# Evaluating Usability Evaluation Methods: Criteria, Method and a Case Study

P. Koutsabasis, T. Spyrou, and J. Darzentas

University of the Aegean
Department of Product and Systems Design Engineering
Hermoupolis, Syros, Greece, GR-84100
Tel.: +30 22810 97100, Fax: +30 22810 97109
{kgp, tsp, idarz}@aegean.gr

**Abstract.** The paper proposes an approach to comparative usability evaluation that incorporates important relevant criteria identified in previous work. It applies the proposed approach to a case study of a comparative evaluation of an academic website employing four widely-used usability evaluation methods (UEMs): heuristic evaluation, cognitive walkthroughs, think-aloud protocol and co-discovery learning.

**Keywords:** Usability evaluation methods, comparative usability evaluation, case study.

## 1 Introduction

There are many case studies of application of usability evaluation methods (UEMs). However, knowledge about particular methods has not been extensively validated comparatively. Van den Haak et al [17] note for the two widely used UEMs of think-aloud protocols and co-discovery learning that '*while there is a substantial body of literature, which describes the methods as a tool for uncovering cognitive processes, hardly any research has focused on comparing the methods as a tool for usability testing*'.

The value of comparative studies of UEMs is indisputable since that they provide a consolidated understanding based on multiple usability evaluations. However, as noted in [4] '*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*'. In addition [1] the validity and reliability of results of the various UEMs have not been studied, while and the methods themselves continue to evolve.

The paper proposes an approach to comparative usability evaluation that incorporates important relevant criteria identified in previous work ([1] [2] [4] [5] [8] [9] [13] [16] [17]); and applies this approach to a case study of a comparative evaluation of an academic website employing of four widely-used UEMs: heuristic evaluation (HE), cognitive walkthroughs (CW), think-aloud protocol (T-AP) and co-discovery learning (C-D).

J. Jacko (Ed.): Human-Computer Interaction, Part I, HCII 2007, LNCS 4550, pp. 569–578, 2007.

## 2 Related Work

A comparative usability evaluation involves multiple evaluators or evaluation teams that employ a single or multiple UEMs to carry out parallel evaluations of the same target system. There are few comparative evaluations in HCI literature.

Hertzum and Jacobsen [5] present a comparative study concerning eleven UEMs evaluations carried out with three of the four methods studied in this paper, namely CW, HE, and T-AP. Their results show that the average agreement between any two evaluators who have evaluated the same system using the same UEM ranges from 5% to 65%, and no one of the three UEMs is in general more consistent than the others. Unfortunately, Hertzum and Jacobsen could not find studies where heuristic evaluation was performed by evaluators who aggregated the results of their individual inspections to a group output (which is the case for our study). Heuristic evaluations are usually applied by a group of inspectors or users and the individual results are then aggregated [12].

Van den Haak et al (2004) make a comparison of T-AP and C-D to test the usability of online library catalogues. The UEMs were compared upon four criteria of comparison related to digital libraries: number and type of usability problems detected; relevance of the problems detected; overall task performance; and participant experiences. The study involved 80 students. The main result of their study was that the UEMs revealed similar numbers and types of problems that were equally relevant.

Molich et al [9] report on the results a comparative evaluation of a single web site (Hotmail) by nine professional teams. The goal of this study was to investigate the consistency of the results obtained. Each team was let alone to select their particular UEM and carry out the evaluation according to their work practices. The results of this evaluation are quite surprising: a large ratio (75% - 232 of 310) of usability problems identified were unique for each team that participated in the experiment, while there were only two usability problems of the target system that were reported from six or more teams.

Other comparative evaluations with different foci are presented in [3] [4] [6] and [10].

These comparative studies differ in terms of goals and the criteria used to compare evaluator performance and/or UEMs. Of particular interest for comparative evaluation work is the work of Hertzum and Jacobsen [5] who investigate the evaluator effect in usability evaluations. The term denotes the fact that multiple evaluators evaluating the same interface with the same user evaluation method detect markedly different sets of problems [6]. They [5] propose three generic guidelines to minimize 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 we strongly recommend using more than one evaluator.
- Reflect on your evaluation procedures and problem criteria.

The work presented in this paper contributes to related work by synthesising a general set of criteria from previous work into a structured approach for comparative usability evaluations. Furthermore, it presents a case study of a comparative usability evaluation that provides various insights about the UEMs employed.

# 3 A Structured Approach for Comparative Usability Evaluations: Criteria and Process

## 3.1 Criteria for Comparative Usability Evaluations

The criteria that can be taken into account for comparative usability evaluations can be distinguished by whether they refer to the evaluation target or to the UEMs themselves. An example of the first category of criteria is [17] that evaluate a web-based digital library focusing on: layout, terminology, data entry and comprehensiveness. However, criteria that are related to the target system are quite different for systems that follow different user interface paradigms. On the other hand there are also generic criteria that refer to the UEMs and not the target system. Among these (a useful review is provided by [4]), the paper identifies as most important the following:

**Realness (or relevance)** refers to whether a usability finding is a real usability problem or not (or to what degree, i.e. a severe or not important problem). According to [4] the realness of usability findings can be determined by: a) comparing with a standard usability problem list; b) expert review and judgment; and c) by end-user review and judgement. Any approach includes advantages and drawbacks regarding applicability, cost-effectiveness and trustworthiness. In this respect further research includes: severity ratings [11] and combinations of severity and probability of occurrence [15].

**Validity (or accuracy)** can be defined as the ratio of the number of real usability problems with respect to the total number of findings (i.e. real and 'false alarms') for each application of UEM [4] [16].

**Thoroughness (or completeness)** is identified in [4] and [16] as the ratio of the number of (real) usability problems found by the application of a UEM with respect to the total number of usability problems that exist in the target system. Obviously validity requires that the total number of real problems has been identified through a detailed cross-examination of results produced by all UEMs.

**Effectiveness.** The criterion of effectiveness for UEMs has been synonymous to thoroughness and validity of usability findings by most related work [2] [4] [8]; this is also in line with the definition of effectiveness by the ISO 92412 standard for usability as the '*accuracy and completeness with which users achieve specified goals*'. Thus, the effectiveness of UEMs can be identified as the product of thoroughness and validity [4]. Some related work goes even further to the definition of effectiveness by adding the issue of predictive power of UEMs in relation to the uptake of usability findings by developer teams [7] [8]. The latter perspective has to cope with additional methodological considerations not only about the persuasiveness of usability findings reporting, but also about the nature of usability findings themselves e.g. 'objective' usability problems (such as broken links in a web site) are far more likely to be addressed by development teams, rather than 'subjective' findings (such as findings related to terminology), which are the most difficult to explain in usability reporting anyway.

**Consistency** has been related to reliability [4] and repeatability [13]. In our work, we use a working definition of consistency of UEMs in terms of repeatability, as the extent to which multiple applications of different usability inspection methods produce 'reasonably similar' results. This working definition is similar to the approach of [9]. Again, the need for the identification of means for trustworthy interpretation of the similarity of usability findings is required and may be addressed by the same ways as with the realness problem.

### 3.2 Essential Process Steps for Carrying Out Comparative Usability Evaluations

The set up and carrying out of any comparative usability evaluation needs first of all to ensure that it has controlled as much as possible the aspects of the experiment that are related to the evaluator affect, thus conform with the guidelines proposed by [5]. Furthermore, the processing of results needs to ensure effective decision making about the problems of realness or relevance of the results and about the similarity of the results obtained by the parallel usability evaluations. In order to address the issues above, we propose the following guidelines for comparative usability evaluations:

**Ensure Commons when Carrying out the Usability Evaluations:** A number of issues related to the preparation and carrying out of the parallel usability evaluations need to be addressed uniformly for each evaluation. In particular:

▪ *Select evaluators that have a similar level of experience for usability evaluations*. This can be achieved by selecting professional evaluators for carrying out the experiments. When this is not possible and novice evaluators must participate, then ensure that they work in teams and that they are closely supervised. Having more than one evaluator to carry out a usability evaluation is also proposed by [5] to maximise the number of results that can be obtained; when novice evaluators are employed, then working in teams can also assist their interaction towards resolving issues about the carrying out of the usability evaluation provided they are supervised by an experienced evaluator.

▪ *Assign UEMs to evaluators according to their experience*. It is generally better to allow evaluators to select a method in which they have experience or feel most comfortable in using it.

▪ *Provide a common set of tasks to carry out*. Unless a common set of tasks is provided, there is no way to ensure that evaluators have examined the same or at least similar areas of the target system.

▪ *Provide a common format for documentation - reporting of usability findings*. As Hartson et al [4] remark "*many UEMs are designed to detect usability problems but problem reporting is left to the developers/evaluators using the UEM... problem report quality will vary greatly according to the skills of the individual reporter at communicating complete and unambiguous problem reports*." Reporting of usability problems can aid significantly the processing of results, especially for the case of parallel usability evaluations.

**Ensure effective decision making when processing the results of multiple usability evaluations:** In particular,

▪ *Select criteria for comparative usability evaluation*: As discussed in related work there are various criteria that can be considered for comparative evaluations. We make use of the criteria list presented in section 3.1, in order to draw more general conclusions about UEMs. However, aspects related to the target system affect the performance of UEMs, such as the user interface paradigm (e.g. hypertext, WIMP, 3D, etc.) and the level of maturity of the target system (e.g. application or prototype). For example, it has been argued that usability inspection methods may be more appropriate for finding problems in the early design stage of an interactive system [3]. Therefore, any conclusions drawn need to be interpreted carefully in the context of the particular class of target systems.

▪ *Select a decision criterion for relevance of usability findings*: making decisions about the realness or relevance of usability findings has been addressed in various as discussed above. We have addressed the relevance problem in a two-stage approach: first, the evaluation teams provided as part of their documentation their argumentations upon each usability finding; secondly, all usability findings were rated by an experienced usability evaluator (the first author of the paper) upon a three-scale severity scheme: 0 – not a problem; 1: minor problem; 2: serious problem.

▪ *Select a decision criterion for similarity of usability findings*: when all evaluations are available, there is a need to go through the reports in order to identify the similarity of usability findings. Again, we followed an expert-based approach for this task, which is the most usual condition in parallel usability evaluations. In this case it is however generally advisable that more than one expert performs this task. Van den Haak et al [17] have used five experts to interpret the results of their comparative study. However, there are various practical problems with involving more than a single expert. The amount of time that is needed to go through all evaluation reports, to process the large pool of data in terms of relevance and similarity and to resolve ambiguities and disagreements actually requires a lot of synchronous work. Therefore, we have used a single expert to go through the data, as well as others (e.g. [9]).

## 4 A Case Study of Comparative Usability Evaluation

### 4.1 Evaluation Object

The web site evaluated is that of the Department of Product and Systems Design Engineering (www.syros.aegean.gr), University of the Aegean and 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.

### 4.2 Participants

The usability evaluations were carried out by the MSc students of the department in terms of partial fulfilment of their obligations for the course on interaction design.

The students have a wide range of backgrounds about design having graduated from departments such as arts, graphic design, industrial engineering and information systems. Only two (out of a total of 27) students had limited experience on usability from their bachelor studies and had carried out a usability evaluation before. However, all students had considerable knowledge about the web site since they had been using it repeatedly.

Thus the selected subjects had a similar level of usability experience (novice) but a good knowledge of the target system. According to Nielsen [10] 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. Furthermore, the progress of the exercise was reviewed in weekly sessions with all teams in order to allow for resolution of queries and guide the smooth progress of the usability evaluations. Finally the fact that there was a team that carried out the evaluation instead of single novice designers encouraged critical discussion and group decision making about the findings of the usability evaluation.

### 4.3 Tasks and Methods Selected

The evaluation teams were assigned one from the four usability evaluation methods of heuristic evaluation (HE - 3 teams), cognitive walkthroughs (CW - 3 teams), think-aloud protocol (T-AP - 3 teams) and co-discovery learning (C-D 1 team) according to their degree of confidence for carrying out a usability evaluation with each one of these methods. All four methods are widely used in industry and academia for usability evaluation. The evaluation teams were provided with an analytic template for documenting the results, which included table of contents for the usability report and a predefined categorization of types of usability problems.

The evaluation teams should test the system by following two given user tasks:

- For a student, to locate information about a specific course: course description, instructor and online notes.
- For a visitor of the department, to locate necessary information about visiting the department at Hermoupolis, Syros, Greece: the map of the town, accommodation information and travel information.

The evaluation teams were given a two-month period to organise, carry out and document the usability evaluation. Their main deliverables were the usability report and their presentation of their results to an open to all discussion session.

### 4.4 Results

**Realness or relevance:** The realness of usability findings (**Table 1**) is generally high in most methods even reaching 100% in one case of HE. However, three UEMs were identified with a rather large number of false (not real) usability findings HE2, CW2 and C-D1. The fact that this variability appeared in three different UEMs leads to the conclusion that it cannot be safely related to intrinsic characteristics of methods themselves but rather to the inexperience of the evaluation teams.

**Table 1.** Realness of usability findings and severity ratings

| UEMs | Usability findings | 0: not a problem | | 1: minor problem | | 2: major problem | | Real usability problems (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 | |

**Validity:** The validity of UEMs (**Table 2**) can be directly measured out of the process of identifying the realness (or relevance) of usability findings. Baring in mind that the evaluator teams had little experience in usability evaluations, the validity of UEMs was quite satisfactory besides the three applications of methods that were discussed above.

**Table 2.** Validity of Usability Evaluation Methods

| UEMs | Usability findings | Real usability problems | Validity (%) |
|---|---|---|---|
| HE1 | 18 | 17 | 94.4% |
| HE2 | 28 | 17 | 60.7% |
| HE3 | 14 | 14 | 100.0% |
| CW1 | 21 | 18 | 85.7% |
| CW2 | 24 | 17 | 70.8% |
| T-AP1 | 21 | 19 | 90.5% |
| T-AP2 | 18 | 17 | 94.4% |
| T-AP3 | 17 | 14 | 82.4% |
| C-D1 | 39 | 29 | 74.4% |

**Thoroughness:** Thoroughness can be specified by the total number of real usability problems identified by each UEM divided by the total number of real problems that exist in the system, which is the sum of unique real problems identified by all methods. The eight out of nine UEMs demonstrated similar performance regarding the thoroughness measure (**Table 3**): they identified about 1/4 to 1/5 of the total number of the usability problems found throughout the system. The last UEM (co-discovery learning) resulted to an impressive (in comparison to the other UEMs) 41.4% of usability problems identified.

**Effectiveness:** The effectiveness of UEMs can be identified as the product of thoroughness and validity (**Table 4**). The effectiveness of UEMs has demonstrated wide ranging results:

- Five out of nine UEMs identified about 1/4-1/5 of the total number of usability problems effectively (HE1: 22.9%; HE3: 20%; CW1: 22%; T-AP1: 24.6%; T-AP2: 22.9%)

- Another three out of nine methods identified 1/6 of the total number of usability problems effectively (HE2: 14.7%, CW2: 17.2% and T-AP3: 16.5%).
- Only one UEM identified almost 1/3 the total number of usability problems effectively (C-DL1: 30.8%)

The overall results about the effectiveness of UEMs are unsatisfactory with regard to one of the central questions in usability evaluation: whether the application of a single UEM can identify a considerable amount of usability problems. This was also shown by the comparative usability evaluation work of [9] that uses professional design teams.

A second interesting result, regarding the comparison of the effectiveness of UEMs themselves is that the co-discovery learning method was significantly more effective than all other methods. Thus, it seems that this method seems to significantly help young teams to perform better than the other three methods. On the other hand, the fact that only one team selected this method constraints the safety of the conclusion, which can also be further pursued in other comparative usability evaluations.

**Table 3.** Thoroughness of Usability Evaluation Methods

| UEMs | Total number of real usability problems | Total number of 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 4.** Effectiveness of UEMs

| UEMs | Effectiveness |
|---|---|
| HE1 | 22.9% |
| HE2 | 14.7% |
| HE3 | 20.0% |
| CW1 | 22.0% |
| CW2 | 17.2% |
| T-AP1 | 24.6% |
| T-AP2 | 22.9% |
| T-AP3 | 16.5% |
| C-D1 | 30.8% |

**Consistency:** The consistency of UEMs was not satisfactory (Table 5). 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.

**Table 5.** Consistency across UEMs

| Total number of usability problems | 70 | % |
|---|---|---|
| ... found by 9 teams / UEM | 0 | 0.0% |
| … found by 8 teams / UEM | 1 | 1.4% |
| … found by 7 teams / UEM | 3 | 4.3% |
| ... found by 6 teams / UEM | 5 | 7.1% |
| … found by 5 teams / UEM | 0 | 0.0% |
| … found by 4 teams / UEM | 5 | 7.1% |
| … found by 3 teams / UEM | 5 | 7.1% |
| … found by 2 teams / UEM | 16 | 22.9% |
| … found by 1 team / UEM | 35 | 50.0% |

### 4.5 Discussion

The main conclusions that stem out of the case study are that:

- The employment of a single method is not enough for comprehensive usability evaluation. If it is important to find most problems, parallel evaluations can be carried out.
- No method was found to be significantly more effective or consistent than others.
- The realness and validity of evaluation results was considerably high for most teams, which counts for young designers' supervised participation to usability evaluations.

In the case study presented, we have followed the proposed approach to inform current practice regarding the use of UEMs. The educational setting in which the case 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. an expert-based approach was followed to make final decisions about the relevance and similarity of the usability findings). On the other hand, the educational setting was convenient for a number of other reasons including that: UEMs were applied according to a common set of lecture notes; evaluators followed a common format for reporting; and they followed the same tasks to evaluate the system. These conditions are hard to achieve in an industrial setting. For example, Molich et al [9] perform a comparative usability evaluation where the evaluator teams use different UEMs (actually combinations of UEMs that have evolved by practice) and different templates for reporting.

## 5 Summary and Conclusions

Comparative usability evaluations are important for the throrough identification of usability problems and the comparison of UEMs in particular contexts. The paper contributes to the understanding of criteria for comparative usability evaluation both in terms of providing a method for this task and by presenting a relevant case study

for a web-based system. It is envisaged that the approach taken can be applied to other comparative studies as well. Also the results of the case study can inform the selection of UEMs particularly when young designers need to be employed in comparative usability evaluations.

## References

1. Andre, T.S, Hartson, H.R., Belzand, S.M., McCreary, F.A.: The user action framework: a reliable foundation for usability engineering support tools. Int. J. Human-Computer Studies 54, 107–136 (2001)
2. Cockton, G., Woolrych, A.: Understanding inspection methods. In: Blandford, A., Vanderdonckt, J., Gray, P.D. (eds.) People and Computer, vol. XV, pp. 171–192. Springer, Heidelberg (2001)
3. Doubleday, A., Ryan, M., Springett, M., Sutcliffe, A.: A comparison of usability techniques for evaluating design. In: Proceedings of Designing interactive systems (1997)
4. Hartson, H.R., Andre, T.S., Williges, R.C.: Criteria for evaluating usability evaluation methods. International Journal of Human-Computer Interaction 15, 145–181 (2003)
5. Hertzum, M., Jacobsen, N.E.: The Evaluator Effect: A Chilling Fact about Usability Evaluation Methods. International Journal of Human-Computer Interaction 13(4), 421–443 (2001)
6. Jacobsen, N.E., Hertzum, M., John, B.E.: The evaluator effect in usability tests. In: Summary Proceedings of the ACM CHI 98 Conference, pp. 255–256. ACM Press, New York (1998)
7. John, B.E., Marks, S.J.: Tracking the effectiveness of usability evaluation methods. Behaviour and Information Technology, 16(4/5), 188–202 (1997)
8. Law, E.L-C., Hvannberg, E.T.: Analysis of strategies for estimating and improving the effectiveness of heuristic evaluation. In: Proceedings of NordiCHI 2004, Tampere, Finland (October 23-27, 2004)
9. Molich, R., Ede, M.R., Kaasgaard, K., Karyukin, B.: Comparative usability evaluation. Behaviour and Information Technology 23(1), 65–74 (2004)
10. Nielsen, J.: Finding Usability Problems Through Heuristic Evaluation. In: Proceedings of CHI Conference on Human Factors in Computing Systems, pp. 373–380. ACM, New York (1992)
11. Nielsen, J.: Usability Engineering. Academic Press, San Diego (1993)
12. Nielsen, J.: Usability Inspection Methods. In: CHI'94, Boston, Massachusetts (1994)
13. Öörni, K.: What do we know about usability evaluation? - A critical view, In: Conference on Users in the Electronic Information Environments, September 8 - 9, 2003 Espoo, Finland (2003)
14. Rosson, M.B., Caroll, J.M.: Usability Engineering: Scenario-Based Development of Human-Computer Interaction. Morgan-Kaufmann, San Francisco (2002)
15. Rubin, J.: Handbook of Usability Testing. John Wiley & Sons, Inc. New York (1994)
16. Sears, A.: Heuristic Walkthroughs: Finding the Problems Without the Noise. International Journal of Human-Computer Interaction 9(3), 213–234 (1997)
17. Van den Haak, M.J., De Jong, M.D.T., Schellens, P.J.: Employing think-aloud protocols and constructive interaction to test the usability of online library catalogues: a methodological comparison. Interacting with Computers 16, 1153–1170 (2004)