Difference between Evaluators and Users in Remote Asynchronous Web Usability Evaluation

Noboru Nakamichi
Nanzan University
nakamiti@nanzan-u.ac.jp

Toshiya Yamada
NTT IT Corporation
yamada.toshiya@ntt-it.co.jp

Mikio Kiura
Canon Inc.
kiura.mikio@canon.co.jp

Susumu Kuriyama
Mitsue-Links Co., Ltd.
kuriyama-susumu@mitsue.co.jp

Hidetake Uwano
Nara National College of Technology
uwano@info.nara-k.ac.jp

ABSTRACT
Some tools and environment may have applicability to remote asynchronous Web usability evaluation is being put into place. We experimented for comparing between users' evaluation and evaluators' evaluation. From comparison result, it is difficult to evaluate same evaluation for evaluators. But a usability expert in evaluators has a correlation with users' evaluation. And his evaluation only is not over-evaluating.

Author Keywords
Gazing point, Eye information, Performance, Usability testing.

ACM Classification Keywords
H.5.2 [Information Interfaces and Presentation]: User Interfaces - Evaluation/methodology

General Terms
Experimentation.

INTRODUCTION
The usability of a Web site is quite important because users are unwilling to read Web pages with low usability, such as hard to operate or understand, or react differently from their expectations. Discovering problems from the Web site by usability testing [1] are generally performed. However usability testing has a problem that recruiting costs of subjects is high. If evaluators can check only using interaction data in remote asynchronous environment, recruiting cost will decline.

In this paper, we confirm evaluators can check Web usability only using interaction data. The interaction data are gazing point information, mouse movement and browsing history. Subjects' interaction data was recorded, and they evaluate Web pages they visited. And evaluators check the Web pages based on subjects' interaction data. We compare the subjects' evaluation results and evaluators' evaluation result.

Copyright is held by the author/owner(s).
APCHI '12, August 28–31, 2012, Matsue-city, Shimane, Japan.

RECORDING OF USABILITY EVALUATION
We experimented for recording the subjects' interaction data. Recording tool is ITR-Recorder[2]. Subjects are 9 frequent users of the Internet. They have never visited the sites used in the experiment. We requested the subject to perform 5 task of looking for the starting salary of a master from the site of companies.

The Web pages that a subject visited are displayed to the subject. We requested the subject to choose the ease of use for every visited Web page from the four levels. In the experiment, we recorded the interaction data and subjects' evaluation result for 51 pages which the subject visited.

Evaluators reproduce the interaction data using ITR-Player (show fig. 1), and the evaluator checks the Web pages. We requested the evaluator to choose the ease of use for every visited Web page from the four levels. Evaluators are an expert A, a novice B and a novice C. The expert A has experience of eyetracking and usability testing. The novices are college students.

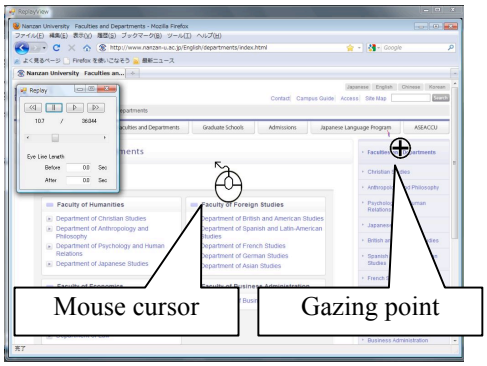

Figure 1. Replay Function of ITR-Player.

	Evaluation Result (pages)			
	Subjects	Evaluator		
		A	B	C
Usability level		**1 2 3 4**	**1 2 3 4**	**1 2 3 4**
1. hard to use	11	1 7 2 1	2 6 2 1	1 4 1 5
2. relatively hard to use	11	0 4 4 3	1 2 5 3	1 2 1 7
3. relatively easy to use	17	0 6 8 3	1 1 9 6	1 3 1 12
4. easy to use	12	0 3 6 3	1 2 4 5	0 1 2 9
Average	2.588	2.765	2.882	3.333

Table 2. Summary of Subjects' evaluation and Evaluators' evaluation.

Analysis method	A's evaluation	B's evaluation	C's evaluation
kappa statistic p-value	0.437	0.169	0.802
Correlation test p-value	**0.044**	**0.014**	0.064
Pearson's correlation coefficient	0.283	0.341	0.261
Sign test (one-tail) p-value	0.308	**6.62e-05**	**0.008**

Table 2. Comparison analysis between Subjects' evaluation and Evaluators' evaluation.

Table 1 shows the summary of the subjects' evaluation and the evaluators' evaluation.

COMPARISON OF THE EVALUATION RESULT

We analyzed the experimental result to compare difference between the subjects' evaluation and the evaluators' evaluation.

Kappa coefficient is a statistical measure of inter-rater agreement for qualitative items. Kappa measures the agreement between the subjects' evaluation and the evaluators' evaluation. The results of the kappa statistic in Table 2 show that the mean of each usability level for the cases with the subjects' evaluation is statistically non-agreement from that for the cases with all evaluators' evaluation. This result showed clearly that there is difference between the subjects' evaluation and the each evaluator's evaluation. It is difficult for evaluators to evaluate Web usability level which was in agreement with subjects from subjects' interaction data.

Next we measure correlation between the subjects' evaluation and the each evaluator's evaluation. The results of the correlation coefficient in Table 2 show poor positive correlation between the subjects' evaluation and the each evaluator's evaluation. Expert A's evaluation and novice B's evaluation are statistically significant. From these results, it is difficult for evaluators to evaluate Web usability level which was in agreement with subjects but evaluators can evaluate the same tendency of subjects' evaluation.

There is a positive correlation between the subjects' evaluation and the each evaluator's evaluation. From the difference between the average of subjects' evaluation and the average of evaluators' evaluation, there is a possibility that the evaluators have overvaluation. The result of one-tail sign test in Table 2 show that the mean of each usability level for the cases with the subjects' evaluation is statistically different from that for the cases with novice B's evaluation and novice C's evaluation. But expert A as a usability expert has not overvaluation.

From these results, it is difficult for evaluators to evaluate same Web usability level of subjects' evaluation. However the evaluation results by the expert A have poor positive correlation with the subjects' evaluation. And he has not overvaluation.

CONCLUSION

We analyzed difference between evaluators' evaluation and users' evaluation to confirm evaluators can check Web usability only using interaction data. We showed clearly that there is difference between the users' evaluation and the each evaluator's evaluation. It is difficult for evaluators to evaluate Web usability level which was in agreement with users from users' interaction data. However the evaluation results by the usability expert have poor positive correlation with the users' evaluation. And he has not overvaluation. From these result, a usability expert possible evaluate Web usability using only users' interaction data in remote asynchronous place.

ACKNOWLEDGMENTS

This research was partially supported by Nanzan University Pache Research Subsidy I-A-2 for the 2012 academic year.

REFERENCES

1. Dumas, J. S., Redish, J. C., A Practical Guide to Usability Testing. Ablex Publishing, (1993).
2. Nakamichi, N., Kiura, M., Yamada, T., Uwano, H. Collaborative Visualization of Web Interactions for Usability Testing, *9th Pan-Pacific Conference on Ergonomics(PPCOE2010)*, (2010).