The Effect of Severity Ratings on Software Developers' Priority of Usability Inspection Results

Asbjørn Følstad
SINTEF ICT
Forskningsveien 1
0314, Oslo, Norway
+47 22067515

asf@sintef.no

ABSTRACT

Knowledge of the factors that affect developers' priority of usability evaluation results is important in order to improve the interplay between usability evaluation and software development. In the presented study, the effect of usability inspection severity ratings on the developers' priority of evaluation results was investigated. The usability inspection results with higher severity ratings were associated with higher developer priority. This result contradicts Sawyer et al. [7], but is in line with Law's [5, 6] finding related to the impact of user test results. The findings serve as a reminder for HCI professionals to focus on high severity issues.

Categories and Subject Descriptors

H5.m.Information interfaces and presentation (e.g., HCI): Miscellaneous.

Keywords

Usability evaluation, usability inspection, developers' priority, impact, severity ratings.

1. INTRODUCTION

One important indicator of successful interplay between usability evaluation and software development is the extent to which evaluation results are associated with subsequent changes in the system under development. This indicator, termed the "impact" [7] or "persuasive power" [4] of usability evaluation results, may reflect whether or not a usability evaluation has generated results that are needed in the development process.

Problem severity is a characteristic of usability evaluation results that has been suggested to affect the impact of usability evaluation results. There is, however, divergence in the literature regarding the actual effect of severity ratings on developers' prioritizing of usability evaluation results. Sawyer et al.'s [7] study of the impact of usability inspection results indicated that usability inspectors' severity ratings had no effect on the impact of the evaluation results; reported impact ratios were 72% (low severity issues), 71% (medium severity issues), 72% (high severity issues). In contrast to this finding Law [5, 6], in a study of the impact of user tests, reported a tendency towards higher severity results having higher impact; reported impact ratios were 26% (minor problems), 42% (moderate problems), 47% (severe problems). Law's findings, however, were not statistically significant [5]. Hertzum [3] suggested that the effect of severity classifications may change across the development process, e.g. high severity evaluation results may have relatively higher impact in later phases of development. Law's study was conducted relatively late in the development process, on the running prototype of a digital library. Sawyer et al. did not report in which development phases their usability inspections were conducted.

In order to complement the existing research on the effect of severity ratings on the impact of evaluation results, an empirical study of the impact of usability inspection results is presented. The data of the present study was collected as part of a larger study reported by Følstad [2], but the results discussed below have not previously been presented.

2. RESEARCH PROBLEM AND HYPOTHESIS

The research problem of the present study was formulated as:

What is the effect of usability inspectors' severity ratings on developers' priority of usability inspection results?

The null hypothesis of the study (no effect of severity ratings) followed the findings of Sawyer et al., and the alternative hypothesis (H1) was formulated in line with the findings presented by Law:

H1: High severity issues will tend to be prioritized higher by developers than low severity issues.

3. METHOD

Usability inspections were conducted as group-based expert walkthroughs [1]. The objects of evaluation were three mobile work-support systems for medical personnel at hospitals, politicians and political advisors, and parking wardens respectively. All systems were in late phases of development, running prototypes close to market. The usability inspectors were 13 HCI professionals, all with >1 year work experience (*Mdn*=5

I-USED'08, September 24, 2008, Pisa, Italy

years)[1]. Each inspector participated in one of three evaluation groups, one group for each object of evaluation. The walkthroughs were conducted as two-stage processes where (1) the individual evaluators noted down usability issues (usability problems and change suggestions) and (2) a common set of usability issues were agreed on in the group. All usability issues were to be classified as either *Critical* (will probably stop typical users in using the application to solve the task), *Serious* (will probably cause serious delay for typical users …), or *Cosmetic* (will probably cause minor delay …). The output of the usability inspections was one report for each object of evaluation, delivered to each of the three development teams respectively.

Three months after the evaluation reports had been delivered individual interviews were conducted with development team representatives. The representatives were requested to prioritize all usability issues according to the following: *High* (change has already been done, or will be done no later than six months after receiving the evaluation report), *Medium* (change is relevant but will not be prioritized the first six months), *Low* (change will not be prioritized), *Wrong* (the item is perceived by the developer to be a misjudgment). In order to align the resulting developers' priorities with the impact ratio definitions of Law and Sawyer et al., the priority *High* was recoded as "Change", and the priorities *Medium, Low and Wrong* were recoded as "No change".

4. RESULTS

The evaluation groups generated totally 167 usability issues. The three objects of evaluation were associated with 44, 61, and 62 usability issues respectively. The total impact ratio (number of issues associated with change/total number of issues [following 7 and 6]) was 27%, which is relatively low. The relationship between the developers' priorities and the usability inspectors' severity ratings is presented in Table 1.

Table 1. Usability issues distributed across developers' priorities and usability inspectors' severity ratings

	Not Classified	Cosmetic	Serious	Critical
Change	6	9	18	12
No change	46	31	26	16
Impact ratio	12%	23%	41%	43%

Visual inspection of Table 1 shows a tendency towards higher priority given to usability issues with severity ratings serious and critical. A Pearson Chi-Square test showed statistically significant differences in priority between severity rating groups; X^2=14.446, df=3, p(one-sided)=.001.

5. DISCUSSION

The presented results indicate that severity ratings may have significant impact on developers' priority of results from usability inspections. This finding contributes to our understanding of severity ratings as a characteristic of usability evaluation results that may help to identify which usability evaluation results that are needed in software development.

The finding is particularly interesting since it contradicts the conclusions of Sawyer et al. and therefore may provoke necessary rethinking regarding usability inspectors ability to provide severity assessments that are useful to software engineers.

It is also interesting to note that the results are fully in line with Law's findings related to severity ratings of user test results. The present study may thus serve to strengthen Law's conclusions. Curiously, the impact ratios of the different severity levels in Law's study and the present study are close to being identical.

Why, then, do the present study indicate that the severity ratings of usability inspection results may have an effect on the developers' priority, when Sawyer et al. did not find a similar effect? One reason may be the relatively high impact ratios reported by Sawyer et al., something that may well result in a greater proportion of low severity issues being prioritized. Another reason may be that the present study, as the study of Law, favored high severity evaluation results since the usability evaluations were conducted relatively late in the development process [cf. 3]. Sawyer et al. do not state which development phases their usability inspections were associated with, but their relatively high impact ratios suggest that their inspections possibly may have been conducted in earlier project phases.

The present study, as the study of Law, indicates that the identification of a low severity usability issue typically is of less value to software developers than the identification of a high severity issue. This should serve as a reminder for HCI professionals to spend evaluation resources on identification and communication of higher severity usability issues.

6. ACKNOWLEDGMENTS

This paper has been written as part of the RECORD project, supported by the VERDIKT program of the Norwegian Research Council.

7. REFERENCES

[1] Følstad, A. 2007. Group-based Expert Walkthrough. In: D. Scapin, and E.L.-C. Law, Eds. R3UEMs: Review, Report and Refine Usability Evaluation Methods. Proceedings of the 3rd. COST294-MAUSE International Workshop, 58-60.

[2] Følstad, A. 2007. Work-Domain Experts as Evaluators: Usability Inspection of Domain-Specific Work-Support Systems. International Journal of Human-Computer Interaction 22(3), 217-245.

[3] Hertzum, M. 2007. Problem Prioritization in Usability Evaluation: From Severity Assessments Toward Impact on Design. International Journal of Human-Computer Interaction, 21(2), 125–146.

[4] John, B.E., and Marks, S.J. 1997. Tracking the effectiveness of usability evaluation methods. Behaviour & Information Technology, 16, 188–202.

[5] Law, E. L.-C. 2004. A Multi-Perspective Approach to Tracking the Effectiveness of User Tests: A Case Study. In Proceedings of the NordiCHI Workshop on Improving the

[1] The study reported by Følstad also included separate evaluation groups with work-domain experts. The results of these groups were not included in the current study, in order to make a clear-cut comparison with the findings of Law and Sawyer et al.

Interplay Between Usability Evaluation and User Interface Design, K. Hornbæk, and J. Stage, Eds. University of Aalborg, HCI Lab Report no. 2004/2, 36-40.

[6] Law, E. L.-C. 2006. Evaluating the Downstream Utility of User Tests and Examining the Developer Effect: A Case Study. International Journal of Human-Computer Interaction, 21(2), 147-172.

[7] Sawyer, P., Flanders, A., Wixon, D. 1996. Making a Difference - The Impact of Inspections. In Proceedings of the CHI'96 Conference on Human Factors in Computing Systems, 376–382.