For Usability Quantification

IKEGAMI Teruya, OKADA Hidehiko

Abstract

The authors of this paper propose a method of quantitative evaluation of the ease of use (usability) of systems by using a checklist evaluation method that guides detailed procedures and thresholds and which is improving its accuracy by verification tests. In the tests, the authors conducted experimental evaluations with the participation of several persons and have confirmed that high reproducibility can be achieved by determining the results based on deliberations. This paper is intended to introduce the proposed checklist evaluation method, details of its tests and its future perspectives.

Keywords

usability, evaluation technique, quantification, checklist, AHP

1. Introduction

A variety of evaluation techniques are practiced in order to shape and improve the usability of systems. One of the representative usability evaluation techniques [1], the checklist evaluation method [2] is known to have the advantage of being suitable for application at the upstream stage of development. However, this method also has some problems, such as the dependency of the evaluation results on the skill, experience and subjectivity of the evaluator and the difficulty of obtaining consistent, reproducible evaluation results.

Aiming at the quantification of usability based on the checklist evaluation method, we are currently developing a checklist that can eliminate as much as possible the fluctuations of results that depend on the evaluator. In the following, we will describe the proposed method, its verification testing and finally its future perspectives.

2. Checklist Construction

Before development of quantification using a checklist, we arranged the problems linked to quantification and decided on the course of our efforts.

2.1 Issues of Quantification

(1) Eliminating the Fluctuation of Results between Evaluators

The checklist evaluation method generally checks the targets with respect to each evaluation item and scores the degree of conformity. If the conformity is assessed on a one-to-five scale, for example, the results tend to fluctuate at the discretion of evaluators. Some evaluators may even be unable to understand the meanings of some items or assess them erroneously. An earlier study [3] attempted to solve this problem by letting the evaluators learn and experience the usability-related matters to a certain degree so that they could patch up any inadequacies in knowledge and experience before applying the checklist. Nevertheless, it is difficult for all the evaluators to have the same degree of knowledge. In addition, the names of the UI components such as the list box and pull-down menu tend to vary, even between skilled persons. As a result, the minimization of the fluctuations in results between evaluators remains an important issue.

We described the targets and procedures of each evaluation item in detail and prescribed the judgment criteria so that conformity could be judged as "Approved (no problem)," "Unapproved (problem)" or "Inapplicable (no evaluation target)." We also put together a glossary of the checklist-related terminology and case examples in order to reduce fluctuations in understanding and interpretation between the evaluators.

(2) Presenting the Effects in an Easy-to-Understand Manner for Users

In general it is assumed that the checklist is to be used by experts or developers of UI design. As a result the evaluation items are often composed of elements related directly to design and development, such as layouts and buttons and the effects are often difficult for the user to understand. In addition, as the content and degree of the effects exerted on the

user in fulfilling requirements vary according to specific checklist items, it is important to assign valid weightings to them.

We adopted weightings by using AHP (Analytic Hierarchy Process) [4] and decided to output the evaluation results from four viewpoints. These were; "learnability (easy to learn)," "errors (few errors)," "memorability (easy to remember)," and "efficiency (efficient to use)."

2.2 Arrangement of the Checklist

Based on several guidelines, standards and expertise obtained through actual operations, we built a checklist composed of 126 items arranged in 5 sections (**Table 1**).

Although detailed evaluation procedure is specified for each item in the checklist in order to prevent fluctuations in the results that depend on the evaluator, the evaluation of some items will still require specific knowledge of the targets. Accordingly, the checklist items are classified into "basic items" that can be evaluated with a certain degree of correctness by anyone observing the defined procedures and the "extended items" that require expert knowledge of the work in order to answer them correctly. For example, in the case of one of the basic items called "contrast of color scheme," a procedure for confirming sufficient contrast between the text and the background color is described together with the judgment formula. This is so that the same answer can be obtained from all, regardless of the degree of knowledge of the work. On the other hand, the item on "emphasized information expression" is an extended item because it requires a particular knowledge of the work in order for the judgment on the important information to be emphasized.

Table 1 Checklist configuration (89 basic items and 37 extended items)

1. Consistency of display/operation
If the display and operating procedures are consistent throughout the system. Items related to the visual effects, layout/screen transition, data output, operation, and response/notification.
2. Ease of viewing and distinction of information
If information is easily visible, different information is easily distinguishable. Items related to the visual effects, layout/screen transition, data output and operation.
3. Presentation of current status
If information is displayed according to the present display (work) or operation status. Items related to the data output, operation and response/notification.
4. Conformity to users/environment
If flexible adaptation to the characteristics of the user and environment is possible. Items related to operation, response/notification and customization.
5. Conformity to work
If information and means of operation matching the user's work are presented. Items related to the layout/screen transition, data output and operation.

2.3 Weighting of Items

We adopted AHP to determine the weighting of checklist items as well as to calculate the results of evaluation of the items configured from the viewpoints of design and development as well as from the effects on the user. When used in decision-making, AHP calculates the weighting of each involved element of the entire checklist of items by analyzing the hierarchical structure of associated elements and by assigning weighting of the numerical values at each hierarchical level. This process features a paired comparison of evaluation targets with respect to specific criteria (including those that are hard to measure directly). The weighting calculated by this procedure is more valid than the weighting of elements determined by a more general method.

Among the five usability attributes proposed by Nielsen [1] , we selected four of them, which were "learnability (easy to learn)," "errors (few errors)," "memorability (easy to remember)," and "efficiency (efficient to use)" as the judgment criteria and determined the weighting of each checklist item from

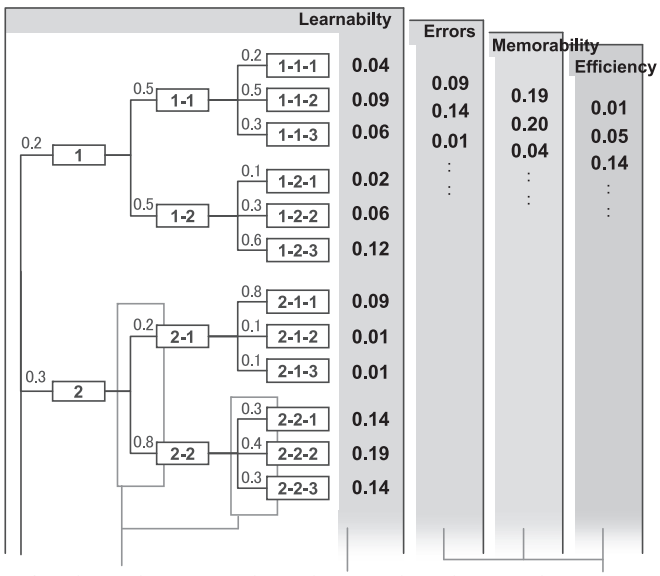

Fig. Checklist weighting.

each viewpoint (**Fig.**). We excluded "satisfaction" because this is a subjective attribute that is determined based on the integration of various factors, including the richness of functions in addition to the four other attributes and varies greatly depending on the personal taste and sense of value of the user. This procedure makes the checklist suitable for calculating the evaluation results based on the four viewpoints described above. It also makes the improvement process efficient by allowing items with a higher weighting to be improved as a priority from a viewpoint that stresses the conceptual framework of a specific product.

3. Checklist Evaluation Tests

We conducted evaluation tests using the checklist under development on multiple evaluators, verified the consistency of the results and extracted issues requiring improvement.

3.1 Test Method

The first test was made in September 2007 and a total of five tests were conducted by December. Each test was performed according to the method described below.

(1) Evaluated Check Items

We first selected four or five items from the basic items that do not require knowledge of work or skill. We have conducted 5 test sessions and 19 items (22 items in total) were evaluated.

(2) Evaluators

Each test was conducted with 5 or 6 evaluators. 3 or 4 of the evaluators were university students, who were beginners without a prior experience of checklist evaluation, who had little knowledge of usability. On the other hand, the others were NEC researchers, who were skilled practitioners with knowledge and experience of the target evaluations. Comparison of the results reported from the two groups of evaluators would make it possible for us to verify if similar results could be obtained, regardless of them being beginners or skilled persons. The total number of the skilled persons that participated in the 5 test sessions was 3 and that of the beginners participating in them was 10.

(3) Evaluation Targets

We selected 4 to 5 windows from the windows displayed by the GUI (Graphical User Interface) of certain e-mail software, and set these windows as the population of the evaluation targets.

(4) Evaluation Method

For each checklist item, we compiled an instruction document describing the evaluation target windows (all or part of the windows in the population), types of evaluated UI components (menu, button, text entry, etc.), method/criteria of judgment, examples of nonconformity to checklist, etc. and delivered them to the evaluators. Each evaluator independently (under prohibition of cooperation or information exchange with other evaluators) operated the specified e-mail software and evaluated it by following the indications in the instruction document.

With checklist items that can be evaluated on a per-window basis such as the visibility of information and the validity of UI component use, we had the evaluators check problems and report the results on a per-window basis.

With checklist items that should be evaluated as a whole by comparing multiple windows such as the consistencies of display and operation, the evaluators evaluated and reported the results on the set of windows.

3.2 Test Results

Since the results of evaluation using the checklist is one of the "Acceptable," "Unacceptable" and "Inapplicable" options as described above, comparing the results of beginners with those of skilled persons makes it possible to perceive if the results are consistent between beginners and skilled persons and if consistent results can be obtained from all of the multiple evaluators. For example, if the results are as shown in **Table 2** (a) (which were obtained from the four beginners 1 to 4 and two skilled persons 1 and 2), the results of the 4 beginners match completely those of the skilled persons, indicating that these checklist items are effective for introducing consistent results. On the other hand, if the results are as shown in Table 2(b), 1 of the 4 beginners reported a result different from the skilled persons, this indicated that this checklist item was problematic for obtaining consistent results.

We defined the match rate as the index representing how far

Table 2 Evaluation result examples.

	Beginner				Skilled	
	1	2	3	4	1	2
Acceptable	●	●	●	●	●	●
Unacceptable						
Inapplicable						

(a)

	Beginner				Skilled	
	1	2	3	4	1	2
Acceptable		●				
Unacceptable	●		●	●	●	●
Inapplicable						

(b)

Table 3 Average match rate (Per test count).

1st	2	3	4	5
62.50%	77.60%	72.00%	61.90%	66.70%

the results of beginners match those of skilled persons. With the results of Table 2(a) and (b), for example, the match rates are 100% and 75% respectively. We also obtained the average match rates of the 5 test sessions and obtained results as shown in **Table 3** .

$$\text{Match rate} = 100 * \frac{\text{Number of beginners matching skilled persons}}{\text{Number of beginners}} (\%)$$

4. Considerations

The results of the tests are fed back to support checklist development in improving the accuracy of the checklist. In addition, we also extracted the points to be considered and are using them in advancing studies of appropriate measures.

4.1 Improvement of Checklist Items

Although we described the evaluation procedures of the checklist in as detailed a manner as possible, this resulted in increasing the burden on evaluators. Also, in the case of items with which the descriptions of the evaluation procedures were inadequate, for example when the evaluation targets were not specified clearly, the differences in interpretations between evaluators produced fluctuations in results as seen in Table 2(b).

For example, in the case of the item on the "contrast of color scheme," we instructed to check all colors in the evaluation target and perform judgment using mathematical expressions. As a result, the evaluators had to take tens of times longer periods for the evaluation than was expected and the match rate they achieved was 0%. We dealt with this by clarifying the evaluation target, permitting visual judgment of points that clearly do not involve any problem, and by providing a graphical means for use in judgments. This procedure has made it possible to significantly reduce the evaluation time and improve the match rate to 83%.

As described above, we improved the evaluation procedures and graphical materials of the items that presented low match rates in the 1st to 3rd test sessions, and presented the improved items in the 4th and 5th test sessions to another group of beginners and thus confirmed improvement of the match rate. In the future, too, we will continue to implement such improvements by considering the evaluation both of labor and match rates.

4.2 Determination of Evaluation Results Based on Deliberations

The present match rates are still not so high, as far as Table 3 shows, but many of the causes of non-matching are already found to be caused by human errors such as omissions and misunderstandings. Therefore, we selected 12 combinations of checklist items and evaluation target windows that might produce non-matching due to human errors in the 5 test sessions, and presented them to the beginners, who participated in the evaluation tests, so that they could deliberate on them with the aim of obtaining the final results. After the deliberations, the match rates of 10 of the 12 combinations were improved to 100% (they were between 0% and 70% before introduction of the deliberations). This fact witnessed the high probability of solving unmatching due to human errors by determining the final evaluation results based on deliberations participated by several evaluators. Incidentally, in the cases of 2 of the 10 checklist items with which the match rates were improved, omissions and other errors were also found in the evaluation results of skilled persons. However detailed the evaluation procedures are, human errors are unavoidable even among skilled evaluators, in as much as they are after all human beings. For the present, we recommend dealing with this issue by using several evaluators and determining the final results based on deliberations after evaluations. However, it may also be considered necessary to prepare a tool to provide further support for the work of evaluators in the future.

5. Conclusion

In the above, we have described the checklist construction by defining the evaluation procedures of each item in detail as well as the strict judgment criteria and the verification tests of the checklist. This procedure aims at eliminating any fluctuations in results that depend on the evaluators. As it has been found that the match rate can be improved by the participation of multiple evaluators and by determination of the final results after deliberations, this checklist can be expected to offer results with a certain degree of reliability when it is applied to such a methodology.

In the future, we will test the checklist items that have not so far been verified and feed back the results so that we can publicize the checklist as a method that can be used confidently by all. In addition, we are also planning to provide easier evaluation/scoring methods as well as various tools for improving the usability of the checklist itself.

References

1) Nielsen, J.; "Usability Engineering", Academic Press, 1993.
2) Ravden, S., Johnson, G.; "Evaluating Usability of Human-Computer Interfaces: A Practical Method", Prentice Hall 1989.
3) Kato, S. et al.: "HI SEKKEI CHEKKURISUTO TO SONO YUYOSEI NO HYOKA (A Human Interface Design Checklist and Its Effectiveness)," Transactions of Information Processing Society of Japan, Vol. 36, No. 1, pp.61-69, 1995.
4) Tone, K.: "GEEMU KANKAKU ISHI KETTEI-HO (Making a Decision, Feeling Like Playing a Game)," JUSE Press, pp.8-46, 1986.

Authors' Profiles

IKEGAMI Teruya
Assistant Manager,
Human Interface Center,
Common Platform Software Research Laboratories,
NEC Corporation

OKADA Hidehiko
Assistant Professor,
Kyoto Sangyo University