Problems of Consolidating Usability Problems

Effie Lai-Chong Law
University of Leicester/ ETH Zürich
LE1 7RH Leicester/ Institut TIK
UK/Switzerland
+44 116 2717302
law@tik.ee.ethz.ch

Ebba Thora Hvannberg
University of Iceland
107 Reykjavik
Iceland
+354 525 4702
ebba@hi.is

ABSTRACT
The process of consolidating usability problems (UPs) is an integral part of usability evaluation involving multiple users/analysts. However, little is known about the mechanism of this process and its effects on evaluation outcomes, which presumably influence how developers redesign the system of interest. We conducted an exploratory research study with ten novice evaluators to examine how they performed when merging UPs in the individual and collaborative setting and how they drew consensus. Our findings indicate that collaborative merging causes the absolute number of UPs to deflate, and concomitantly the frequency of certain UP types as well as their severity ratings to inflate *excessively*. It can be attributed to the susceptibility of novice evaluators to persuasion in a negotiation setting, and thus they tended to aggregate UPs leniently. Such distorted UP attributes may mislead the prioritization of UPs for fixing and thus result in ineffective system redesign.

Categories and Subject Descriptors
H.5.2 [**User Interfaces**]: *Evaluation/Methodology*

General Terms
Measurement, Performance, Experimentation, Theory

Keywords
Usability problems, Merging, Filtering, Consensus building, Downstream utility, Severity, Confidence, Evaluator effect

1. INTRODUCTION
The extent to which UPs identified by different users/analysts overlap seems unpredictable, despite the persistent research efforts of formalizing the cumulative relation between the numbers of users/analysts and UPs ([7], [8], [10]). The practical implication of these concerns is to recruit as many users/analysts as the project's resources allow, thereby maximizing the probability of identifying most, but impossibly all, UPs.

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

One concomitant procedure of involving multiple users/analysts in usability evaluation is to consolidate UPs identified by different users/analysts to produce a master list. Such a consolidation process can serve two purposes: (i) providing a design team with neat and clean information to facilitate system redesign, and (ii) enhancing the validity of comparing the effectiveness of different (instances of) usability evaluation methods (UEMs). This process consists of two phases [1]: The first step is known as *filtering*, that is, to eliminate duplicates *within* a list of UPs identified by a user when performing a certain task with the system under scrutiny or by an analyst when inspecting it. The second step is *merging*, that is, to combine UPs *between* different lists identified by multiple users/analysts, to retain unique, relevant ones, and to discard unique, irrelevant ones. While such consolidation procedures are commonly practised by usability professionals and researchers, little is known about how it is exactly done and what impact it can have on final evaluation outcomes and eventually on system redesigns, especially when severity ratings play a non-trivial role in the prioritization strategy for UP fixing ([2], [3]).

In the HCI literature, the UP consolidation procedure is mostly described at a coarse-grained level. Nielson [9], when addressing the issue of multiple users/analysts, highlighted the significance of merging different UP lists, but he did not specify how this should be done. Connell and Hammond [1], in comparing the effectiveness of different UEMs, delineated the merging procedure at a rather abstract level. Further, Hertzum and Jacobsen [4] coined the notion of *evaluator effect* that has drawn much attention from the HCI community towards the reliability and validity issues of usability evaluation. Nonetheless, their work focused on problem extraction on an *individual* basis rather than problem merging on a *collaborative* basis. More recently, a tool for merging and grouping UPs has been developed [5], which, however, supports the work of individual evaluators but neglects the collaborative aspect of usability evaluation.

In summary, the actual practice of UP consolidation is largely open, unstructured and unchecked. With the major goals to examine the impact of the UP consolidation process and to understand the mechanism underlying the consensus building process, we have conducted a research study. In this paper we summarize the main findings on the first issue while leaving out the second one as the data are still being analyzed.

2. RESEARCH METHODS
The empirical study was conducted at a university in the UK. Ten students (one female) majored in computer science were recruited. All have acquired reasonable knowledge of HCI and

experience in user-based evaluation through lectures and projects. They were grouped into five pairs. An e-learning platform was usability evaluated (i.e. think aloud) with representative end-users one year ago. Among different types of data collected, we employed for this current study the observational reports written by the experimenter who was present throughout the testing sessions and registered the users' behaviours in very fine detail. We also developed several structured forms to register the participants' findings in the different steps of our study. All the participants had to attend two testing sessions: In the first one they performed *Individual Problem Extraction* and *Individual Problem Consolidation*, and about a week later, they paired up to perform *Collaborative Problem Consolidation.*

2.1 Individual Problem Extraction

Each participant was given the narrative observational reports (printed texts) how the users P1 and P2 performed Task 1 (T1) "Browse the Catalogue" and Task 2 (T2) "Provide and Offer a Learning Resource". For each UP extracted, the participant was required to record in a structured analysis form five attributes:

1. Develop UP identifier with a given format;
2. Provide a UP description as detailed as possible;
3. Select criteria from a given list to justify the UP;
4. Judge the severity level of UP: minor, moderate, severe;
5. How confident the evaluator was that the UP identified was true: 1 lowest – 5 highest;

After completing the analysis form for T1, the participant was asked to apply the same procedure to P1's T2, and then to P2's T1 and T2 (Figure 1). In other words, each participant was required to analyse four sets of data (P1-T1, P1-T2, P2-T1 and P2-T2).

2.2 Individual Problem Consolidation

With the four lists of extracted UPs, the participant was required to filter out any duplicate within the lists and then merge similar UPs, resulting in two sets of UPs (i.e. P1-T1 and P2-T1 as one set; P1-T2 and P2-T2 as another set). Unique UPs identified would be retained or discarded during this process. The participants were asked to record the outcomes in the same form for problem extraction, but they needed to indicate explicitly in the column UP-identifier which UPs were combined. Severity and confidence levels could also be adjusted. No time limit was imposed.

2.3 Collaborative Problem Consolidation

With a break of several days, two participants of a group came together to merge their respective lists of UPs prepared in the individual sessions into a master list. They could access all the materials used in the earlier sessions. They were asked to track every item (i.e., a single UP or combined UPs) in their own consolidated list by recording in a structured form which of the three possible changes was made - merged (with which one), retained or discarded. No time limit was imposed on any of the above procedures. While individual and collaborative problem consolidation basically involved similar sub-tasks, the latter was conducted to observe how the collaborative setting influenced an individual's merging strategies.

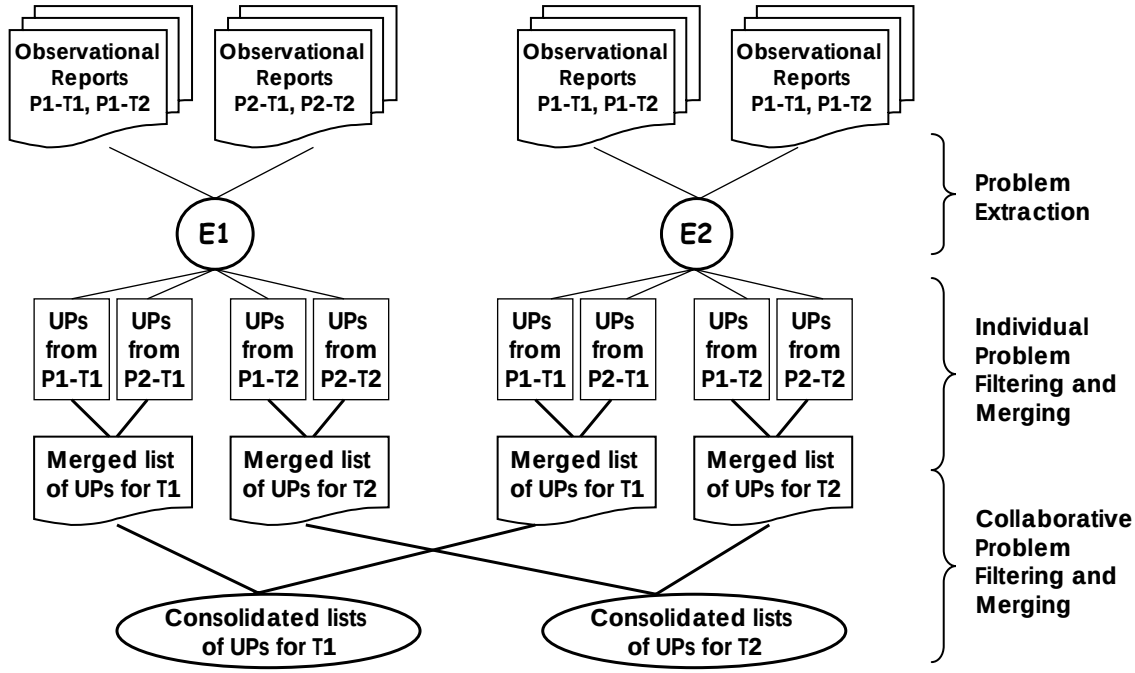

Figure 1: **The workflow of problem consolidating process**

3. RESULTS

3.1 Individual Problem Consolidation

The ten participants extracted from the observational reports altogether 98 and 81 UPs for T1 and T2 over the two users (P1 and P2), respectively. Furthermore, they individually consolidated their UPs. Table 1 shows the extent to which the participants merged, discarded and retained the UPs extracted.

Table 1. **Distribution of outcomes in the individual filtering**

	Merged	Discarded	Retained
T1	39%	13%	48%
T2	51%	10%	39%

For the merged and retained UPs, there were changes in severity ratings and/or confidence levels or no changes at all. To simplify the results, we collapse different degrees of increase/decrease (e.g. minor → moderate/severe or vice versa) into INC or DEC, respectively, and denote no change with SAME.

Table 2. **Severity/confidence changes in merged UPs (Indiv.)**

	Severity		Confidence	
	T1	T2	T1	T2
DEC	4 (10%)	3 (7%)	6 (15%)	4 (10%)
SAME	20 (53%)	29 (71%)	15 (40%)	18 (44%)
INC	14 (37%)	9 (22%)	17 (45%)	19 (46%)

The same notations are applied to the confidence level. In merging the UPs, the participants tended to increase the severity ratings by one or two degrees (i.e. 37% for T1 and 22% for T2; Table 2). In contrast, it seemed they did not bother to adjust the severity of the UPs retained (i.e., 2% and 6% for T1 and T2, respectively). In the post-filtering interviews, most participants explained that when a UP was both identified in P1 and P2, it could indicate that the UP was more severe than originally estimated and that it rectified the realness of the problem, thereby boosting their confidence. Interestingly, the correlation between the original severity ratings and confidence levels (r = 0.25, n = 179, p = 0.001) was found to be significant, implying that the participants were more confident that they judged the severe UPs correctly but less so when judging minor or moderate UPs. In contrast, the correlation between the changes in both variables (r = 0.19, n = 26) was insignificant. In other words, changing the severity of a UP does not imply that the participant has become more (or less) confident about the realness of the UP.

3.2 Collaborative Problem Consolidation

In comparison, the participants demonstrated an even stronger tendency to merge UPs in a collaborative setting (Table 3), which is higher than that (cf. 39% vs. 81% for T1; 51% vs. 77% for T2) observed in an individual session. The participants tended to negotiate at a higher abstract level where broad problem types can accommodate a variety of problem instances, thus mitigating direct confrontation with partners over controversial similarities. The participants tended to receptive to their partners' proposals, especially when the agreement thus reached would not cause any actual economic or personal gain (or loss). When negotiating to merge or retain UPs, the participants adjusted the severity and confidence ratings. For each aggregate we averaged the ratings of the original set of to-be-merged UPs and compared it with the corresponding final ratings. Table 4 displays the results for the merged UPs. Similar patterns to Table 1 were observed.

Table 3. **Distribution of outcomes in the collaborative filtering**

	Merged	Discarded	Retained
T1	81%	10%	9%
T2	77%	15%	8%

Table 4. **Severity/confidence changes in merged UPs (collab.)**

	Severity		Confidence	
	T1	T2	T1	T2
DEC	2 (5%)	2 (7%)	2 (5%)	3 (11%)
SAME	23 (52%)	16 (57%)	22 (50%)	13 (46%)
INC	22 (43%)	10 (36%)	19 (45%)	12 (43%)

4. DISCUSSION

The empirical findings of this study enable us to draw comparisons between the individual and collaborative UP consolidation processes, which presumably involve the core mechanism of judging similarity among UPs. One notable distinction is the lenience towards merging in the collaborative setting, as shown by the high merging rate. Indeed, quite a number of participants combined UPs that had not been merged in their individual sessions to merge with their partners'. It may be attributed to social pressure that coerces them to reach consensus. The data indicate that as a result of the merging process, severity ratings of UPs tend to inflate and the number of UPs tends to deflate excessively in the collaborative setting. In contrast, confidence levels, in which personal experience plays a role, do not fluctuate with the merging process. Previous research studies indicate that severity ratings influence how developers and project managers prioritize which UPs to fix ([3], [6]). Invalid severity ratings presumably lead to the fixing of less urgent UPs. Consequently, the quality of the system may still be undermined by more severe as well as more urgent UPs.

The implication for the future work is to look into relevant theories on similarity (an age-old issue), communication, and social interaction. Further, we aim to extend our empirical studies by systematically comparing merging through negotiation (i.e. the consolidation procedure is to be implemented by a group of two or three usability specialists or a group of developers or an integrated team) versus merging through authority (i.e. only one person-in-charge is to combine different lists of UPs). The quality of the consolidated usability outcomes will be compared, thereby enabling us to identify valid and reliable methods for consolidating UPs and to develop objective measures of the cost-effectiveness of such methods. Findings thus obtained will also contribute to our ongoing research endeavour on downstream utility.

5. REFERENCES

[1] Connell, I., & Hammond, N. (1999). Comparing usability evaluation principles with heuristics: Problem instances vs. problem types. *Proc. INTERACT 1999*.

[2] Hassenzahl, M. (2000). Prioritizing usability problems: data-driven and judgement-driven severity estimates. *Behaviour & Information Technology, 19*(1), 29-42.

[3] Hertzum, M. (2006). Problem prioritization in usability evaluation: From severity assessments toward impact on design. *International Journal of Human Computer Interaction (IJHCI), 21*(2), 125-146.

[4] Hertzum, M., & Jacobsen, N.E. (2003). The evaluator effect: A chilling fact about usability evaluation methods. *IJHCI, 15(*1).

[5] Howarth, J. (2007). *Supporting novice usability practitioners with usability engineering tools*. PhD thesis (VT).

[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* (IJHCI), 21(2), 147-172.

[7] Law, E. L-C., & Hvannberg, E. T. (2004). Analysis of combinatorial user effect in international usability test. *Proc. CHI 2004*

[8] Lewis, J.R. (1994). Sample sizes for usability studies: Additional considerations. *Human Factors*, *36*(2), 368-378.

[9] Nielsen, J. (1994). Heuristic evaluation. In J. Nielsen & R.L. Mack (Eds.), *Usability inspection methods*. New York: Wiley

[10] Virzi, R.A. (1992). Refining the test phase of usability evaluation: How many subjects is enough? *Human Factors*, *34*(4), 457-468