

Guiding Usability Evaluators During Hypermedia Inspection

M.F. Costabile*, A. De Angeli⁺, M. Matera[°]

**Dipartimento di Informatica, Università di Bari, Bari, Italy*
+NCR FSG, SSSS Advanced Technology & Research, Dundee, UK

°Dipartimento di Elettronica e Informazione, Politecnico di Milano, Milano, Italy
costabile@di.uniba.it, antonella.de_angeli@ncr.com, matera@elet.polimi.it

Abstract

This paper presents an empirical validation of the Abstract Tasks (ATs), which are operational guidelines driving the inspection activities during a usability evaluation. Two groups of inspectors evaluated a multimedia CD-ROM, one group using the ATs, the other group using a list of heuristics. Results demonstrated a better performance of the AT group but raised some issues concerning the acceptability of the technique.

1. Introduction

Usability inspections are emerging as promising solutions for cost-effective evaluations [1]. However, their success still depends on skills and experience of the inspectors. Further, current methods are too general, focusing on surface features of the graphical interface. To address the usability of hypermedia systems, taking also into account organisation of the information elements and functionality, we have developed SUE, a methodology whose aim is helping evaluators share and transfer their expertise. The very novelty of SUE is the proposal of an inspection technique that exploits operational guidelines, called Abstract Tasks (ATs), to drive the evaluation activities. ATs help evaluators concentrate on specific features of hypermedia applications without however neglecting the surface aspects. As originally defined [2], SUE required evaluators to be familiar with a Hypermedia Design Model – HDM. This was used to represent schematically the application, so that to highlight the main components corresponding to the objects that are worth analysing.

The value of the SUE inspection has already been empirically demonstrated [2]. A controlled experiment (in the following Exp_1) showed that, compared to a standard heuristic approach, SUE enhances effectiveness and efficiency of the evaluation, as well as the satisfaction of the evaluator. However, a post-experiment questionnaire suggested that the HDM model was difficult to understand. Hence, we realised that HDM could be both the strength and the weakness of our approach. Advantages relate to the availability of a systematic method for identifying evaluation objects; drawbacks relate to the complexity of the model.

Applying the same experimental design, we have run a new study aimed at investigating the value of the ATs as a stand-alone technique (Exp_2).

2. Experimental Method

Twenty students from the University of Bari participated in the experiment as part of their credit for an HCI course. They were familiar with heuristic evaluations but not with the HDM model. Half of the sample was randomly assigned to the Heuristic Inspection (HI) condition, the other half to the Abstract Task (AT) condition. A few days before the experiment, participants were introduced to the conceptual tools to be used in the inspection. A 2-hour seminar presented to both groups the hypermedia-specific heuristics proposed in SUE. Then, a 15-minute demo of the application to be evaluated was given. Finally, half an hour seminar introduced participants of the AT group to the key concepts of the AT-based technique. The experiment lasted 3 hours. Working individually, participants had to inspect a commercial CD-ROM [3] applying the technique they were assigned to and recording the usability problems on a booklet. All participants received the SUE heuristics; the AT group also received ten ATs chosen among the forty defined in [4]. At the end, all participants were invited to fill in a satisfaction questionnaire.

3. Results

The study aimed to assess effectiveness, efficiency and evaluator satisfaction. The statements reported in the booklets were scored as *Problems* (actual usability flaws) or *Non-Problems* (personal preferences, misjudgments, not understandable statements). A severity scoring, modulated on a 5 points Likert scale, was performed on all the problems.

Effectiveness took into account *completeness* and *accuracy* of the evaluation. Completeness corresponded to the percentage of problems detected by a single inspector out of the total number of problems present in the application (N = 38). Accuracy referred to the quality of individual reports. It was further broken down into *precision* (percentage of problems detected by an individual inspector out of the total number of statements she reported) and *severity* (average of 3 independent

ratings). As regards completeness, Exp_2 entirely confirmed the advantage evinced in Exp_1. On the average, inspectors assigned to the AT condition found 23% of all the usability problems, whilst inspectors in the HI condition only 16%. A Mann-Whitney U test demonstrated that the difference is significant ($U = 16.0$ ($N = 20$), $p < .01$). Moreover, a cross experiment comparison showed no difference in the performance of participants applying the complete SUE Inspection and those using only ATs ($U = 62.5$ ($N = 24$)). It follows that the mere use of the ATs increased the completeness of the evaluation.

The analysis of the two dimensions defining accuracy gave raises to a more complex framework. In both cases, the advantage evinced in Exp_1 was not confirmed. As expected, the distribution of precision is affected by a severe skewness, since most evaluators performed well. The variable ranges from 63 to 100 with a median value of 87.5. The AT group scored an average of 89%, the HI group an average of 80%. Despite being in the expected direction, the difference is not significant ($U = 30.0$ ($N = 20$), $p = .14$). The analysis of the severity index evinced a similar trend. Problems detected in the AT condition were slightly more serious than those detected in the HI condition (mean difference = 0.33). A t-test demonstrated that the expected advantage of the AT over the HI condition was just a tendency ($t_{(18)} = -1.56$, $p = .14$). These results should be carefully weighted considering the lower number of participants tested in Exp_2 as compared to Exp_1. Moreover, we must consider that the ATs were tested in non-optimal conditions. They were not fully adapted to their new stand-alone function, since their linguistic formulation still relied on terms and concepts of HDM. Nevertheless, we cannot exclude that the accuracy of SUE heavily relies on the use of the HDM, which could be instrumental in precisely identifying the application constituents to be evaluated.

Efficiency was assessed by the Nielsen's cost-benefit curve. It relates the proportion of usability problems to the number of evaluators, according to the following formula $Found(i) = n(1-(1-\lambda)^i)$, where $Found(i)$ is the number of problems found by aggregating reports from i independent evaluators, n is the total number of problems in the application, and λ is the probability of finding the average usability problem when using a single evaluator. Figure 1 shows that the AT condition tended to reach better performance with a lowest number of evaluators. Assuming the Nielsen's 75% threshold, AT reached it with five evaluators, while HI needed eight evaluators. The advantage evinced from Exp_1 has been confirmed.

The dimension regarding the *evaluator-satisfaction* gave rise to the most unexpected result. Indeed, a t-test evinced a significant difference ($t_{(17)} = .231$, $p < .05$) favoring the HI group. Evaluators who applied the traditional heuristic method were more satisfied than those who applied the ATs (mean difference = 0.5). The effect was probably due to the HDM terminology that inspectors in the AT condition had to face.

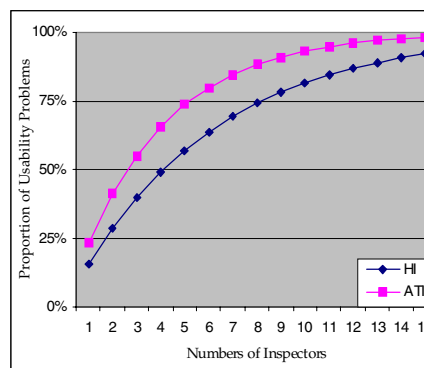


Figure 2. Nielsen's cost-benefit curve for the experimental groups ($n = 38$, $\lambda_{HI} = 0.15$, $\lambda_{SI} = 0.23$).

4. Conclusions

Two validation experiments have so far demonstrated that ATs help hypermedia inspectors to be effective and efficient. In particular, the final experiment has demonstrated that the evaluation performance is not seriously compromised if the HDM model is not used by evaluators, as originally prescribed by SUE. This result extends the portability of the method, which can be easily employed without an extensive training on the model. We believe that the problems related to acceptability and accuracy of the AT technique can be solved by a revision of the linguistic formulation of the ATs. Indeed, in their current form they still relies on the HDM terminology, thus increasing the demand on the inspectors. We are very confident on the validity of our results, since by no means the evaluators were influenced by our attachment to the ATs. We are planning further experiments involving expert evaluators, in order to prove whether ATs provide greater power to them as well. In addition, we are reformulating ATs to make explicit HDM concepts also to people who have never been exposed to the model.

Acknowledgements

The support of grants MURST Cofin 2000 and EU FAIRWIS project IST-1999-12641 is acknowledged.

References

- [1] Nielsen, J., and Mack, R.L. (1994), *Usability Inspection Methods*, John Wiley & Sons, New York.
- [2] De Angeli A., Matera M., Costabile M.F., Garzotto F., Paolini P., "Validating the SUE Inspection Technique", *Proc. AVI 2000*, Palermo, Italy, May 24-26 2000, ACM Press, pp. 143-150.
- [3] Mondadori New Media. (1997), *Camminare nella Pittura*, CD-ROM.
- [4] Matera, M. (1999), "SUE: A Systematic Methodology for Evaluating Hypermedia Usabilità", Ph.D. Thesis, Dipartimento di Elettronica e Informazione, Politecnico di Milano.