the basis of recognised usability principles (the ten heuristics suggested by Nielsen were employed);
- Cognitive walkthroughs: described by Polson et al (1992) as the evaluators' attempt to identify those actions that would be difficult to choose or execute for the average member of the proposed user population.
- The think-aloud protocol: described by Van Someren et al (1994) as "*thinking aloud while solving a problem and analysing the resulting verbal protocols."*
- Co-discovery learning is described by (Miyake, 1986) as the process of having a pair of individuals (users) discuss a topic and work collaboratively on a solution.

For the first two inspection methods, each member of the team first worked alone by walking through the target system in order to fulfill the given tasks and then the team gathered to interpret and organize the presentation of their results. In a similar vein, the teams that carried out the user testing methods started by recruiting users and the preparation of recording material and then carried out the usability evaluation atomically at first and then interpreting and aggregating the results. The user testing teams had the option (for didactic purposes) to enrich the application of the user testing methods with recording equipment and questionnaires. This degree of freedom was allowed for didactic purposes. The methods selected along with their basic features are outlined in (Table 1).

Teams / methods	T1	T2	T3	T4	T5	T6	T7	T8	T9
HE	√	√	√						
CW				√	√				
T-AP						√	√	√	
C-DL									√
Evaluators	3	3	3	3	3	3	3	3	3
Users	-	-	-	-	-	11	8	6	8

Table 1: Methods selected by evaluation teams

The evaluation teams had to test the system by following two user tasks that required substantial navigation within the web site:

- For a student, to locate information about a specific course, such as course description, instructor, online notes, etc.
- For a visitor of the department, to locate necessary information about visiting the department at Hermoupolis, Syros, Greece.

The evaluation teams were provided with an analytic template for documenting the results, which included table of contents for the usability report and a categorisation of types of usability problems. The evaluation teams were given this assignment after the first six lectures on interaction design. They had a two-month period to organise, conduct and document the usability evaluation. Their main deliverables were the usability report and their presentation of their results in an open discussion session. Weekly meetings were arranged with the supervisor of the work to ensure progress, to answer queries and to ensure that UEMs were applied as described by their authors.

3.4. Processing of results

The processing of usability evaluation results refers to making decisions about the realness of results and about interpreting multiple or similar results into single statements of usability problems. When these decisions are made by experts, it is generally advisable that more than one expert performs this task. However the tasks of reviewing and interpreting the data from single usability reports to a single one are tedious for more than one expert. The amount of time required to go through evaluation reports, to process the large pool of data in terms of relevance and similarity and to resolve ambiguities requires many hours of synchronous work. Also, the study of Molich et al (2004) has also used a single expert to go through the data.

In this study, the first author went through the results of the usability evaluation reports. The total number of usability findings (200) by all nine usability evaluations had to be compared with one another in order to decide about their relevance (i.e. whether they were problems) and similarity (i.e. whether they were reporting on the same problem). The relevance of usability findings was determined in terms of a three-scale severity scheme: 0 – not a problem; 1: minor problem; 2: serious problem.

4. Results

4.1. Validity

The validity of usability evaluation results (Table 2) is considerably high in most teams (six from nine, namely: HE1: 94%; HE3: 100%; CW1: 85.7%; T-AP1: 90.5%; T-AP2: 94.4%; T-AP3: 82.4%). This reveals that the large majority of usability evaluation results were valid and useful and that there were not significant false alarms within the reports. This finding is encouraging for the novice evaluators in that it showed that most novice evaluators do not make too many mistakes when they have to decide about whether a finding at the user interface is a usability problem.

However, three teams were identified with a rather large number of false (not real) usability findings: HE2: 28.6%; CW2: 25%; and C-DL: 25.6%. The large majority of falsely identified usability findings were rephrased or re-emphasised instances of a unique usability problem that appeared in more than one part of the web site, which is an issue of evaluator experience.

Teams[3]	Total findings	0: not a problem		1: minor problem		2: major problem		Validity (1+2)	
HE1	18	1	5.6%	6	33.3%	11	61.1%	17	94.4%
HE2	28	8	28.6%	8	28.6%	9	32.1%	17	60.7%
HE3	14	0	0.0%	4	28.6%	10	71.4%	14	100.0%
CW1	21	3	14.3%	4	19.0%	14	66.7%	18	85.7%
CW2	24	6	25.0%	4	16.7%	13	54.2%	17	70.8%
T-AP1	21	1	4.8%	6	28.6%	13	61.9%	19	90.5%
T-AP2	18	1	5.6%	3	16.7%	14	77.8%	17	94.4%
T-AP3	17	3	17.6%	4	23.5%	10	58.8%	14	82.4%
C-D1	39	10	25.6%	14	35.9%	15	38.5%	29	74.4%
	200							162	

Table 2: Validity of usability findings (and severity ratings)

4.2. Thoroughness

The thoroughness of usability evaluation results is specified by the total number of real usability problems identified by each team divided by the total number of real problems that exist in the system (Table 3), which is the sum of unique real problems identified by all methods.

[3] HE - Heuristic evaluation; CW: Cognitive Walkthrough; T-AP: Think-Aloud Protocol; C-D: Co-Discovery

The eight out of nine teams demonstrated quite similar performance regarding the thoroughness criterion: they identified about 1/4 to 1/5 of the total number of the usability problems found throughout the system. Thus, it seems that novice evaluators can identify a consistent 20%-25% of the usability problems that exist in a target system.

Teams	Real usability problems	Usability problems that exist in the system	Thoroughness (%)
HE1	17	70	24,3%
HE2	17		24,3%
HE3	14		20,0%
CW1	18		25,7%
CW2	17		24,3%
T-AP1	19		27,1%
T-AP2	17		24,3%
T-AP3	14		20,0%
C-D1	29		41,4%

Table 3: Thoroughness of Usability Evaluation Methods

At a first sight this might seem a rather poor result. Certainly, a single usability evaluation, when conducted by novice evaluators is simply not enough to identify the amount of usability problems that exist in a system. However, when professional evaluators are employed the situation does not dramatically improve. Öörni, (2003) reports on six usability evaluations that were performed by professional teams and significantly the thoroughness range from 6% up to 70%. Experienced designers do not usually follow strictly a single UEM but prefer to rely on their experience by setting up their plans for usability evaluation that combine aspects of known UEMs and depend on the domain and the particular conditions of evaluation.

The last team (co-discovery learning) achieved an impressive (in comparison to the other applications of UEMs) 41.4% of usability problems identified. It also needs to be noted that the background of the designers that used this method was probably the weakest overall regarding their relationship to usability evaluation. Thus, this method seems to significantly help inexperienced teams to perform better than the other three methods. On the other hand, the fact that only one team selected this method constrains that conclusion, which can also be further pursued in other comparative usability evaluations.

4.3. Consistency

The consistency of evaluation results is considered as the extent to which the multiple usability evaluations produce similar results (Table 4). The consistency of UEMs was not satisfactory. About the half of usability problems found (50.7%) were uniquely

reported by the application of just one UEM. Furthermore, only 2 of a total of 9 teams found a consistent set of about 1/4-1/5 of the total number of usability problems (22.9%). On the contrary there was not a single usability problem that was identified by all UEMs.

Total number of usability problems	70	%
... found by 9 teams	0	0,0%
… found by 8 teams	1	1,4%
… found by 7 teams	3	4,3%
... found by 6 teams	5	7,1%
… found by 5 teams	0	0,0%
… found by 4 teams	5	7,1%
… found by 3 teams	5	7,1%
… found by 2 teams	16	22,9%
… found by 1 team	35	50,0%

Table 4: Consistency across usability evaluation teams

The comparative usability evaluation of the same object from professional teams should normally produce highly compatible results (this is implied in many usability studies such as those of Lewis (1994) and Nielsen (2000)). However this assumption has been challenged by results from more recent studies such as those of Hertzum and Jacobsen (2001) and Molich et al (2004). The evaluator effect in usability evaluations is described by Hertzum and Jacobsen (2001) as the fact that multiple evaluators evaluating the same interface with the same user evaluation method detect different sets of problems. Furthermore, Molich et al (2004) report on the results of a comparative evaluation of a single web site by nine professional usability evaluation teams: 75% (232 of 310) of usability problems identified were unique for each team that participated in the experiment, while only 2 problems were reported from six or more teams!

Hertzum and Jacobsen (2001) propose three guidelines aimed at minimising the evaluator effect:

- Be explicit on goal analysis and task selection.
- If it is important to the success of the evaluation to find most of the problems in a system, then it is strongly recommended that more than one evaluator is used.
- Reflect on evaluation procedures and problem criteria for each case.

These guidelines were applied to the greatest extent possible given the educational setting to which this study was conducted. Thus the first guideline was intentionally not strictly followed to allow for didactic trade-offs, i.e. a significant aspect of the educational process was to let students formulate some of the goals and tasks of the evaluation to gain experience; designed otherwise, the comparative study would have

probably provided more consistent and thorough results, but it would have resulted into a mechanistic process of checking a quite similar set of web pages for usability. However, the second and third guidelines were followed: each usability evaluation was a three-person team project with aggregated results; while the reflection on the evaluation procedures and problem criteria happened throughout the project continuously and at the end via interactive presentations and discussion.

5. Summary and conclusions

Although many usability evaluations are conducted with the participation of novice evaluators, knowledge about the effectiveness of the results of these studies is not readily available. The paper presented a comparative study of nine usability evaluations of the same evaluation target that were conducted by corresponding three-person teams of novice evaluators. The results show that novice evaluators can perform considerably valid and thorough usability evaluations, with some caveats.

The results of this comparative evaluation indicate a number of recommendations for usability evaluation, when novice evaluators are employed. First, it seems that there is no way to be sure that all usability problems can be identified by the use of a single group or method, with the consequence that there is a strong need to emphasise quality and clarity in usability evaluation along with completeness. This result has been also identified for the case of parallel usability evaluations that employed expert evaluators (Molich et al, 2004). Furthermore, when novice evaluators have to be employed for usability evaluations and it is important to find most usability problems then it seems that only multiple parallel usability evaluations can provide overall valid and thorough results. However, in order to review and interpret these results, the automated support of usability professionals could be of great help - this is an ongoing task for the authors. The interpretation and consolidation of the results of multiple usability evaluations is a tedious, time consuming task and its effectiveness can be considerably enhanced with (automatic) tools that can assist the participation of more than one expert.

The educational setting in which the comparative study was carried out imposed restrictions regarding the selection of evaluators (i.e. supervised teams of novice evaluators), the assignment of UEMs (i.e. only one team felt confident to carry out the usability evaluation following co-discovery learning) and the processing of results (i.e. a single expert-made final decisions about relevance and similarity of findings). On the other hand, the educational setting was convenient for reasons such as: UEMs were applied according to a common set of lecture notes; evaluators followed a common reporting format; and they followed the same tasks to evaluate the system. These conditions are hard to achieve in an industrial setting, e.g. Molich et al (2004) perform a comparative usability evaluation where the evaluator teams use different UEMs and different templates for reporting.

6. References

Hartson, H.R. Andre, T.S. and Williges, R.C. (2003) Criteria for evaluating usability evaluation methods. Int. Journal of Human-Computer Interaction, 15, 145-181.

Hertzum, M. and Jacobsen, N.E., (2001) The Evaluator Effect: A Chilling Fact about Usability Evaluation Methods, Int. Journal of Human-Computer Interaction, vol. 13, no. 4 (2001), pp. 421-443.

Lewis, C., & Wharton, C. (1997) Cognitive walkthroughs. In M. Helander, T.K. Landauer, & P. Prabhu (Eds.), Handbook of Human-Computer Interaction. Second, completely revised edition (pp. 717-732). Amsterdam: Elsevier.

Lewis, J.R. (1994) Sample sizes for usability studies: additional considerations. Human Factors 36, 369–378.

Miyake, N. (1986) Constructive interaction and the iterative process of understanding. Cognitive Science 10, 2 (1986), pp. 151–177.

Molich, R. Ede, M.R. Kaasgaard, K. and Karyukin, B. (2004) Comparative usability evaluation, Behaviour and Information Technology, Vol. 23, No. 1, 65–74.

Nielsen, J. (1992) Finding Usability Problems Through Heuristic Evaluation. Proceedings of CHI Conference on Human Factors in Computing Systems. New York: ACM, 373-380.

Nielsen, J. (1993) Usability Engineering. Academic Press, San Diego, CA.

Nielsen, J. (1994) Estimating the number of subjects needed for a thinking aloud test. Int. Journal of Human-Computer Studies 41, 385–397.

Nielsen, J., 1994. Usability Inspection Methods, CHI'94, Boston, Massachusetts, April 24-28, 1994.

Nielsen, J. (2000) Why you only need to test with 5 users. http://www.useit.com/alertbox/20000319.html

Öörni, K. (2003) What do we know about usability evaluation? – A critical view, Conference on Users in the Electronic Information Environments, September 8-9, 2003, Espoo, Finland.

Polson, P.G. Lewis, C. Rieman, J. and Wherton, C., (1992) Cognitive walkthroughs: a method for theory-based evaluation of user interfaces, Int. J. Man-Machine Studies 36, 741-773.

Rubin, J. (1994). Handbook of Usability Testing New York: John Wiley & Sons, Inc.

Sears, A. (1997). Heuristic Walkthroughs: Finding the Problems Without the Noise. International Journal of Human-Computer Interaction, 9(3), 213-234.

Van den Haak, M.J. De Jong, M.D.T. and Schellens, P.J., (2004) Employing think-aloud protocols and constructive interaction to test the usability of online library catalogues: a methodological comparison, Interacting with Computers, 16, 1153–1170.