A Two-Phase Model for Usability Evaluation of Software User Interfaces

Chee-Hwan Lim [*] and Kyung S. Park [**]

[*] Department of Management Information Systems, Seowon University

[**] Department of Industrial Engineering, Korea Advanced Institute of Science and Technology

ABSTRACT

There is currently a focus on usability of interactive computer software. Previous research in software ergonomics has indicated the importance of evaluating the usability of software user interfaces. Software developers, interface designers or human factors engineers often confront the task of comparative evaluation among systems, versions or interface designs. This study presents a structured model for comparative evaluation of user interface designs using usability criteria and measures. The proposed model consists of two main phases: the prescreening phase and the evaluation phase. The first phase involves expert judgment-based approach with qualitative criteria. The prescreening phase uses absolute measurement analytic hierarchy process to filter possible alternative interfaces to a reasonable subset. The second phase involves user-based approach such as usability testing, with quantitative criteria. The objective of the evaluation phase is to evaluate a subset of alternatives using objective measures. A set of criteria and measures for evaluating the usability of computer software designs is presented. The proposed model provides practitioners with a structured approach to select the best interface based on usability criteria and measures.

1. Introduction

The reasons for evaluating software user interfaces can be divided into some broad classes. These include assisting in design decisions, and measuring quality of use. In short, the distinction is between whether or not one is looking for new information about possible alternatives (e. g., user interface designs), or looking to report on the value of some measure for comparative purposes. In some situations, one is interested in evaluation to select between two or more choices. For example, software developers or human factors engineers often confront the task of comparative evaluation among systems, versions or interface designs. In this type of evaluation multiple criteria are used because it is difficult to derive a single measure that effectively characterizes the overall usability of an interactive system.

Some authors have studied computer interfaces evaluation in terms of an integrated assessment of several interface characteristics. Mitta (1993) has suggested using the analytic hierarchy process (AHP) to rank-order computer interfaces based on multiple evaluative criteria. Stanney and Mollaghasemi (1995) also used the AHP to assess the relative importance of a realistic and an unrealistic desktop interface design. These approaches are able to allow simultaneous consideration of multiple criteria and to readily quantify consistency in the decision maker's judgments.

Mitta's approach, however, was based solely on subjective assessments. With this approach experimenters evaluated users with respect to their abilities to make satisfactory judgments regarding three criteria. Stanney and Mollaghasemi (1995) pointed out that the repeatability of this assessment procedure is questionable, and evaluated the interface attributes in terms of objective measures. The objective of this paper is to extend the earlier research, and to describe a structured approach for evaluating the usability of user interfaces. The decision framework as proposed in this paper, is made up of two phases, namely the prescreening phase (expert judgment-based approach) and the evaluation phase (user-based approach).

2. The first phase

The first phase is mainly concerned with qualitative and subjective assessment. The objective of this phase is to filter possible alternatives to a reasonable subset. In this phase, experts mostly rely on their experience to make a judgment on the ergonomic quality of alternative interfaces. Lacking experience, they appraise alternatives with user interface design guidelines (Smith & Mosier, 1986; Shneiderman, 1992; Nielsen, 1993), established human factors principles and standards (e. g., ISO, ANSI, DIN, etc.), and criteria (Ravden & Johnson, 1989; Scapin, 1990). This assessment also requires clear criteria against which to assess the quality of the user interfaces. Criteria should be agreed upon and identified before evaluation efforts begin. As a result of the review and the comparison, eight criteria have been identified. If all criteria might be considered for alternatives, all criteria are not equally important: they should be weighted according to their relative importance. Experts (i. e., members of the evaluation team) understand all alternatives under consideration and then evaluate them in terms of the usability criteria.

The proposed approach in the first phase uses absolute measurement AHP. Absolute measurement AHP requires a pairwise comparison procedure between indicator categories (for each lowest level criterion) to establish the relative weights for these categories using eigenvector approach. The objective of the first phase is to discard certain inferior alternatives and to reduce the number of alternatives under consideration. This gives us economical efficiency of analysis. After all, the high ranked (leading) alternatives are selected at the end of the prescreening phase. The results obtained from the prescreening phase are taken into the evaluation phase, which aims to evaluate the alternatives using objective measures.

3. The second phase

The objective of the second phase is to evaluate a subset of alternatives using quantitative criteria and to select the best alternative. The evaluation phase involves user-based assessment such as user testing, with quantitative criteria and measures. Since usability is a multi-dimensional concept that cannot be characterized by a single criterion, multiple measures are used in usability assessment. A wide variety of quantifiable measures may be used, based on such factors as the specific interface to be tested, laboratory or field conditions, available test equipment, or aspects of the product (Nielsen, 1993; Whiteside et al., 1988).

ISO 9241-Part 11 (Guidance on usability) gives the following definition of usability (Bevan, 1995): "usability is measured by the extent to which the intended goals of use of the overall system are achieved (effectiveness); the resources such as time, money, mental effort that have to be expended to achieve the intended goals (efficiency); and the extent to which the user finds the overall system acceptable (satisfaction)". Effectiveness, efficiency and satisfaction can be seen as critical criteria that influence usability of interfaces. To evaluate these criteria, they need to be decomposed into sub-criteria, and finally, into usability measures. According to characteristics of usability measures, thus, each criterion (e. g., effectiveness, efficiency, satisfaction) can be decomposed into sub-criteria (or usability measures).

The relative importance of components of usability also depends on the context of use and the purposes for which usability is being described. Effectiveness and efficiency are usually a prime concern, but satisfaction may be even more important, for instance where usage is discretionary. A method is needed that integrates user testing results, thus providing a composite measure of usability to facilitate direct comparisons between interface design options. We also use the AHP to assess the relative importance for each set of usability components (e. g., criteria and sub-criteria) with the objective of selecting the best interface. The priorities (i. e., relative importance weights) of usability components are based on experts' judgments because users have not enough sense to judge the relative importance of them. The priorities of alternatives with respect to each of measures (sub-criteria) are based on user testing data. It is necessary to unite the first phase and the second phase outcomes (i. e., the priorities of each alternative) for more reliable analysis. The alternative with the highest overall rating is ranked the best choice, taking into account user testing data as well as experts' assessments.

4. An Example

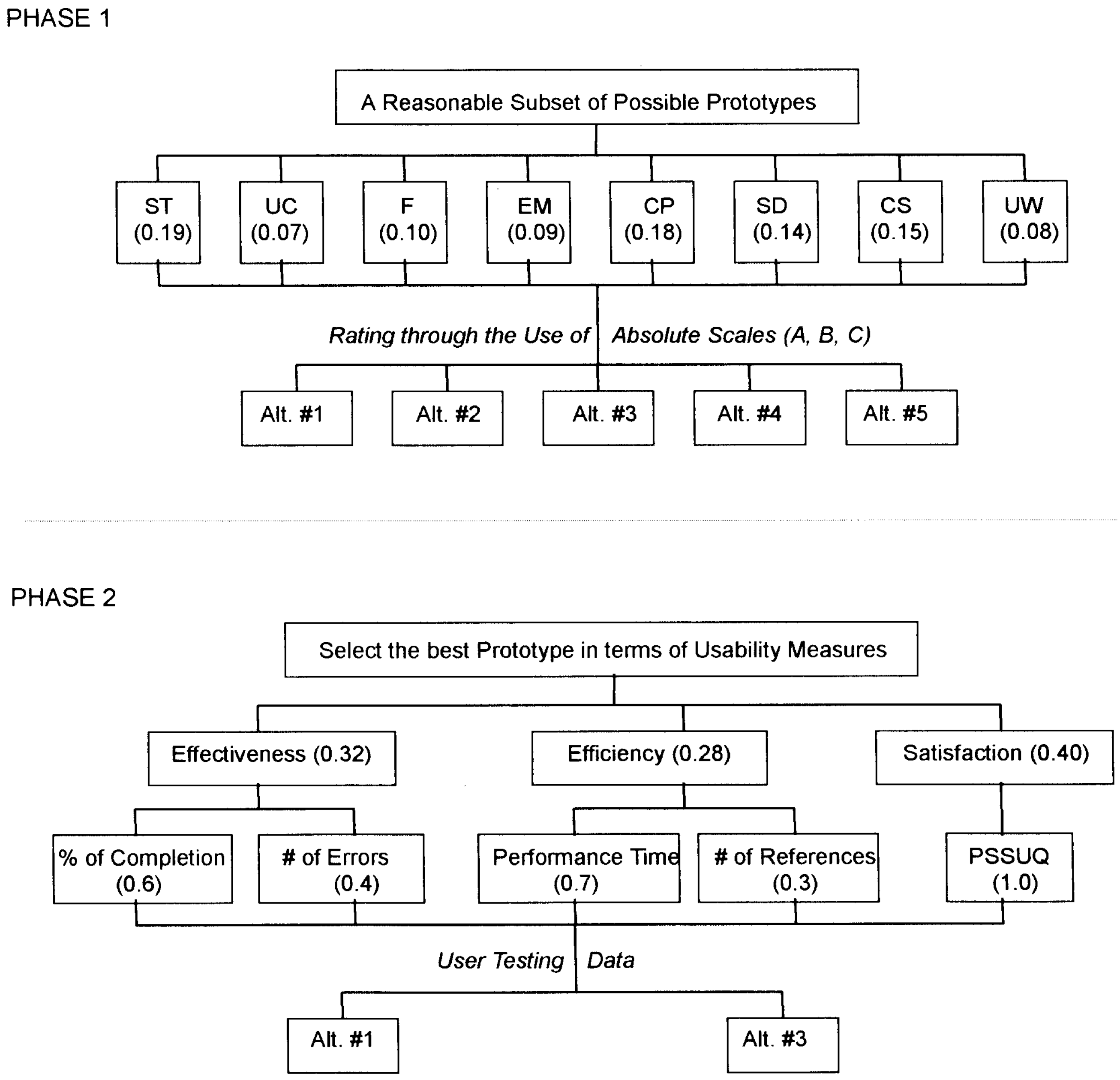
An application of the proposed approach involved a comparative assessment of several interface prototypes of an interactive system. For demonstration, five interface prototypes were designed to support a database system. This system running on an IBM-PC provides information obtained from the analysis of nuclear power plant trips, such as component failures or human errors that induced the trips, the sequence of unit events, and problems that contributed to the trips. The five alternatives, prototypes written in Multimedia ToolBook ^{TM} for Windows, incorporated several types of user interface technology, with windows, screen layouts, system command modes (e. g., menu, button, and icon), colors, and information feedback. Two hierarchical trees for selecting the best alternative are shown in Figure 1.

The first phase started with five prototype alternatives. The lowest level of the hierarchy represents the indicator categories comprising the scales for the criteria. For simplicity, three grade levels such as A, B, and C were provided for each criterion scale. The alternatives were rated through the use of absolute scales where grades have been assigned to alternatives according to how they fulfill the criteria. Consensus was reached on all ratings. Table 1 provides an overview of criteria weights and alternative weights resulting in a rating of the proposed altematives. The prototype altemative 3, with 24% relative priority, was favored. The analysis outcome in the first phase, however, indicates reasonably close outcomes especially between the alternative 3 and the alternative 1. These two alternatives were taken into the evaluation phase, which aims to evaluate them using objective measures.

The two alternatives were assessed based on the evaluation criteria. This analysis used the percentage of successfully completing tasks and the number of errors as measures of effectiveness. Efficiency was quantified in terms of performance time and the number of references to help. Satisfaction was obtained by the Post-Study System Usability Questionnaire (PSSUQ) that is a kind of IBM questionnaires (Lewis, 1995). The mean performance times for the subjects were calculated to be 913 seconds and 876 seconds for the altemative 1 and the altemative 3, respectively. The mean number of tasks successfully completed for the subjects were determined to be 92 % and 96 % for the altemative 1 and the altemative 3, respectively. The summarized results are shown in Table 2. The weights of alternatives and weights for each set of criteria and sub-criteria were synthesized into a composite score for each alternative. The results from the two-phase model showed that the overall usability of the altemative 3 was better than the alternative 1.

5. Conclusions

This study presents a structured methodology for comparative evaluation of user interface designs using usability criteria and measures. Usability evaluation usually has been conducted by


PHASE 1

PHASE 2

Figure 1. Hierarchical trees in the two-phase model

Table 1. Overview of criterion and alternative weights in the first phase

Criteria	Weights	Alt. #1	#2	#3	#4	#5
Suitability for the Task (ST)	0.19	0.25	0.48	0.34	0.23	0.31
User Control (UC)	0.07	0.56	0.29	0.36	0.48	0.23
Flexibility (F)	0.10	0.48	0.31	0.56	0.16	0.48
Error Management (EM)	0.09	0.23	0.16	0.16	0.24	0.29
Compatibility (CP)	0.18	0.29	0.23	0.23	0.25	0.16
Self-descriptiveness (SD)	0.14	0.16	0.13	0.25	0.23	0.20
Consistency (CS)	0.15	0.48	0.25	0.56	0.29	0.23
User Workload (UW)	0.08	0.31	0.29	0.23	0.21	0.25
Rating	1.00	0.23 *	0.19	0.24 *	0.16	0.18

Table 2. Overview of measure and alternative weights in the second phase

Criteria and Sub-criteria	Weights	Alt. #1		Alt. #3	
		Actual Measure	Rel. Wts.	Actual measure	Rel. Wts.
Effectiveness	0.32				
% of Completion	0.6 [0.192]	92 (%)	0.49	96 (%)	0.51
# of Errors	0.4 [0.128]	4.2	0.46	3.6	0.54
Efficiency	0.28				
Performance Time	0.7 [0.196]	913 (sec)	0.49	876 (sec)	0.51
# of References	0.3 [0.084]	2.4	0.54	2.8	0.46
Satisfaction	0.40				
PSSUQ	1.0 [0.40]	3.81(pts.)	0.45	3.17 (pts.)	0.55
Rating			0.47		0.53

human factors experts or by using user testing. Because each of two approaches has its advantage, we combine two approaches to evaluate user interfaces effectively. The proposed model consists of two phases: the prescreening phase (expert judgment-based approach) and the evaluation phase (user-based approach). It is particularly useful when multiple criteria and several alternative interfaces are considered and when usability evaluation must be performed under limited resources.

References

Bevan, N., 1995. Human-computer interaction standards. Proceedings of the 6th International Conference on Human-Computer Interaction, Yokohama, Japan, Elsevier, pp. 885-890.

Lewis, J. R., 1995. IBM usability satisfaction questionnaires: psychometric evaluation and instructions for use. International Journal of Human-Computer Interaction, 7(1): 57-78.

Mitta, D. A., 1993. An application of the analytical hierarchy process: a rank-ordering of computer interfaces. Human Factors, 35(1): 141-157.

Nielsen, J., 1993. Usability Engineering. Academic Press, London.

Ravden, S. J. and Johnson, G. I., 1989. Evaluating Usability of Human-Computer Interfaces. Ellis Horwood, Chichester.

Scapin, D. L., 1990. Organizing human hactors knowledge for the evaluation and design of interfaces. International Journal of Human-Computer Interaction, 2(3): 203-229.

Shneiderman, B., 1992. Designing the User Interface. Addison-Wesley, pp. 471-500.

Smith, S. L. and Mosier, J. N., 1986. Guidelines for designing user interface software. Report No. MTR-10090, Esd-TR-86-278, The Mitree Co., Beford, MA.

Stanney, K. and Mollaghasemi, M., 1995. A composite measure of usability for human-computer interface designs. Proceedings of the 6th International Conference on Human-Computer Interaction, Yokohama, Japan, Elsevier, pp. 387-392.

Whiteside, J., Bennett, J. and Holzblatt, K., 1988. Usability engineering: our experience and evolution. In: M. Helander (Eds.), Handbook of Human-Computer Interaction. Elsevier, pp. 791-817.