Outliers in usability testing: How to treat usability problems found for only one test participant?

Asbjørn Følstad
SINTEF
P.O. Box 124, Blindern,
N-0134 Oslo, Norway
asf@sintef.no

Effie Lai-Chong Law
Depart. of Computer Science
University of Leicester,
Leicester, LE1 7RH, UK
elaw@mcs.le.ac.uk

Kasper Hornbæk
Department of Computing,
University of Copenhagen,
Njalsgade 128, DK-2300
Copenhagen S, Denmark
kash@diku.dk

ABSTRACT

In usability testing, usability problems are often found for only one test participant. The literature does not help in deciding whether such single-user problems should be accepted or rejected as usability problems. To help us understand how such decisions are made in practical usability testing, 89 practitioners described how they dealt with single-user problems in their latest usability test. Single-user problems was accepted, rejected, or reported as outliers. This decision depended on problem severity, participant profile, sample size, and judgments on whether the problem is an artifact of the test situation.

Author Keywords
Usability testing, single-user problems, questionnaire survey, usability practitioner

ACM Classification Keywords
H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION

Identifying usability problems is a key motivation for usability testing. The quality of usability problem identification depends on (a) identifying a sufficiently high proportion of relevant usability problems and (b) minimizing the number of false alarms, that is, avoiding interpreting non-relevant problem instances as usability problems [5]. Consequently, for any potential usability problem a usability professional needs to weigh the potential benefit of including the problem in the final report against the risk of making a false alarm.

For the purpose of this paper, we distinguish three phases of usability problem identification. (1) Collecting data about the test participants' interaction (e.g., thinking aloud, observations of body language, logging of interaction). (2) Identifying potential usability problems; sometimes this is straightforward, sometimes it requires combination of data and interpretation. (3) Problem merging, where usability problems are combined, consolidated, and prioritized across a test. In this study we address how to deal with usability problems identified in the second phase, but only backed up with data from a single participant. We refer to such problems as *single-user problems*.

Concluding on single-user problems is challenging as their relevance and validity are difficult to assess, in particular for small samples. Even so, the literature provides little guidance on how to deal with single-user problems. This lack of guidance is critical as single-user problems can represent a nontrivial proportion of the usability problems reported from a usability test. For example, Law and Hvannberg [9] reported that 41 of 88 usability problems identified in a usability test with 17 participants were found for one user only. Similarly, Nielsen and Landauer [11] reported that in their 15-participant usability test of the Office system, 77 of 145 usability problems were single-user problems.

We present the findings from a survey on how practitioners decide whether incidents observed for only one test participant are to be considered usability problems. On this basis we suggest guidelines for dealing with single-user problems.

BACKGROUND

Usability testing is often considered the gold standard of usability evaluation methods [7]. Dumas and Fox [1] even claim that they are unaware of studies that question the validity of usability tests. How, then, can single-user problems represent a challenge? Two reasons may be given, the first concerns relevance, the second concerns validity.

First, single-user problems may reflect highly infrequent usability problems. Imagine a usability problem that is just as likely to be detected with any one user. If the problem is detected by only one in 15-test participants a best estimate of the frequency of this problem to occur in the user population is .12 (LaPlace Method) and a 95% confidence interval of the population frequency is .00-.32 (Adjusted Wald) [14,15]. The problem may well be so infrequent that it would be considered irrelevant in many development projects.

Second, and more important, a problem instance detected by just a single test participant may be an artifact of the test situation. For example, if the test participant misunderstood the task instructions or is not representative of the user population.

NordiCHI '12, October 14-17, 2012 Copenhagen, Denmark
Copyright © 2012 ACM 978-1-4503-1482-4/12/10... $15.00"

Such a problem instance should most likely not be interpreted as a usability problem. As stated by Nielsen [10] *"there is always a risk of being misled by the spurious behavior of a single person who may perform certain actions by accident or in an unrepresentative manner"*.

The literature on usability evaluation contains different and conflicting views on how to deal with single-user problems. In their study of Instant Data Analysis, Kjeldskov et al. [8] argued for "*viewing [...] unique problems as noise rather than 'real' usability problems*". Taking the opposite stance, Woolrych and Cockton [17], in a 12-participant stress-test of problem predictions from Heuristic evaluations, treated their five single-user problems as reflecting real usability problems. In their text book on usability testing, Dumas and Redish [2] made a short discussion of single-user problems where they recommend reporting these as outliers. We are not aware of thorough discussions in the literature that can advise practitioners on the factors that determine how to deal with single-user problems, such as sample size, estimated problem frequency, and possible test-specific issues.

METHOD

To gain knowledge on how single-user problems are dealt with, we explored the practices of a large number of usability practitioners. We collected data as part of a larger online survey on analysis in usability evaluation [4]. We asked usability practitioners to give a free-text response to the question: "*If an incident was observed with only one of the users participating in your latest usability test, how did you decide whether this was a usability problem or not?*"

Usability practitioners were invited to participate in the survey in the following ways: E-mail invitations distributed in local SIGCHI and UPA chapters, to an international mailing list for usability practitioners, to usability experts in the European COST project TwinTide, on social media, and in CHI'11 fliers. The respondents to our question had conducted at least one usability test within the last six months and reported on their latest usability test only. As incentives, the respondents were included in a lottery of a USD 250 gift card and were promised a report of the survey findings.

In total, 89 respondents answered our question on single-user problems. The respondents had a median of 6 years working experience as usability practitioners (25th percentile = 4 years; 75th percentile = 12 years). They worked in 17 different countries (40 in the US, 41 in Europe and 8 in other parts of the world). The usability tests on which they reported involved a median of 8 test participants (25th percentile = 6; 75th percentile = 13).

The respondents' answers were broken down into a total of 134 items and coded in a thematic analysis [3]. The items were coded independently by two analysts (the first and second authors of this paper). The second coding was done with a modified version of the code set used in the first; this code set was then mapped onto the code set of the first coding. Free-marginal *kappa* for inter-rater agreement was .77 which is regarded as adequate [12].

RESULTS

The respondents' answers concerned (a) resources and strategies used to reach a judgment on single-user problems, (b) relevant conditions to consider when making the judgment, and (c) potential outcomes of the judgment. These three clusters of descriptions are treated below.

Resources and strategies

Sixty-three items referred to resources and strategies used to interpret single-user problems (see Table 1).

The most frequently mentioned resource was general references to the respondents' *"professional judgment and past history in the field"*. This highlights the importance of experience when making judgments in usability evaluation, and also serves to explain the importance attached to discussing single-user problems with experts or *"within the project team"*.

Respondents also frequently checked against other sources of usability knowledge. Some reported subjecting particularly interesting single-user problems *"to evaluation with a greater number of users"*. Others reported making a *"review against previous tests"* or checking *"against design patterns*" or *"heuristics and guidelines"*. The latter is particularly interesting because it refers to usability evaluation resources typically used in usability inspection.

A few respondents made general notes of estimating the probability of recurrence in other users, but without explaining how this was done. *"We tried to decide if others would be confused by the same issue, or if there was something unique about that individual"*. Such estimation may be in line with Sauro and Lewis' [14] recommendation to report an estimate of the problem probability with a 95% Adjusted Wald confidence interval.

Resources and strategies	Freq.
The practitioners' own professional knowledge and experience	20
Discussing with experts or team members	9
New or extended evaluations	9
Checking against heuristics, guidelines, or design principles	8
Checking against previous evaluation results	5
Estimating the probability of recurrence with other users	4
Following a specific process or policy	3
Comparing the problem instance against assumptions or previous experiences	2
Discussion with test participant	2
Identifying a solution	1
Total	**63**

Table 1. Frequency of items across resources and strategies.

Other respondents described how they used a specific process or policy to deal with single-user problems. One respondent wrote: *"A kind of simple decision tree: a. Any other user had it? (you already said no) b. Is it a "logical" error (ie, other users may think/understand the same)? If yes, fix it. If no... c. Does it have severe conse[q]uences? If yes, check with more users, if no, let's pay attention to this possibility in the future."*

The three least frequently reported resources or strategies provide valuable insights. Two respondents mentioned the importance of comparing single-user problems with assumptions or previous experiences, as an expected problem instance is more likely to be interpreted as a usability problem than a non-expected instance. *"If it fitted previous experiences, it was included"*. This line of reasoning makes sense from a hypotheses-testing point-of-view, as hypotheses make it less likely that a finding is spurious.

Two other respondents mentioned the test participant as a resource that can help explain the source of the problem instance. *"Usually on the post-test discussion we can ask more on the detail of the accident and decide whether or not that has been an usability issue"*. This strategy is in line with text-book recommendations on using debriefing sessions to get additional information on observations made during the test [13].

A single respondent reported that the identification of a fix will be important when judging single-user problems. This line of reasoning suggests that the evaluation should be closely integrated with design and serve as a means for generating design ideas [6].

Relevant conditions

Relevant conditions for the interpretation of single-user problems were treated in 49 of the 134 items (see Table 2).

The respondents frequently reported that single-user problems of high severity (*"if it was a particularly bad issue"*) are likely to be judged as usability problems. It seems reasonable to discard low-severity single-user problems, as it is unlikely that these reflect important usability problems. At the same time, although none of the respondents reported this, it should be noted that high-severity single-user problems may also be irrelevant; they may be artifacts of the test situation or they may be highly infrequent in the user population. As noted by Turner, Lewis, and Nielsen [16], problem frequency and severity are not correlated.

Relevant conditions	Freq.
Severity / consequence of the problem instance	18
Test participants' profile	9
Sample size	6
Artifact of the test situation?	6
Task importance	5
Other conditions (misc.)	5
Total	**49**

Table 2. Frequency of items across relevant conditions.

The test participant profile was also a frequently mentioned condition, and was used in two lines of argument: Either, a single-user problem is more likely to be interpreted as a usability problem if the test participant's profile is closely aligned with the primary user group. Or, if the test participant is one of a few representatives of a particular user group a single-user problem may trigger further investigations for this particular user group; *"if the participant was underrepresented as a persona type we will evaluate the issue further."*

Sample size was reported to be an important consideration as single-user problems in small test samples were more likely to be judged as usability problems than single-user problems in large samples. *"With a small sample, I usually include it"*. This consideration makes sense from point of view of the binomial distributions of usability problem identification, as the best estimate for problem frequency in a small sample will be higher than the best estimate of the same in a large sample [14].

The respondents described that single-user problems may be *"an artifact of the test design"* or test situation in different ways. In particular it is relevant whether or not (a) the test participant is *"paying attention to the test"* or displays odd behavior throughout the test situation, and (b) things such as technical breakdowns happened during the test session.

It is interesting, yet puzzling, how some respondents mention task importance; we would assume all tasks included in a usability test to be important. However, it may mean that single-user problems in, for example, a warm-up task is less likely to be interpreted as a usability problem than such problem instances in other tasks. It could also be that the task was user-defined rather than predefined by the evaluator (i.e., open vs. structured test design).

Potential outcomes

Twenty-four percent of the respondents commented on the specific outcome of their judgment of single-user problems (see Table 3). Four respondents are in line with Dumas and Redish's [2] recommendation to report single-occurrence instances as outliers. *"I included it on a list of findings, but specified that only one user encountered it"*.

No respondents reported to include estimates of the probability of single-user problems occurring in the user

Potential outcomes	Freq.
Accept	8
Classify as low priority	4
Record as outlier	4
Reject	6
Total	**22**

Table 3. Frequency of items across potential outcomes.

population in the problem report. However, as noted above, four respondents mentioned to have made such estimates.

DISCUSSION AND RECOMMENDATIONS

The survey show varied practices in handling single-user problems. We see that practitioners are aware of relevance and validity issues related to single-user problems, possibly to a greater degree than the current literature in which, according to Dumas and Fox, no studies have questioned the validity of usability testing [1].

Based on the respondents' answers we propose five recommendations on how to deal with single-user problems.

(1) Establish a procedure for handling single-user problems. Such a procedure may include details on strategy, knowledge resources, relevant considerations, and possible outcomes.

(2) Sample size is important. Single-user problems in small samples should be given more weight than such problem instances in large samples. Estimate the problem probability, including its confidence interval [14].

(3) Check single-user problems against knowledge resources such as heuristics and guidelines, as well as against results from previous evaluations. Consider making extended evaluations to better understand particularly relevant single-user problems.

(4) Seek advice. Use the knowledge and experience of experts and team members. Use debriefing sessions to access the test participants' interpretations and explanations.

(5) Be aware that single-user problems sometimes do not reflect actual usability problems. Consider whether the problem can be interpreted as an artifact of the test situation. Use a low threshold for rejecting low-severity single-user problems.

LIMITATIONS AND FUTURE WORK

The findings and recommendations of this study are limited. In particular, as the recommendations are based on existing practices rather than research-based knowledge, we do not know how effective they are. If current practice is suboptimal, this will also be the case for the recommendations. Future research is needed to validate the findings and recommendations in empirical studies.

Even so, the findings and recommendations serve as useful input for usability practitioners on how to interpret single-user problems in usability testing. We also hope they will spur discussion on this virtually unexplored topic in usability testing and motivate future research.

ACKNOWLEDGMENTS

This work was done within the R2D2 Networks project, supported by the Norwegian Research Council program VERDIKT, and the COST Action TwinTide: IC0904.

REFERENCES

1. Dumas, J.S., Fox, J.E. Usability testing: Current practices and future directions. In A. Sears, J. Jacko (eds.) *The Human-Computer Interaction Handbook*, 2nd edition, Lawrence Erlbaum Associates (2008), 1129-1150.
2. Dumas, J.S., Redish, J. *A practical guide to usability testing*. Ablex Publishing Corporation, 1993.
3. Ezzy, D. *Qualitative Analysis: Practice and Innovation*. Routledge, 2002.
4. Følstad, A., Law, E.L-C., Hornbæk, K. Analysis in Practical Usability Evaluation: A Survey Study. In *Proc. CHI '12*, ACM Press (2012), 2127-2136.
5. Hartson, H.R., Andre, T.S., Williges, R.C. Criteria for Evaluating Usability Evaluation Methods. *International Journal of Human-Computer Interaction 13*, 4 (2001), 373-410.
6. Hornbæk, K. Usability Evaluation as Idea Generation. In E.L-C. Law, E.T. Hvannberg, G. Cockton (Eds.) *Maturing Usability: Quality in Software, Interaction and Value*, Springer (2008), 267-286.
7. Hornbæk, K. Dogmas in the assessment of usability evaluation methods. *Behaviour & Information Technology 29*, 1 (2010), 97-111.
8. Kjeldskov J., Skov M. B., Stage J. Instant Data Analysis: Evaluating Usability in a Day. In *Proc. NordiCHI '04*, ACM Press (2004), 233-240.
9. Law, E.L.-C. Hvannberg, E.T. Analysis of Combinatorial User Effect in International Usability Tests. In *Proc. CHI '04*, ACM Press (2004), 9-16.
10. Nielsen, J. Why You Only Need to Test with 5 Users, *Jakob Nielsen's Alertbox*, March 19, 2000, http://www.useit.com/alertbox/20000319.html
11. Nielsen, J., Landauer, T.K. A mathematical model of the finding of usability problems. In *Proc. CHI '93*, ACM Press (1993), 206-213.
12. Randolph, J.J. *Online Kappa Calculator*. http://justus.randolph.name/kappa
13. Rubin, J. and Chisnell, D. *Handbook of usability testing* (2nd.edition). Wiley Publishing, 2008.
14. Sauro, J, Lewis, J.R. Estimating completion rates from small samples using binomial confidence intervals: Comparisons and recommendations. In *Proc. HFES*, (2005), 2100-2104.
15. Sauro, J. Confidence Interval Calculator for a Completion Rate. http://www.measuringusability.com/wald
16. Turner, C.W., Lewis, J.R., Nielsen, J. Determining usability test sample size. In W. Karwowski (ed.) *International Encyclopedia of Ergonomics and Human Factors*, 2nd Edition, CRC Press (2006), 3084-3088.
17. Woolrych, A., Cockton, G. Why and when five test users aren’t enough. In *Proc. IHM-HCI 2001*, Cépadèus Éditions (2001), 105-108.