Novice Evaluators' Behavior when Consolidating Usability Problems Individually or Collaboratively

Ebba Thora Hvannberg
University of Iceland
Reykjavik, Iceland
ebba@hi.is

Effie Lai-Chong Law
University of Leicester
Leicester, UK
lcl9@leicester.ac.uk

ABSTRACT

An important, but resource demanding step in analyzing observations from usability evaluations is to consolidate usability problems (UPs) that were identified by several evaluators into one master list. An open question is whether consolidating UPs in pairs is cost-effective. A within-subject study examined if evaluators merge UPs differently when working in pairs than individually and what motivates their decisions. Eight novice evaluators took part. The number of discarded, retained and merged UPs, evaluators' confidence and severity of UPs in the two settings were measured. The results showed that UPs merged or discarded in the collaborative setting would rather be retained in the individual setting. Participants increased confidence and UP severity in the collaborative setting but decreased UP severity and confidence in the individual setting.

Author Keywords

Usability problems; Collaboration; Consolidation; Severity; Confidence; Criteria; Usability Evaluation; Novice; Group

ACM Classification Keywords

H.5.2. [User Interfaces]: Evaluation/Methodology

INTRODUCTION

An important, but resource demanding step in analyzing observations from usability evaluations is to consolidate usability problems (UPs) that were identified by several evaluators into one master list. When conducting usability evaluations an evaluator needs to extract a list of UPs from observations collected from each user. For each user, the evaluator needs to search for duplicates and filter them out. When all the user lists are ready, the evaluator or evaluators need to merge them into one.

Several methods have been suggested to generate output from usability evaluations in less time but with the same or better quality than traditional usability methods such as user testing. Instruments have included involving software developers and end users with minimum training [4], performing instant data analysis [13] and applying discount video data analysis [19].

For decades social scientists have studied group performance and interaction to understand if people make different decisions individually or in groups. Sauer et al. have referred to the theory as behavioral theory of group performance [16]. The studies, which are either within-subject or between-subject experiments, usually comprise two stages. In the first one individuals perform some task, make an assessment or a decision based on the material given to them. During the second stage, individuals meet in dyads or in groups to share information and repeat the task from the first stage. In research on team work, one of the questions is if working in groups is more effective than working individually [5, 6, 16, 18]. For example, the issue has been investigated by researching how team members with different power and control categorize concepts [6, 18] and by surveying the effect of team size, training and expertise when detecting software defects [16]. A noteworthy study was conducted by Heath and Gonzalez [7], who found that interactions increased people's confidence in their decisions about sports predictions, but was not supported by increased accuracy. Consistent with this finding Schuldt et al. [17] found that dyads were more confident in making decisions on tasks (i.e. deciding if statements are true or false) than individuals but that the pattern varied according to the confidence types of the dyads, i.e. low confidence individuals, high confidence individuals or mixture of low and high confidence individuals [17].

Hertzum et al. [9] reported a within-subject study of eleven participants where evaluators individually extracted UPs and then met in a group of three. The results showed that evaluators viewed the group work as multiple evidence in support of the same UPs and that the group work increased evaluators' confidence. To further study group evaluation, Hertzum et al. [10] studied how four groups of four or five evaluators merged UPs of critical severity, UPs of serious severity and bugs that they had extracted individually. The results showed that there was a substantial difference in which UPs were extracted between members of the group. A second result was that from the individual UP extraction

EICS '17, June 26-29, 2017, Lisbon, Portugal
© 2017 Association for Computing Machinery.
ACM ISBN 978-1-4503-5083-9/17/06…$15.00
http://dx.doi.org/10.1145/3102113.3102129

until the merging by group, 17% of the issues were discarded from the list of UPs and the severity of another 34% of the UPs were decreased from critical, serious or bugs to minor severity. The recommendation made as a consequence of this result was to have groups consolidate severity of UPs.

An open question is whether merging UPs and assessment of their severity in pairs is worth the additional resources. In the area of user interface evaluation, research studies have been conducted on the effectiveness of working in groups vs. individually. Law and Hvannberg [14] reported on a within-subject study comparing how individual novices merged UPs to novices collaborating in pairs. The main results of that study was that novices tended to merge more UPs when collaborating compared with working individually and that the severity of the UPs tended to increase in the collaborative setting. In comparison to problem extraction, the severity of merged UPs increased as a result of the activities individual filtering and collaborative merging. Similar results were found for participants' assessment of confidence in their rating a problem as a UP. While these results of severity rating were contrary to the results of Hertzum et al. [10] who found in a within-subject study that the UP severity tended to be rated lower in groups, they harmonized with the previous results that individuals in a group increased their confidence [7]. Hertzum and Jacobsen [8] gave an overview of studies on evaluators' agreement on severity and found that their agreement varied from 20% to 28%, and that evaluators' agreement on a variety of characteristics varied from 5%-65%. Therefore, Hertzum and Jacobsen [8] suggested that an extra evaluator resource might improve the reliability of the results. The variability in severity ratings has been confirmed in a more recent study [10]. In a within-subject study, Brajnik, et al. [3] compared novices reviewing conformance to Web Content Accessibility Guidelines 2.0 individually to novices working in pairs. The results showed that groups were better in identifying all the true UPs, but given some tolerance (11% in a validity measure that includes correctness and sensitivity), then the overall effectiveness of individuals were as good as pairs. Thus, the improved accuracy of groups in the study of Brajnik, et al. [3] was inconsistent with that of Heath and Gonzalez [7]. From the above reviews we observe that the results are contradictory at least in some aspects. Furthermore, none of these studies have attempted to find relationships between severity and confidence vs. evaluators' decision to discard, retain or merge a UP during the merging process.

Motivated by the need to understand processes of merging, the following four questions were posed. We hoped to understand if evaluators merge UPs differently in pairs compared to when working individually:

R1: Do evaluators filter or merge more UPs in a collaborative setting than when working individually?

R2: Do evaluators become more confident in their decisions in a collaborative setting compared to working individually? Do evaluators increase the severity of UPs in a collaborative setting compared to when working individually?

R3: Are there relationships between severity of a UP and evaluator's confidence of their decision regarding a UP before and after problem consolidation?

R4: Can severity or confidence predict whether a UP is merged, discarded or retained?

The study was carried out by asking participants to merge UPs individually and collaboratively. This study is a replication of that reported in Law and Hvannberg [14] with an additional research question R4.

RESEARCH STUDY

Research protocol

For the study, eight participants were recruited among bachelor and masters students of computer science and software engineering. They all had at least one course in human computer interaction, including skills on design, design guidelines and usability evaluation. Initially, we recruited ten participants, but two of them could not finish the study. The number of participants was on the low end but it was a within-subject study. The participants were novices from a homogenous group, diminishing the need for a large number of participants. Furthermore, as is customary in analysis of such studies the UPs extracted were the units of study and some of the statistical tests were significant. When Lewis [15] explained the cost effectiveness and different expected results in having many vs. few participants, he noted that e.g. Bailey [1] had eight participants in a between-subject study where he compared three different groups that used three different prototypes and received significant results.

Prior to this study, a usability evaluation was conducted on a Learning Management System (LMS), using the think-aloud protocol. In each of these usability sessions, users were asked to carry out two tasks: Browsing a catalogue of learning resources and Providing a learning resource. As an output of that protocol, with two users carrying out two tasks each, four text documents and screen capturing videos from the sessions resulted. These materials were given to participants of this study. This procedure may seem artificial to usability work, but since participants extracted the problems themselves based on the text protocol and videos, they did get a good idea of the context. A similar protocol was used by [11] and [10] where participants extracted problems from three and five video files respectively. This design was chosen in the original study [7] to minimize bias.

After a pre-study training meeting on usability evaluation and familiarization of the system under evaluation, participants were asked to attend twice with one week apart.

The main study comprised three sessions. In the first one, participants worked individually and extracted problems from the text and videos given to them. Participants were given a list of six criteria [12] which they used to help them determine if an issue was indeed a UP. From the protocol:

> A user aims to achieve a sub-goal of a given task; he/she articulates the intention or is interpreted to have it through his/her action), but he/she: **C1**: cannot continue without external help.**C2**: tries several things and then explicitly gives up.**C3**: fails to achieve it or gets a wrong output.**C4**: commits an error that makes him/her pause for thought before he/she can continue (i.e. the duration of the pause is an indicator of problem severity).**C5**: expresses frustration, anger or surprises.**C6**: makes some negative comments on an interface element or proposes a design alternative.

They were asked to rate the UPs' severity (minor, moderate, severe) and their own confidence (five points, from very low to very high) in their decision to extract a UP. In the second session, participants worked again individually and filtered duplicate problems, i.e. from the two users. Again, they were asked to rate the problems' severity and their own confidence in their decision regarding filtering. In one week's time the participants were asked to return and work in pairs to merge the UPs across two users. UPs are only filtered or merged within tasks, and not between tasks. Once again, the participants were asked to re-rate the severity and their confidence in their decision to merge, discard or retain problems. The within-subject design was chosen to best match the way evaluators work, i.e. first extracting problems alone and then pairing with another evaluator to merge problems.

Research Model

In this study, the process of UP extraction and merging has three steps: Problem extraction (PE), Individual Filtering (IF) and Collaborative Merging (CM). Besides the textual description of a UP, each problem extracted is characterized by four variables: Severity of the problem, participant's Confidence in his/her decision, Criteria for a problem and the problem's Fate during filtering or merging. The first three variables have already been described. The variable Fate describes whether an evaluator decides to retain a problem, discard it or merge it with one or more other problems. Furthermore, the variables SeverityChange and ConfidenceChange take on the values DEC, SAME and INC denoting that, from problem extraction until after filtering or merging, severity of a problem or a participant's confidence in his or her decision has decreased, stayed the same or increased.

RESULTS

Problem extraction

Eight participants extracted 71 problems. Over half of the problems were rated severe (54%), a quarter moderate (25%) and one fifth was rated minor (20%). Participants felt they were confident (high, very high) in their rating for less than half of the problems (44%), medium confident for 24% of the problems and less confident (low, very low) for 33% of the problems.

After problem extraction, we were interested in seeing if there was a relation between participants' confidence and problem severity. Since the variables Confidence and Severity are ordinal we could compute Spearman's correlation (ρ=.483, p=.000***, N=71). The results showed that there was a significant correlation between Confidence and Severity. This means that high severity was rated with high confidence and low severity tended to be rated with low confidence.

Changes in Severity and Confidence

From Problem Extraction to Individual Filtering

Participants did not change their rating of severity significantly after individual filtering (ρ=.529, p=.077, N=12), but changed their confidence significantly at the .05 level (ρ=.600, p=.039, N=12). Correlations between changes in both variables were insignificant (Table 1, (IF)). Examining the relationship between severity and confidence again, the results showed that the correlation found after individual filtering was similar after problem extraction, with Spearman's ρ=.423, N=68 and p=.000***.

Table 1 Comparing changes: individual filtering vs. collaborative merging of merged problems only

	Severity (IF)	**Severity (CM)**	**Confidence (IF)**	**Confidence (CM)**
DEC	25%	14%	17%	2%
SAME	75%	70%	58%	42%
INC	0%	16%	25%	56%

From Problem Extraction to Collaborative Merging

For collaborative merging, evaluators rated their confidence in the merged UPs. We were interested in seeing if there were differences in severity and confidence between problem extraction and collaborative merging. For this analysis we used a K related Friedman test, a non-parametric alternative to repeated measure ANOVA (Table 2). Only merged UPs were considered. The results showed that confidence increased significantly as a result of collaborative merging but not severity. While participants assessed confidence again after the collaborative filtering, severity was computed from the averaged severity of the original to-be-merged UPs.

Table 2 Change in characteristics from problem extraction to collaborative merging

	Mean Rank PE	**Mean Rank CM**	**Friedman's Q**	**N**	**p**
Severity	1.49	1.51	.040	43	.841
Confidence	1.23	1.77	22.154	44	.000***

Comparing Individual Filtering with Collaborative Merging

In the two previous subsections we looked at changes in severity and confidence between problem extraction and to individual filtering and collaborative merging. Statistical analysis using Fisher exact test showed that the increase or decrease in severity or confidence were not significant with respect to the two settings, individual and collaborative. For severity the Fisher exact test gave p=.110, N=55 and for confidence it gave p=.073, N=55. Table 1 shows the comparison between the two activities: individual filtering and collaborative merging with respect to relative changes in each of severity and confidence.

Characteristics that Influence the Fate of a Problem

So far, we have looked at how severity and confidence change after individual filtering and collaborative merging. These results are useful for learning about the difference between individual filtering and collaborative merging. In addition, it would be valuable to know if characteristics of problems could predict participants' decisions to discard, retain or merge problems. We built a generalized linear mixed model fit by maximum likelihood (Laplace Approximation) to see how severity, confidence and the two settings (individual, collaborative) could predict the fate of a problem. Random effects are UP and Participant ID. Fixed effects were Group (individual, collaborative), Severity (minor, moderate, sever) and Confidence (very low, low, moderate, high, very high), all categorical variables (factors). The response variable was Fate (discarded, merged, retained). We ran the analysis in two parts. First Discarded vs. Retained and then Merged vs. Retained. A binomial (logit) distribution was used. The computations were done in R, using the lme4 package [2].

The only characteristic that could significantly predict Retained vs. Discarded problems was Group, where problems in the collaborative setting were less likely to be retained than in the individual setting (OR .10; 95% CI .02-.52), N=80 (see Table 3). We used Likelihood Ratio test to test significance between models. Comparing a model with the Group as predictor with the null model (no predictors) we found that the Group attribute (individual or collaborative) did affect the fate of the UP (χ^2 = 8.83, p=.00296**). The BICs (Bayesian Information Criterion) for the two models were 147.3 vs. 185.14 for the null model.

Table 3 Group as a predictor for Retained vs. Discarded

Fixed effects	**OR (95% CI)**
Intercept	28.00 (6.83, 114.737) ***
Group	
Individual	1.00 (referent)
Collaborative	.10 (.02, .52) ***

*** p < .001

Similarly, the only characteristic that could significantly predict Retained vs. Merged problems was Group, where problems in the collaborative setting were less likely to be retained than in the individual setting (OR .07; 95% CI .03-.18), N=128 (see Table 4). In this model we dropped the random effect of UP since the model did not converge because of singularity. We used Likelihood Ratio test to test significance between models. Comparing the model with the Group as predictor with the null model (no predictor) we found that the Group attribute (individual or collaborative) did affect the fate of the UP (χ^2 = 42.87, p=0***). The BICs (Bayesian Information Criterion) for the two models were 60.71 vs. 65.16 for the null model.

Table 4 Group as a predictor for Retained vs. Merged

Fixed effects	**OR (95% CI)**
Intercept	4.86 (2.41, 9.81) ***
Group	
Individual	1.00 (referent)
Collaborative	.07 (.03, .18) ***

*** p < .001

We ran models with all variables simultaneously as predictors and subsets thereof, e.g. Group, Severity and Confidence. We opted for using multinomial logistic regressions without any random effects. Significance was measured using a Wald test. Neither severity nor confidence alone had a significant effect on the fate of a UP. We examined the odds ratio between moderate and severe UPs vs. minor ones. The results showed that compared to minor UPs, moderate ones had almost the same odds to be merged vs. retained (OR .91, 95% CI .34-2.44, p=.85) but severe UPs had less odds to be merged (OR .49, 95% CI .16-1.55, p=.23).

None of the conditions of Confidence (five levels) were significantly different for either Merged or Discarded vs. Retained. Figure 1 shows the results of analysing effects (probabilities) and confidence intervals of Confidence on the Fate of a problem. Since no UPs are discarded and of confidence in the fourth category (Confidence=high, level 4) the confidence interval becomes very high. For high confidence (level 4) the UPs are less likely to be merged than retained (OR .024, CI .05-1.07, p=.06).

We ran a multinomial logistics regression model between the interaction of Group and Severity. The results showed a significant model for Group interacting with Severity predicting the Fate of the problem (χ^2=11.375, p=.02). Figure 2 shows the effects (probabilities and confidence intervals). As an example, UPs with minor severity have a higher probability to be merged in the collaborative setting than in the individual setting. One would expect that evaluators would retain severe UPs over minor ones. The merged problems in the individual setting break this pattern.

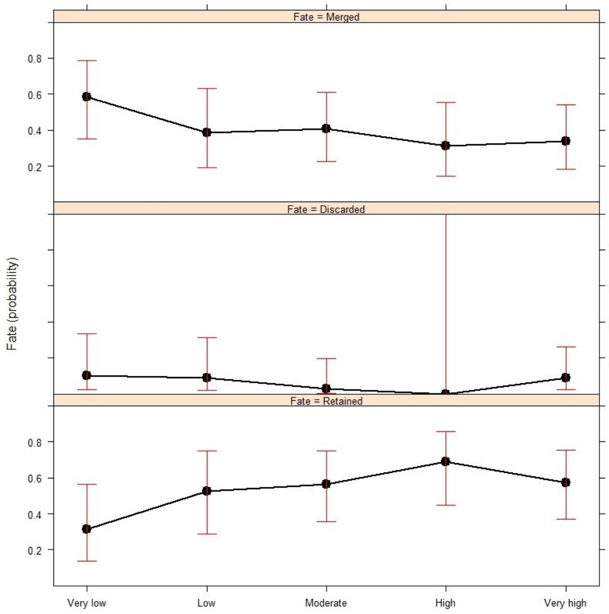

Figure 1 Effect of Confidence on Discarded, Merged or Retained

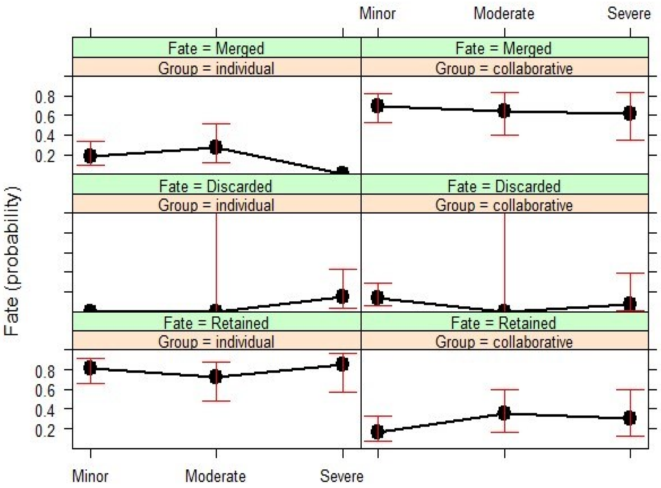

Figure 2 Effects of Interaction of Group and Severity

In the individual setting minor UPs have lower probabilities of being merged than the moderate ones, but in the collaborative setting, they follow the expected pattern and have a higher probability of being merged than the moderate ones. In the collaborative setting severe UPs have a lower probability of being retained than moderate ones.

SUMMARY AND DISCUSSION

Here we summarize answers to the research questions posed earlier:

R1: Do evaluators filter or merge more problems in a collaborative setting than when working individually?

Evaluators tended to merge and discard more problems in the collaborative setting compared to retaining them in the individual setting. Comparing the results to [14] we see similar trends except in their case there was not a tendency to discard more problems in the collaborative setting.

The settings, individual vs. collaborative could significantly predict the fate of the problem.

R2: Do evaluator raise confidence in their decisions or increase severity in a collaborative setting compared to when working individually?

Analysis showed that there were no significant differences between changes in confidence or severity between the two settings individual or collaborative. Evaluators increased their confidence considerably in the collaborative setting but changes in severity were insignificant. The changes in confidence after individual filtering were significant but insignificant for severity. Evaluators had a tendency to decrease severity of problems and lowered or raised confidence of some decisions in the individual setting. Whereas in this study there seems different patterns between individual and collaborative settings, in Law and Hvannberg [14] evaluators showed similar patterns in collaborative merging vs. individual filtering, i.e. that of increasing confidence and severity.

R3: Are there relationships between severity and confidence of a problem before and after problem consolidation?

Severity had a moderate effect on confidence before and after individual filtering, meaning that low severity tends to imply low confidence and high severity high confidence. Same results were found in [14].

R4: Can severity or confidence predict whether a problem is merged, discarded or retained?

Neither severity nor confidence alone could significantly predict discarded problems over retained. Group and Severity together could significantly predict the fate of a problem. However, more research is needed to see if these results are stable or if more data or predictors need to be collected.

The main limitations of this study were the number of participants and their novice background. However, we have replicated another study [14] and novices are seen as important evaluators in usability evaluation [4].

CONCLUSION

This study has contributed to the question whether it is worth spending the additional resources on consolidating usability problems. In agreement with previous work, the study has shown that during collaboration many problems may be aggregated together, making individual problems invisible to designers during redesign. Further studies are needed to study whether this has a negative downstream effect or the positive effect of simplifying the results to designers.

Another result of this study is that evaluators working together did not excessively discard problems. If collaborating evaluators increase the severity of problems as was the case in this study, it may give problems undeservedly higher priority during redesign.

That evaluators raised confidence when working in pairs may indicate that the scheme could be a way for training novices, regardless of their HCI skills. This contribution may synergize well with recent research that suggests that it may be resource efficient to have others than HCI experts perform usability evaluation [4]. Working individually and then in pairs during training may encourage novice evaluators to reflect on their behavior and make them more conscious that they might treat UPs differently in the two settings.

REFERENCES

1. Gregg Skip Bailey. 1993. Iterative methodology and designer training in human-computer interface design. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*, (1993), ACM, 198-205.
2. Douglas Bates, Martin Mächler, Ben Bolker and Steve Walker. 2014. Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.
3. Giorgio Brajnik, Markel Vigo, Yeliz Yesilada and Simon Harper. 2016. Group vs Individual Web Accessibility Evaluations: Effects with Novice Evaluators. *Interacting with Computers*, *Online, Advance access*. http://iwc.oxfordjournals.org/content/early/2016/04/25/iwc.iww006.abstract
4. Anders Bruun and Jan Stage. 2015. New approaches to usability evaluation in software development: Barefoot and crowdsourcing. *Journal of Systems and Software*, *105*. 40-53. http://doi.org/10.1016/j.jss.2015.03.043
5. Asbjörn Fölstad. 2008. The effect of group discussions in usability inspection: a pilot study *Proceedings of the 5th Nordic conference on Human-computer interaction: building bridges*, ACM, Lund, Sweden, 2008, 467-470. http://doi.acm.org/10.1145/1463160.1463221
6. Rebecca W. Hamilton, Stefano Puntoni and Nader T. Tavassoli. 2010. Categorization by groups and individuals. *Organizational Behavior and Human Decision Processes*, *112* (1). 70-81. http://www.sciencedirect.com/science/article/pii/S0749597810000142
7. Chip Heath and Rich Gonzalez. 1995. Interaction with others increases decision confidence but not decision quality: Evidence against information collection views of interactive decision making. *Organizational Behavior and Human Decision Processes*, *61* (3). 305-326.
8. Morten Hertzum and Niels Ebbe Jacobsen. 2001. The Evaluator Effect: A Chilling Fact About Usability Evaluation Methods. *International Journal of Human–Computer Interaction*, *13* (4). 421-443. http://dx.doi.org/10.1207/S15327590IJHC1304_05
9. Morten Hertzum, Niels Ebbe Jacobsen and Rolf Molich. 2002. Usability inspections by groups of specialists: perceived agreement in spite of disparate observations *CHI '02 Extended Abstracts on Human Factors in Computing Systems*, ACM, Minneapolis, Minnesota, USA, 2002, 662-663.
10. Morten Hertzum, Rolf Molich and Niels Ebbe Jacobsen. 2014. What you get is what you see: revisiting the evaluator effect in usability tests. *Behaviour & Information Technology*, *33* (2). 144-162.
11. Kasper Hornbæk and Erik Frøkjær. 2008. Comparison of techniques for matching of usability problem descriptions. *Interacting with Computers*, *20* (6). 505-514.
12. Bonnie E. John and Matthew M. Mashyna. 1997. Evaluating a Multimedia Authoring Tool. *Journal of the American Society for Information Science*, *48* (11). 1004-1022. http://doi.org/10.1002/(SICI)1097-4571(199711)48:11<1004::AID-ASI4>3.3.CO;2-T
13. Jesper Kjeldskov, Mikael B Skov and Jan Stage. 2004. Instant data analysis: conducting usability evaluations in a day. In *Proceedings of the third Nordic conference on Human-computer interaction*, (2004), ACM, 233-240. http://doi.org/10.1145/1028014.1028050
14. Effie Lai-Chong Law and Ebba Thora Hvannberg. 2008. Consolidating usability problems with novice evaluators. In *Proceedings of the 5th Nordic conference on Human-computer interaction: building bridges*, (2008), ACM, 495-498. http://doi.org/10.1145/1463160.1463228
15. James R Lewis. 2014. Usability: lessons learned… and yet to be learned. *International Journal of Human-Computer Interaction*, *30* (9). 663-684.
16. Chris Sauer, D. Ross Jeffery, Lesley Land and Philip Yetton. 2000. The effectiveness of software development technical reviews: a behaviorally motivated program of research. *IEEE Transactions on Software Engineering*, *26* (1). 1-14. http://doi.org/10.1109/32.825763
17. Jonathon P Schuldt, Christopher F Chabris, Anita Williams Woolley and J Richard Hackman. 2015. Confidence in Dyadic Decision Making: The Role of Individual Differences. *Journal of Behavioral Decision Making*.
18. Pamela K. Smith, Rachel Smallman and Derek D. Rucker. 2016. Power and Categorization: Power Increases the Number and Abstractness of Categories. *Social Psychological and Personality Science*, *7* (3). 281-289. http://doi.org/10.1177/1948550615619760
19. Jody Wynn and Jeremiah D Still. 2011. Motivating Change and Reducing Cost with the Discount Video Data Analysis Technique. In *International Conference of Design, User Experience, and Usability*, (2011), Springer, 321-328. http://doi.org/10.1007/978-3-642-21708-1_37