Developers' Needs and Severity Conceptions of Usability Problems

Ngoc-Huy Truong, Daniel Brand, Carine Ewert, Laura Wächter, Rul von Stülpnagel

Center for Cognitive Science, University of Freiburg

ngoc-huy.truong@cognition.uni-freiburg.de, daniel.brand@cognition.uni-freiburg.de, carine.ewert@mars.uni-freiburg.de, waechtel@tf.uni-freiburg.de, rul.von.stuelpnagel@cognition.uni-freiburg.de

Abstract

To improve products effectively, usability testing reports must be comprehensible for developers. However, it remains unclear which information is helpful and whether developers agree on the severity of usability issues with usability researchers. In this case study, a team of software developers rated three forms of usability feedback: (1) a low-detailed presentation of obtrusive usability problems, (2) a low-detailed list of aggregated problems, and (3) a high-detailed list of unique and aggregated usability problems. Furthermore, developers and usability researchers independently created criteria for low, medium, and high severe usability problems, based on their respective expertise. Our analyses indicate that (1) and (3) were both perceived as helpful. Moreover, agreement within each group and between both groups was high, indicating that developers and usability researchers can have a similar understanding regarding the severity of usability issues.

1 Introduction

Usability testing has been proven to be an effective means to reveal design limitations and usability issues of a product (Lewis, 2014). Testing results are particularly valuable for developers to decide which changes to make in future iterations. But instead of conducting usability studies themselves, developers often receive summarized reports from usability practitioners. This raises two questions: Firstly, as there are no defined standards for reporting usability issues during development (*formative reports*), it remains unclear which information is particularly helpful for developers (Dumas, Molich, & Jeffries, 2004; Theofanos & Quesenbery, 2005). Secondly, it is unknown to what extent developers agree with usability recommenda-

Veröffentlicht durch die Gesellschaft für Informatik e. V. 2018 in
R. Dachselt, G. Weber (Hrsg.):
Mensch und Computer 2018 – Tagungsband, 02.–05. September 2018, Dresden.
 https://doi.org/10.18420/muc2018-mci-0200

tions. If developers have different concepts about the severity of usability issues to practitioners, recommendations could be less valuable - or worse - ignored altogether (Ferre, Juristo, Windl, & Constantine, 2001).

1.1 Reporting Usability Issues during Development

What type of information should be reported under which circumstances in a formative usability report? In workshops with usability professionals, Theofanos and Quesenbery (2005) identified several aspects of usability reports. They found that reports varied strongly in format (presentation, tables, or full reports), elaborateness (varying from 5 to 55-pages), level of detail, and the way recommendations were presented. Reports also differed concerning included metrics (e.g. severity ratings of usability problem, task success or time on task). The structure of a report was highly dependent on the recipient's needs, knowledge, and role in the team. In other words, usability reports were and should be tailored to the audience. Molich and colleagues (Molich, Jeffries, & Dumas, 2007) proposed several guidelines to write useful recommendations: They should be precise and elaborate enough so that recipients can understand them without further background information, improve overall usability (and not only specific cases of the application), not cause new problems, and consider technical constraints. Hornbaek and Stage (2006) argued that bare lists of usability problems without further information were not helpful for developers because they could not be prioritized. Without prioritization, no useful recommendations can be provided. Finally, Norgaard and Hornbaek (2009) concluded that different types of formats, (e.g. problem lists or multimedia presentations) have different strengths and weaknesses. For example, usability lists provide succinct information about simple problems and can be supplemented with metrics but are rather short and imprecise. All in all, different authors proposed several formats, metrics and guidelines. It is still unclear under which circumstances what type of feedback is particularly helpful. In this paper, we thus aim to shed further light on the question:

RQ1: Which information in a formative usability feedback report meets developers' needs and is perceived as particularly helpful?

1.2 Severity Ratings

Severity of usability problems is an important metric to decide which product changes should be made in future iterations (Hornbaek & Stage, 2006). However, there is an ongoing controversy on how to define severity of usability problems and about the consistency of severity ratings between raters. Several severity scales have been proposed (Dumas & Redish, 1999; Hertzum, Molich, & Jacobsen, 2014; Rubin & Chisnell, 2008; Wilson & Coyne, 2001). All scales comprised of categories for low, medium or high levels of severity, although different authors emphasized different factors. Problems of low severity are often described as being aesthetic preferences, possible user suggestions, rare errors with no data loss, or problems which are easily fixed by the user or the system. Medium severity problems are characterized as irritating, potentially solvable but time-wasting, or by important features which do not work as intended. Finally, highly severe problems prevent users from completing a task, cause ex-

treme irritation, lead to data loss, or rendering the application unusable. Although this classification seems plausible, the categories remain vague. Indeed, Hertzum, Molich, and Jacobsen (2014) found that raters often disagree in severity ratings. They argued that depending on experience, raters are often not capable of foreseeing the causes and consequences of problems, hence leading them to under- or overestimate their severity.

Next to the difficulties of rating a problem's severity, raters also disagree whether an issue represents a usability problem at all. This so-called Evaluator Effect describes the notion that practitioners identify different sets of issues, although they analyze the same usability data (Jacobsen, Hertzum, & John, 1998). One reason for this phenomenon is that evaluators often lack specialized domain knowledge of the application. Partnering up with software developers could be a possible solution. Another reason could be that evaluators rate test sessions based on different evaluation goals. Thus, it is necessary to phrase evaluation goals precisely (Hertzum et al., 2014). Rating the severity and identifying usability problems are highly influenced by individual differences. Hence, reported usability problems can be unreliable, especially for the recipients. This leads to the question: If usability practitioners already disagree, how strong is the agreement between usability practitioners and developers? If for example agreement is low, usability reports could be perceived as unconvincing. Hoegh (2006) found that developers and practitioners can disagree in severity ratings. Practitioners rated most problems less severe than developers. Since in his study developers used a common severity scale, it remains unclear whether they would use different criteria to judge the severity of problems on their own. This research aims to fill this gap by addressing the following questions:

RQ2a: ***Which criteria do developers use to rate the severity of usability issues?***

RQ2b: ***Do usability researchers and developers agree in their severity ratings of usability issues based on their own criteria?***

2 Methods

2.1 Study Background: Usability Testing and Reports

Usability tests with eight users (seven males, aged 23-76 years, M = 48.33, SD = 15.07, all German) were conducted on a social networking site for neighborhoods. Several actions could be performed on this website, such as setting profile information or writing posts in online groups. Usability sessions were video-recorded. The Perceived Website Usability Scale (PWU, Moshagen, Musch, & Göritz, 2009) and the Visual Aesthetics of Websites Inventory (VisAWI, Moshagen & Thielsch, 2010) were filled out by the users to measure their opinions about website usability and aesthetic. After usability testing, three types of usability reports were created: 1) The first report (*low-detail-presentation* one week after usability testing) was a presentation with explanations, diagrams and interpretations of questionnaire results as well as a first overview of 20 obtrusive usability problems. These usability problems were presented with a short description without metrics for frequency of occurrences, frustration level, and

severity, in order to provide a first overview about the usability study's results. Recorded videos of the usability sessions were provided due to repeated requests by developers. 2) In the second report (*low-detail-list* provided three weeks after testing), 30 aggregated usability problems were described. A problem was included when an issue applied to at least two users and it was negative. This list included a title, a short description (max. two sentences) and the task where the problem occurred. The first and second author of this paper categorized these problems into usability error types, e.g. expectation or design error (see Table 3). No information about further metrics were given. 3) The final report (*high-detail-list* presented four weeks after testing) consisted of information about all unique usability problems, the corresponding usability category (e.g. navigational problems, missing error messages, or irreversible consequences), webpage and task where an issue occurred, severity ratings, and the frequency of occurrence aggregated as well as for each user individually.

2.2 Participants

We created a team of six usability researchers from the University of Freiburg with an academic background in psychology or cognitive science (two males, aged 24-30 years, M = 26.70, SD = 2.42) for the purpose of conducting and analyzing the usability tests presented in this paper. One researcher (female, 25 years old) provided advice based on her experience as a usability consultant. Four developers (all male, aged 22-42 years, M = 32.33, SD = 10.02) involved in backend- and frontend programming of the social networking site also participated. They had at least three years of professional programming experience (range = 3-12 years; M = 9.00, SD = 5.20). The developers worked and communicated with each other on a daily basis in a company not affiliated with the University of Freiburg. All participants were German.

2.3 Procedure

At first, the developers received the low-detail-presentation and recorded videos one week after usability testing. After three weeks, the developers and researchers had two tasks. 1) Each group (developers and researchers) should discuss and think of usability severity criteria (low, medium and high severity) based on their expertise in their respective fields. No examples for usability problems were provided at this point[1]. They were instructed to find general severity criteria impartial to observed emotional responses of the user and actual task completion rates. In addition, both groups were told that later on they would be presented with general usability problems, which only consisted of a title, a short description and the task where it occurred (so no information about task completion rates or emotionality would be given). Both sessions took two hours in the laboratory and were moderated by the first author. 2) After both groups were content with their severity criteria, the second task was to judge the severity of a number of aggregated usability problems based on their respective criteria. For this purpose, each developer and researcher received the randomized list of 30 usability problems (low-detail-aggregated-list, see Table 3 for examples) and were instructed to rate them individually. If unsure or if they thought that two or more categories could match a problem, they were instructed to

[1] Note that researchers had knowledge about potential usability problems due to conducting the usability tests.

choose the more severe category. Additional comments to each usability problem were allowed. Four weeks after usability testing, developers received the high-detail-list of usability problems. After receiving each of the three usability reports, the developers were asked to rate the reports' helpfulness, completeness and comprehensibility on 5-point Likert scales (1 = not at all to 5 = definitively, one item per construct). In addition, level of detail of usability reports should be rated on 5-point bipolar scales (from 1 = low level of detail to 5 = too many details). Questions were adapted from Gollwitzer and colleagues (Gollwitzer, Kranz, & Vogel, 2006). Furthermore, open-ended questions about which information developers needed were asked.

3 Results

3.1 RQ1: Which Information is Perceived as Helpful?

Table 1 shows descriptive data for the evaluation questions of each feedback format. Results are presented descriptively and are based on qualitative data due to the small sample size. As preliminary feedback, the presentation was perceived as rather helpful and appriopriate in its level of detail. Developers remarked that the clear, concise, and visual display of usability results helped them to get an overview about the methods and most important usability issues. The transparency of the procedure was emphazised. Although developers acknowledged that the analysis would take more than a week, they wished to have summarized information about users' emotions and thoughts, a ranking of issues, quantitative metrics (e.g. frequency or severity), elaborate descriptions of issues, as well as user comments and improvement suggestions at this early stage. The low-detail-list of usability issues was perceived as rather unhelpful, neither complete nor incomplete, but rather comprehensible and appriopriately elaborated. Developers wished to have detailed usability descriptions (e.g. which action triggered the problem), problem frequency, frustration level, and completion rates. Screenshots and video clips of issues were also requested. Lastly, the high-detail-list was rated as rather helpful, complete, comprehensible and detailed. Developers remarked that two list views, namely a detailed and an aggregated view, were positive. The order of columns were at times inconvenient, the explanations about metrics too short and the aggregated view lost information. Recommendations and prioritization of issues were also requested.

	Presentation (t1)			**Low-Detail-List (t2)**			**High-Detail-List (t3)**		
Evaluation	***M***	***SD***	**95% CI**	***M***	***SD***	**95% CI**	***M***	***SD***	**95% CI**
Helpful	4.00	.00	[4.00, 4.00]	2.25	.50	[1.76, 2.74]	3.50	.58	[2.93, 4.07]
Complete	3.33	.58	[2.67, 3.99]	3.25	.96	[2.31, 4.19]	3.75	.50	[3.26, 4.24]
Comprehensible	2.00	1.00	[.87, 3.13]	3.75	.50	[3.26, 4.24]	4.25	.96	[3.31, 5.19]
Level of detail*	2.67	.58	[2.01, 3.33]	3.00	1.20	[1.80, 4.20]	3.75	.50	[3.26, 4.24]

*Table 1: Evaluation means, standard deviations and 95% confidence intervals for presentation, low-detail-list, and high-detail-list feedback. *Level of detail was measured on a bipolar scale (1 = low detail to 5 = too many details)*

3.2 RQ2a: Which Criteria Do Developers Use to Rate the Severity of Usability Issues?

Table 2 shows an overview of severity criteria generated by developers and researchers. At first, both groups proved to be hesitant to come up with severity criteria without information about the user's frustration level and task completion rates. However, both groups were able to identify categories. Researchers came up with a higher number of criteria than developers. Comments by the developers and researchers suggested that developers tried to limit their criteria to a short number of the most important usability criteria, whereas researchers tried to find specific categories and error cases.

Severity	Developers	Researchers
High	Blocking of primary functions*	Abort
	High Frustration leads to abort	Strong negative feelings
	Unreadable text leads to abort	Missing feedback or error messages
		Irreversible consequences and data loss
		Critical consequences easily triggered
		No fault tolerance of system
		Loss of control of personal data
		Feeling lost in navigation
Medium	Blocking of secondary functions*	Massively exacerbated use and navigation
	High frustration does not lead to abort	Usage unnecessarily complicated
	Unreadable, important text	Function not found, as expected or comprehensible
	Frequent user wishes indicate other problem	Incomprehensible error messages
		Incomprehensible but relevant terms
		Strong irritation
		Confusing design and aesthetics
Low	Nice-to-have wishes	No (serious) consequences
	Low frustration	Quickly solvable or correctable problems
	Unreadable, unimportant text	Problems irrelevant for general system usage
	Design and aesthetic preferences	Nice-to-have wishes
		Design and aesthetic preferences
		Unfavorable terms and symbols
		Other web-standards preferred

*Table 2: Severity criteria created by developers and researchers. *Blocking is defined as when functions do not work or users do not know how to use them. Primary functions are essential to the website, secondary are not*

3.3 RQ2b: Severity Agreement Within and Between Developers and Researchers

Three intraclass correlation coefficients (ICC) were estimated to measure the inter-rater-reliability of severity ratings of the aggregated 30 usability problems (low-detail-list) between a) all developers b) all researchers, c) and all raters (developers and researchers) together. ICC

estimates were chosen based on mean-rating, two-way-mixed-effects models and consistency. Developers (n = 4) had an inter-rater-reliability of ICC = .89, 95% CI [.80, .94], whereas researchers (n = 6) had an estimate of .84, 95% CI [.72, .91], suggesting that both teams were consistent in their ratings. The inter-rater-reliability between all raters (n = 10) was good, ICC = .83, 95% CI [.72, .91], indicating that both groups may have similar severity concepts. Since intraclass correlations for each group were high, means over all 30 aggregated usability problems (low-detail-list) for each group were calculated and compared. Usability researchers (*M* = 1.98, *SD* = .14, 95, CI [1.86, 2.10]) did not differ from developers (*M* = 2.11, *SD* = .13, 95, CI [1.99, 2.23]), t(8) = -1.43, p > .05, r = .45, suggesting that there may be no difference between researchers and developers regarding their perception of usability severity.

Looking further into usability error types, a high agreement between conformity of expectation and irreversible or unclear consequence errors between both groups was found (for ICC values and confidence intervals see Table 3). A good agreement was reached for understanding, usage and unclear term errors. For imprecise system feedback and design problems we found an acceptable level of agreement.

Error type	**Example of usability description**	**Task**	**ICC [95% CI]**
Expectation	Expected web standards on web forms are not present, e.g. a * for mandatory fields	Please set your profile information	.96 [.83, .99]
Consequences	It is unclear on which parts of the website personal information will appear	Please set your profile information	.88 [.45, .99]
Understanding	It is unclear which posts appear on the newsfeed	Please post a posting	.79 [.23, 99]
Usage	It is unclear, who is allowed to join groups or how group members are moderated	Please join a group	.78 [.31, .98]
Term	The meaning and purpose of the term username is unclear	Please register on the website	.71 [.07, .97]
Feedback	There are no, imprecise or not seen error messages of the registration forms	Please register on the website	.68 [-1.31, 1.99]
Design	The color of the font is not easily readable	Please set your profile information	.62 [-.75, .99]

Table 3: Error types, examples of usability problem descriptions presented to participants, corresponding tasks and the intraclass correlations with 95% confidence intervals for developers and researchers together

4 Discussion

In this case study, we aimed to answer the following research questions. Firstly, which information in usability reports do developers find helpful? Secondly, which concepts of severity do developers have, and do they agree with usability researchers?

Regarding the first question, a quick presentation immediately after usability testing was perceived as helpful by the developers. Presentations are a quick and visual way to explain methods, results and interpretations of usability questionnaires about users' opinions as well as giv-

ing a first overview of obtrusive usability problems. This is in line with Norgaard and Hornbaek (2009), who found that multimedia presentations are a great means giving a convincing and clear overview about usability results. In contrast, an aggregated list of usability problems with rather limited information (title, short description and the task where the problem occurred) was perceived as less helpful because it is difficult to interpret short usability information without further context (Hornbaek & Stage, 2006). The detailed list of aggregated usability problems and every unique case was rated as moderately helpful. This is surprising because we believed that the more details were included in a report, the more helpful the report would be perceived. This contradicts results in the literature (Hornbaek & Stage, 2006; Theofanos & Quesenbery, 2005). We think that two reasons could explain this finding: 1) Due to time constraints and limited work resources, developers wished to have already analyzed and summarized usability recommendations. Since our detailed lists consisted of a lot of data, it could be difficult to understand which metric is important. 2) The second reason could be imprecise descriptions for the metrics. More detailed and transparent explanations should be provided. Comments of developers revealed that they also wished to receive more information about emotional responses, quotes of users and occurrences of experimenters' support during the testing sessions.

Regarding the second group of research questions, we found that the developers show very similar concepts of severity as compared to those introduced in the respective literature (Dumas & Redish, 1999; Hertzum et al., 2014; Rubin & Chisnell, 2008; Wilson & Coyne, 2001). Interestingly, they were able to reach a high level of agreement in severity ratings, although the usability problems lacked essential information about frustration level, task completion rates, and frequency of occurrences. This is insofar surprising since task abortion and user frustration were two of the main criteria the developers came up with. Furthermore, developers did not see the usability problems to be rated beforehand, so their criteria were just based on experience. In comparison to the researchers, their criteria were also not as elaborate and specific. We hypothesize that these results could be explained due to the developers' daily work process. Since team meetings and discussions about the main functions of the website were frequent, we assume that each developer adapted similar concepts about software problems. Furthermore, due to their experience in software development, the developers could have estimated the probability of whether a usability problem could potentially lead to task abortion or unusable primary functions heuristically. Some optional comments like "could potentially lead to abortion" supported this claim. Lastly, all developers remarked that they have seen some of the recorded videos of the usability sessions. It could be possible that viewing the videos could give the developers a feeling of certainty (Hertzum et al., 2014).

Finally, a high agreement between developers and researchers suggest that both groups can have similar severity concepts despite their different backgrounds and usability knowledge. The results contrast with the study of Hertzum and colleagues (Hertzum et al., 2014). We explain these results because researchers designed the usability sessions transparently and asked developers for their goals of the usability tests. Furthermore, both groups worked on this project for two years and had good domain specific knowledge about the social networking site. Looking further into error types, we found that on expectation and consequence errors both groups highly agreed on the severity. This is not surprising, since both error types are critical in a system and can often lead to abort or severe consequences (Wilson & Coyne,

2001). The agreement on understanding, usage, and term error types was good but not as good as expection or consequence errors. We believe that for these types of errors the room for interpretation is higher. Agreement on design errors was low. We assume that opinions on aesthetic aspects are inherently subjective. Finally, the low agreement on feedback errors is surprising. Comments of developers indicated that they were aware of missing error messages which were already set to be fixed in following iterations. Therefore, developers rated the severity low, whereas researchers assigned a high priority to these issues.

4.1 Future Directions and Conclusion

Due to the small sample size and only one developer team, the results must be considered preliminary. Future studies should compare larger samples with more distinct developer teams. Furthermore, it remains open whether different developers (e.g. backend vs. frontend developers, commercial vs. open source developers) have the same severity concepts. In addition, studies to optimize how to measure the developers' needs should be conducted. One goal could be to develop standardized questionnaires to identify severity concepts and to easily compare differences and commonalities between developers and researchers. Finally, further studies could investigate how to derive quick, comprehensible, and effective recommendations from usability testing data and how to integrate recommendations into the workflow of developers.

This study is a first attempt to analyze developers' needs for usability reports. We propose that in practice, usability practitioners should make each analyzing step and method transparent and easily understandable so that usability methods and reports become convincing for developers. We hope that this increases the agreement between developers and practitioners which in turn improves a product in a congruent and productive way.

Acknowledgements

We thank Marc Binz, Simon Jacobs, Kay Schmitteckert, and Simon Wiesmayr from IT Strategen (www.itstrategen.de). This study was supported by a grant of BMBF for the project SoNaTe.

References

Dumas, J. S., Molich, R., & Jeffries, R. (2004). Describing usability problems. *interactions*, *11*(4), 24–29. https://doi.org/10.1145/1005261.1005274

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

Ferre, X., Juristo, N., Windl, H., & Constantine, L. (2001). Usability basics for software developers. *IEEE Software*, *18*(1), 22–29. https://doi.org/10.1109/52.903160

Gollwitzer, M., Kranz, D., & Vogel, E. (2006). Die Validität studentischer Lehrveranstaltungsevaluationen und ihre Nützlichkeit für die Verbesserung der Hochschullehre: Neuere Befunde zu den Gütekriterien des "Trierer Inventars zur Lehrevaluation" (TRIL). In G. Krampen & H. Zayer (Eds.), *Didaktik und Evaluation in der Psychologie* (pp. 90–104). Göttingen: Hogrefe.

Hertzum, M., Molich, R., & Jacobsen, N. E. (2014). What you get is what you see: Revisiting the evaluator effect in usability tests. *Behaviour & Information Technology*, *33*(2), 144–162. https://doi.org/10.1080/0144929X.2013.783114

Hoegh, R. T. (2006). Usability problems: Do software developers already know? In T. Robertson (Ed.), *the 20th conference of the computer-human interaction special interest group (CHISIG) of Australia* (pp. 425–428). https://doi.org/10.1145/1228175.1228264

Hornbaek, K., & Stage, J. (2006). The interplay between usability evaluation and user interaction design. *International Journal of Human-Computer Interaction*, *21*(2), 117–123. https://doi.org/10.1207/s15327590ijhc2102_1

Jacobsen, N. E., Hertzum, M., & John, B. E. (1998). The evaluator effect in usability studies: Problem detection and severity judgments. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *42*(19), 1336–1340. https://doi.org/10.1177/154193129804201902

Lewis, J. R. (2014). Usability: Lessons learned … and yet to be learned. *International Journal of Human-Computer Interaction*, *30*(9), 663–684. https://doi.org/10.1080/10447318.2014.930311

Molich, R., Jeffries, R., & Dumas, J. S. (2007). Making usability recommendations useful and usable. *Journal of Usability Studies*, *2*(4), 162–179.

Moshagen, M., Musch, J., & Göritz, A. S. (2009). A blessing, not a curse: Experimental evidence for beneficial effects of visual aesthetics on performance. *Ergonomics*, *52*(10), 1311–1320. https://doi.org/10.1080/00140130903061717

Moshagen, M., & Thielsch, M. T. (2010). Facets of visual aesthetics. *International Journal of Human-Computer Studies*, *68*(10), 689–709. https://doi.org/10.1016/j.ijhcs.2010.05.006

Norgaard, M., & Hornbaek, K. (2009). Exploring the value of usability feedback formats. *International Journal of Human-Computer Interaction*, *25*(1), 49–74. https://doi.org/10.1080/10447310802546708

Rubin, J., & Chisnell, D. (2008). *Handbook of usability testing: How to plan, design, and conduct effective tests* (2nd ed.). Indianapolis, IN: Wiley Pub. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=233103

Theofanos, M., & Quesenbery, W. (2005). Towards the design of effective formative test reports. *Journal of Usability Studies*, *1*(1), 27–45.

Wilson, C., & Coyne, K. P. (2001). The whiteboard: Tracking usability issues: to bug or not to bug? *interactions*, *8*(3), 15–19. https://doi.org/10.1145/369825.369828