A Multi-Perspective Approach to Tracking the Effectiveness of User Tests: A Case Study

Effie Lai-Chong Law
Computer Engineering and Networks Laboratory (TIK)
Swiss Federal Institute of Technology (ETH Zürich)
Gloriastrasse 35, CH-8092 Zürich, Switzerland
law@tik.ee.ethz.ch

ABSTRACT
In this paper we delineate a multi-perspective approach to tracking the effectiveness of user tests, which have been performed on a web-based educational system. We have identified a definitional issue about the effectiveness of usability evaluation methods and thus proposed that tracking and supporting the integration of usability evaluation results should be an integral part of usability engineering process. We also identified a theoretical void in studying the persuasive power of usability evaluation results and thus proposed to bridge the gap with process theories of persuasion. We have collected data from several sources representing different roles and perspectives – usability practitioner, system developer, system manager, and representative end-users. We have consolidated the multi-perspective data to address several hypotheses that predict the persuasiveness of different qualities of usability problems to induce fixes and the effectiveness of such fixes. Implications for future research on this specific topic are inferred.

Categories and Subject Descriptors
H.5.2. Information Interfaces and Presentation: User Interface – Evaluation/methodology

General Terms
Experimentation, Measurement

Keywords
Effectiveness, User Test, Usability Problem, Persuasion Theories

1. INTRODUCTION

Effectiveness is a key notion in usability research. Nevertheless, up to now there has not yet been any well defined parameter that can be used as a reliable and valid indicator of the effectiveness of usability evaluation methods (UEMs). Given the ultimate goal of usability evaluation is to improve the system of interest, the evaluation tool or method selected can be proved to be effective only if such a goal can be attained, ideally at an optimal cost. A number of studies on comparing the effectiveness of different UEMs were conducted in early 1990s [e.g., 5, 10, 13, 19]. Unfortunately, these studies were harshly criticized as lack of experimental rigor and the outcomes were regarded as dubious [8]. In some recent related works [e.g., 4, 9, 15, 22], two parameters – thoroughness and validity – have been adopted to define the effectiveness of UEMs. However, such a definition seems oversimplified with the primary goal of usability evaluation of improving a system remaining unfulfilled. Usability evaluation should not cease at the point when a list of UPs is produced [23]. More important is to insure that such a list can somehow render the system more usable and useful. Indeed, usability evaluation results can have stronger impacts when developers are provided concrete and feasible improvement suggestions from users and usability practitioners than when they are merely confronted with negative criticisms. We advocate that tracking the effectiveness and supporting the incorporation of usability evaluation results into the improvement of the system tested should be an integral part of the overall usability engineering process.

Further, we observe that there is a theoretical void in studying the persuasive power of usability evaluation results and attempt to bridge the gap with process theories of persuasion. Based on the assumption that the effectiveness of a UEM is a combinatorial parameter, we employ multiple methods to collect data from different stakeholders involved in the usability evaluation. The data thus collected may shed some light onto the significant practical issue about the reliability of user tests, which are normally employed to benchmark other UEMs. In the ensuing text, we will elaborate on the aforementioned theoretical and empirical issues.

2. LITERATURE REVIEW

The literature on tracking the effectiveness of UEMs is actually limited. John and Mark's [12] exploratory work is representative in this area, though its methodology has been challenged [3, 11]. Nonetheless, the case study documented by the two authors illustrates clearly how intricate and resource-demanding such a task can be. According to their model, it is necessary to estimate the values of three key variables: (i) How many of usability problems predicted by a UEM can really be experienced by end-users (*predictive power*)? How many of these usability problems can result in fixes or changes of code (*persuasive power*)? How many of these fixes can really improve the usability of the system (*design-change effectiveness*)? The five UEMs investigated were all predictive or analytic. Respective lists of UPs thus derived were benchmarked with clusters of user tests. John and Mark tracked those UPs predicted by the UEMs of interest and verified by the user tests, but not those UPs that were directly discovered by the user tests and overlooked (i.e. misses) by any of the UEMs.

A method known as RITE (Rapid Iterative Testing and Evaluation) for evaluating the efficacy of fixes of UPs identified in user tests has recently been advocated by Medlock and his colleagues [18, 23]. The key to the success of RITE is the intense participation of at least one member of the development team and he usability engineer. They must communicate seamlessly so that

corrigible UPs can get fixed without having to go through any formal process. Despite its claimed advantages, the generalizability of the RITE method is yet to be demonstrated.

Intuitively speaking, developers and manager are more likely to commit to resolving UPs if they are persuaded about the necessity and utility of potential fixes. This concept of persuasiveness, however, is not well addressed in the literature. In fact, John and Marks [12] do not root their notion of 'persuasive power' in any social cognitive theories. We believe that process theories of persuasion [7], especially their emphasis on distinctive cognitive mechanisms, can shed some light into the issue pertinent to the acceptance and adoption of usability evaluation results by developers, designers, and managers alike. According to N.H. Anderson's information integration theory [2], four general determinants of weight of information are its relevance, salience, reliability and quantity. The heavier the weight, the higher the likelihood the information will be accepted and yield the action (implicitly or explicitly) suggested. Furthermore, McGuire [17] addresses that distal persuasion variables such as recipient's intelligence, motivation and personality; sender's perceived domain-specific expertise; message's fear arousal and communication modality can have effect upon the reception and acceptance of the message content. We infer some implications from these theories to the understanding of usability problems management.

3. RESEARCH HYPOTHESES

We formulate different hypotheses based on the literature perused and summarized above. Four of them are presented here and the others will be reported elsewhere. Put briefly, we assume that the persuasive power of user-test results to induce fixes depends on their saliency and ability to motivate developers (i.e. criticality of UPs) and reliability (i.e., frequency of UPs).

H1a: Due to the saliency effect, UPs rated with higher severity level are more persuasive to induce fixes than those rated with lower severity.

H1b: Fixes of severe UPs are more effective than those of moderate or minor UPs, because developers are more motivated to fix the former than the latter.

H2a: Due to the reliability effect, UPs identified with higher frequency are more persuasive to induce fixes than those with lower frequency?

H2b: Fixes of frequent UPs are more effective than those of rarer UPs, because developers can have more information about the former than the latter.

4. BASELINE MEASUREMENTS

The system on which we performed international user tests (IUT) was a platform (version 0.85; March 2003) designed for enabling the exchange of online educational content among academic and industrial institutions. The interface of this brokerage platform was usability tested with 19 representative end users from four different European countries. Standard user test procedures were adopted [6] and implemented locally with the respective language versions by Local Testers. Each participant was asked to perform ten task scenarios covering the core functionalities of the platform and to *think aloud* to maintain a running commentary as he or she interacted with the system. In the usability evaluation report, for every UP, descriptions (where, what and how), severity level (severe, moderate or minor [1]) and frequency (number of users experienced) were presented.

5. TRACKING EFFECTIVENESS

We tracked the effectiveness of the IUT with the baseline list of 81 UPs, of which 52 were given plausible causes and/or potential redesign solutions by the usability specialist. Specifically, we aim to answer three major questions:

i. How many of the UPs reported have induced fixes?
ii. How persuasive were different UP qualities to induce fixes?
iii. How effective were the fixes?

Three sources of data were collected through different procedures.

5.1 Usability Specialist Review

The usability specialist, who was involved in extracting UPs from the qualitative data of the IUT, re-evaluated each of the 81 UPs observed in the previous version (v. 0.85) with the recent version of the platform (version 1.0; January 2004). She identified those UPs that did not receive any fix and described how the other UPs were fixed.

5.2 Development Team Portfolio

The chief developer and the platform manager, who was heavily involved in deciding which and how UPs to be fixed, were asked to provide data on:

(i) the effort invested or would be invested in fixing the UPs
(ii) the decision-making factors for fixing or not fixing the UPs
(iii) the techniques and references used for implementing the fixes.

The developer described the effort with a five-point scale (very short, short, medium, long, very long). He added brief remarks for 15 UPs with most of them being related to the techniques employed for the actual or would-be changes. The platform manager also added some brief remarks of various natures for 41 UPs.

5.3 End User Retest

Three male participants, who took part in the IUT one year ago, were re-invited to evaluate the current version of the platform. They were all university faculty members with high level of competence in information technology and high level of knowledge about e-Learning (i.e. the domain of the platform evaluated). Their participations were voluntary. In the testing session, they were required to perform a set of 12 task scenarios with nine of them being more or less the same as those they performed in the IUT, and the procedure used was also similar.

6. RESULTS

6.1 Usability Specialist Review

31 out of the 81 UPs identified in the IUT were fixed by the developers. In other words, 50 UPs did not receive any fix. The

Impact Ratio (see Equation 1) is only 38.3%, which is relatively low. We further broke the results down in terms of severity level (Table 1) and frequency (Table 2).

Equation 1 [21]:

$$\text{Impact Ratio (IR)} = \frac{\text{Number of Problems Receiving a Fix}}{\text{Total Number of Problems Found}} * 100$$

Table 1. Impact ratios by problem severity levels

	Minor	Moderate	Severe
With Fix / Change (C)	6	17	8
No Fix / Change (NC)	17	24	9
Impact Ratio (IR)	26.1%	41.5%	47.1%

The IR of severe UPs is higher than that of the other two. However, Chi-Square tests show that there are no significant differences between the cells in Table 1.

Table 2. Impact ratios by problem frequency levels

	Low	Medium	High
With Fix / Change (C)	14	8	9
No Fix / Change (NC)	24	16	10
Impact Ratio (IR)	36.8%	33.3%	47.4%

*Low = single user; Medium = >1 and <=20% of the users; High =>20%

The IR of "High"-UPs is larger than that of the other two, but Chi-Square tests show that there are no significant differences between the cells in Table 2.

6.2 Development Team Portfolio

For each of the 29 out of 31 fixes, the chief developer reported the effort required with the five-point scale mentioned earlier (NB: the detailed results will be reported elsewhere). None of the UPs falls in the category 'very long'. It implies that the developer did not tend to fix any UP entailing much effort.

6.3. End User Retest

The three test participants - P1, P2 and P3 – evaluated the earlier version of the platform about one year ago. The rationales for recruiting "old" participants were to observe whether the UPs they experienced previously would perish or persist and to minimize the user effect [14]. Three separate lists of usability problems were extracted and they were compared with their counterparts obtained in the earlier IUT (Table 3). Note that those UPs associated with the completely new functionalities of the platform to which the three participants had never exposed in the IUT were *not* counted.

Table 3. Main results of end-user retest

	P1	P2	P3
No. of UPs **already** experienced in the **earlier** version	14	26	16
No. of UPs **persistently** experienced in the **current** version	4	3	2
No. of UPs **no longer** experienced in the **current** version	10	23	14
No. of UPs **newly** experienced in the **current** version	5	8	6

An inherent limitation of our study is that the effectiveness of the fixes can only be tracked based on the three users' evaluations. Clearly, the validity and reliability of the results could be higher if more users were involved. Nevertheless, a UP could be experienced by none, one, two or all of the three users in v.0.85, the same UP could also be experienced by none, one, two or all of the three users in v.1.0. We developed a data analysis scheme accordingly (details will be reported elsewhere).

Out of the 31 fixes, 15 were effective or mildly effective, 11 had no effect, four were bad and one was terrible. The reported effort for this terrible fix was "very short". The UP concerned was that the error message was not conspicuous enough to be spotted effectively and its severity level was moderate. The fix involved enlarging the font of the text with the colour remaining the same. The system manager remarked that the fix was based on a 'typical approach' for attracting attention to a message. Two of the five effective fixes involved a relatively high effort (i.e., "long") and both were rated severe, whereas the reported efforts of the other three less severe UPs were "very short" or "short". Moreover, we computed the effectiveness of fixes of UPs of different severity and frequency levels (see Table 4 and Table 5).

Table 4. Fix-effectiveness ratio by severity level

	Severe	Moderat	Minor
Effective# Fixes	5	7	3
Ineffective* Fixes	3	10	3
Fix- Effectiveness Ratio (FER)	38.5%	29.2%	50%

Note: # include mildly effective; * include no effect, bad and terrible fixes

The FER of minor UPs is higher than that of the other two. However, Chi-Square tests show that there are no significant differences between the cells in Table 4.

Table 5. Fix-effectiveness ratio by frequency level

	High	Medium	Low
Effective# Fixes	6	2	7
Ineffective* fixes	3	6	7
Fix-Effectiveness ratio (FER)	66.7%	25%	50%

The FER of "Low"-UPs is larger than that of the other two. However, Chi-Square tests show that there are no significant differences between the cells in Table 5.

7. GENERAL DISCUSSION

In the ensuing text, we will go through the research hypotheses delineated in Section 3. Note that the current work was an exploratory case study aiming to give directions of the related future research. As there was no *a prior* stringent experimental manipulation or control, the results obtained cannot lead to any conclusive claims.

H1a: Severe UPs would be more likely to induce fixes

H1a was not supported statistically. However, results show that the UPs rated with high severity tended to be more persuasive to induce fixes than their less severe counterparts (cf. Impact Ratios in Table 2). Arguing along the line of N.H. Anderson's

information integration theory [2], the weight of a piece of information increases with its saliency and is more likely to capture a recipient's attention. Apparently, a UP tagged with a 'severe' label tends to be more salient than one tagged with a 'minor' label. The heightened saliency and the associated emotional responses (i.e., anxiety or fear) can become a force to drive corrective actions. This mechanism may explain why the severe UPs had a higher rate of receiving fixes.

H1b: Severe UPs would have more effective fixes
H1b was rejected. Fixes of minor UPs tended to be more effective than their more severe counterparts (cf. Fix-Effectiveness Ratios in Table 4), though statistically the difference was insignificant. As minor UPs were generally less complicated than severe UPs, therefore the Fix-Effectiveness Ratio tended to be higher.

H2a: Frequent UPs would be more likely to induce fixes.
H2a was not supported statistically. However, results show that the UPs rated with higher frequency tended to be more persuasive to induce fixes than their less frequent counterparts (cf. Impact Ratios in Table 2). We can again apply the information weight model to explain the observed difference in the tendency to fix. Clearly, it is more convincing that a UP is a real problem if more than one user has experienced it. Indeed, some usability researchers and practitioners tend to discard UPs with single occurrence from further analyses [16], based on the assumption that the peculiarity of users' beliefs and attitudes may play in role in ringing "false alarms".

H2b: Frequent UPs would have more effective fixes
H2b was not supported statistically. However, fixes of highly frequent UPs tended to be more effective than their less frequent counterparts (cf. Fix-Effectiveness Ratios in Table 5). Presumably, the higher the number of users experience a UP, the more elaborated the description of the UP will be, especially the contextual data (cf. Anderson's "relevance"), from which the developer can gain more insights into devising appropriate fixes. This assumption on elaborative-ness (cf. Anderson's "quantity") can somewhat explain the observed difference in the effectiveness of fixes for UPs with different frequencies.

In summary, the results presented above reveal two intriguing facts: First, the outcomes of user tests cannot be effectively incorporated into redesign of a system, considering only 38% of the UPs reported receiving a fix and about 68% (= 15/22) of these fixes were effective or mildly effective (NB: this percentage will be inflated if we take the nine UPs that none of the three users experienced in either of the two versions into account). In other words, approximately only 26% (= 38%*68%) of the results of a user test were applicable in improving the system in question. Second, users could be highly adaptive to the "imperfections" of the system, considering that on average 82.4% (Table 3) of the previously experienced UPs was no longer a nuisance and that 38% (=19/50) of the non-fixed UPs did not cause any further trouble, at least for the three users. Such "self-dissolution" of usability problems can be attributed to different possible reasons: the learnability of the system, the increased tolerance of the user towards design flaws, the giving up of lodging complaints that make no effect (i.e. non-fixed UPs reported in the earlier user test), the overcoming of initial psychological barriers of deploying a new system, etc.

8. CONCLUSION

The current exploratory study is not meant to provide any definitive answers to the issues related to tracking the effectiveness of user tests. Instead, it aims to draw the HCI community to this neglected issue. As demonstrated in the foregoing descriptions, tracking the effectiveness of a user test is very resource-demanding and complex. It is likely to be one of the reasons why usability practitioners do not bother to poke into this question. By the same token, managers do not bother to analyse the ROI (Return On Investment) of usability evaluation [20].

Furthermore, the open problem addressed in the beginning of the paper still remains unanswered: *What is the reliable and valid indicator of the effectiveness of UEM*? While we strongly believe that it should be more than conventionally defined "thoroughness" and "validity", we have not yet been able to derive a neat and tidy mathematical formula, which can reduce a cluster of variables into a single comprehensible and computational entity. Nevertheless, as mentioned above, we posit that the effectiveness of a UEM should be specified with two major terms – *Persuasiveness of Problem*- how many percent of UPs identified can induce a fix and *Efficacy of Fix* - How many of the fixes are effective in the sense that they do not entail any re-fix. Besides, process theories of persuasion [7] should further be explored to study the topic of tracking effectiveness of UEMs.

9. REFERENCES

[1] Artim. J. M. (2003). Usability problem severity ratings. Access at: http://www.primaryview.org/CommonDefinitions/

[2] Anderson, N. H. (1981). *Foundations of information integration theory*. Academic Press.

[3] Carroll, J. M. (1998). On an experimental evaluation of claim analysis. *Behaviour & Information Technology, 17*(4), 242-243.

[4] Cockton, G., & Woolrych, A. (2001). Understanding inspection methods. In A. Blandford, J. Vanderdonckt, & P.D. Gray (Eds.), People and Computer XV (pp. 171-192). Springer-Verlag.

[5] Desurvire, H.W., Kondziela, J.M., & Atwood, M.E. (1992). What is gained and lost when using evaluation methods other than empirical testing. In *Proceedings of CHI'92*.

[6] Dumas, J.S., & Redish, J.C. (1999). *A practical guide to usability testing* (rev. ed.). Exeter: Intellect.

[7] Eagly, A., & Chaiken, S. (1993). *Psychology of Attitudes.* NY: Harcourt, Brace Jovanovic.

[8] Gray, W.D., & Salzman, M.C. Damaged merchandise? *Human-Computer Interaction, 13* (1998), 203-262.

[9] Hartson, H.R., Andre, T.S., & Williges, R.C. (2001). Criteria for evaluating usability evaluation methods. *International Journal of Human-Computer Interaction, 13*(4), 373-410.

[10] Jeffries, R., Miller, J.R., Wharton, C., Uyeda, K.M. (1991). User interface evaluation in the real world: A comparison of four techniques. In *Proceedings of CHI'91*.

[11] John, B. (1998). On our case study of claims analysis and other usability evaluation methods. *Behaviour and Information Technology, 17*(4), 244-246.
[12] John. B., & Marks, S.J. (1997). Tracking the effectiveness of usability evaluation method. *Behaviour and Information Technology, 16*(4/5), 188-202.
[13] Karat, C.-M., Campbell, R., & Fiegel, T. (1992). Comparison of empirical testing and walkthrough methods in user interface evaluation. In *Proceedings CHI'92*.
[14] Law, E. L.-C., & Hvannberg, E.T. (2004). Analysis of the combinatorial user effect in international usability tests. In *Proceedings of CHI'04*, April 2004, Vienna, Austria.
[15] Law, E. L-C., & Hvannberg, E. T. (2004). Analysis of strategies for estimating and improving the effectiveness of heuristic evaluation. In *Proceedings of NordiCHI 2004*, 23-27 October, Tampere, Finland.
[16] Lewis, J.R. (1994). Sample sizes for usability studies: Additional considerations. *Human Factors, 36*(2), 368-378.
[17] McGuire, W. J. (1968). Personality and Attitude Change: An Information Processing Theory. In Greenwald and Brock (eds.), *Psychological Foundations of Attitude*.
[18] Medlock, C. M., Wixon, D., Terrano, M., Romero, R.L., & Fulton, B. (2002). Using the RITE method to improve products; a definition and a case study. In *Proceedings of UPA'02*.
[19] Nielsen, J., & Philips, V.L. (1993). Estimating the relative usability of two interfaces: Heuristic, formal, and empirical methods compared. In *Proceedings of INTERACT'93*.
[20] Rosenberg, D. (2004). The myths of usability ROI. *Interactions, Sept-Oct*, 23-29.
[21] Sawyer, P., Flanders, A., & Wixon, D. (1996). Making a difference – the impact of inspections. In *Proceedings of CHI'96*.
[22] Sears, A. (1997) Heuristic walkthroughs. *Journal of Human-Computer Interaction, 9*, 3, 213-234.
[23] Wixon, D. (2003). Evaluating usability methods: Why the current literature fails the practitioner. *Interactions, 10, 4*, 29-34.