Comparing Usability Problem Identification and Description by Practitioners and Students

Miranda G. Capra
HumanCentric Technologies, Inc.
Cary, NC

Many studies of usability studies count the number of usability problems identified to measure the effectiveness of an evaluation. However, communicating problems is also important to evaluation effectiveness, because a problem found but poorly explained may not be fixed. This study compared lists of usability problems from 21 practitioners and 23 students watching a pre-recorded usability session. Lists were evaluated for the number of problems reported, and for following six guidelines for describing usability problems: be clear and avoid jargon, describe problem severity, provide backing data, describe problem causes, describe user actions, and provide a solution. There was no difference in the number of problems reported by students and practitioners, but there was a difference in their ratings for following several of the guidelines. Using both measures provides a more complete assessment of usability reports.

INTRODUCTION

Problem lists are an important part of the formative usability evaluation report, identifying usability problems present in an interface that designers should fix in the next design iteration. Several studies have compared variability in problem lists created by evaluators conducting usability testing (e.g. Dumas, Molich, & Jeffries, 2004; Molich, Ede, Kaasgaard, & Karyukin, 2004; Molich et al., 1999; Rourke, 2003). These studies have focused on problem identification, counting usability problems found and measuring thoroughness and/or reliability of the problem lists to judge the quality of the evaluations. However, having good problem descriptions is also important. Poor communication of usability problems can reduce acceptance of a usability report and decrease the number of problems the recipients choose to fix (Dumas, Molich, & Jeffries, 2004; Jeffries 1994).

The author developed a set of ten guidelines for describing usability problems in a previous series of studies (Capra, 2005; Capra & Smith-Jackson, 2006; Capra, 2006). The goal of the current study was to assess the usefulness of these guidelines in rating problem lists. The approach was to collect problem lists from both usability practitioners and students to see if degree of following the guidelines was a distinguishing factor between these two groups. The hypothesis was that practitioners write better reports than students and so should identify more problems and receive higher scores for following the guidelines.

METHOD

Participants

The usability problem lists evaluated in this study were written by 44 participants, 21 *practitioner evaluators* and 23 *student evaluators*.

Practitioner evaluators were recruited from several usability mailing lists, and had five years of usability experience or had conducted 10 usability evaluations. Practitioners had 2-20 years of experience (*M*=11.0, *SD*=6.0, Median=10) and had conducted 10-500 usability evaluations (*M*=85.0, SD=117.2, Median=40).

Student evaluators had completed a graduate-level course in usability engineering at either Virginia Tech or Georgia Tech. Students had 0-5 years of experience (*M*=1.8, *SD*=1.5, Median=1.5) and had conducted 0-10 usability evaluations (*M*=5.1, *SD*=2.0, Median=5).

Materials

Evaluators received a CD with a pre-recorded usability session and a report template. The pre-recorded usability session consisted of four participants using the Internet Movie Database (imdb.com) and performing this task: “name all of the movies that both Owen Wilson and Luke Wilson (the actor from Old School) have appeared in together.” The session recording was chosen because it had been reviewed before and was known to contain numerous usability problems, despite its short length (~11 minutes). It was created for a different study (Long et al., 2005) using Morae to capture participants’ screens and comments. The pre-recorded usability session was used to ensure that differences in the reports were due to the individual evaluators, and not task, participant selection, or facilitator. This technique has been used successfully in other studies (Jacobsen, Hertzum & John, 1998; Lesaigle & Biers, 2000; Long, Styles, Andre, & Malcolm, 2005; Skov & Stage, 2005; Vermeeren, van Kesteren & Bekker, 2003).

The report template was based on the template used in the fourth Comparative Usability Study (CUE-4; Molich, 2004). It contained a pre-written introduction, definitions of the problem severity codes used, and space to write problem descriptions, shown in Figure 1. The comment categories were positive finding (PF), minor problem (MP), serious problem (SP), critical problem (CP), good idea (GI) or bug (B). The template was shorter than the one used in CUE-4. The evaluators in CUE-4 designed their own studies and recruited their own users. In contrast, evaluators in the current study were provided with a session recording, and so sections about testing methods and participants were omitted or pre-filled.

> Please copy the following template and use it for each of your comments, filling in the areas highlighted in yellow.
>
> **Comment category [PF/MP/SP/CP/GI/B]: ________**
> **Comment:**
> Provide a complete description of the comment, using as much detail as you would typically include in your own descriptions. If you put images in your own reports you may include them with this description.

Figure 1. Usability Problem Template

A website was used to collect the usability reports and present a post-task questionnaire to the evaluators. The questionnaire collected demographic information, such as years of experience and number of evaluations performed. The questionnaire also asked the evaluators for their opinions about the guidelines, rating each guideline for how relevant, helpful, required and difficult it is to follow.

Procedure

Evaluators received the CD in the mail. At a time of their choosing, the evaluators then did the following:

- Watched the usability session movie and wrote comments in a report.
- Visited a website and uploaded the report.
- Completed the post-task questionnaire.

Evaluators were told that writing the report would take about two hours, but were allowed to spend as much or as little time as they wished.

Counting Usability Problem Identified

Evaluators submitted a list of comments in their usability reports. However, comments and problem descriptions were not equivalent because one comment sometimes described multiple problems, and one problem was sometimes described in multiple comments. In order to count usability problems identified by each evaluator, a master problem list (MPL) was needed, or a list of all the problems present in the usability session. The MPL was created by the author and four independent judges reviewing reports from 58 evaluators. While five evaluators would be a small number of evaluators to create a comprehensive MPL, merging problems in usability reports is a different task from identifying problems during a usability session. Judges become much more familiar with the system being evaluated, since they read reports from many different evaluators and spent more time analyzing the system than the evaluators did (20-40 hours for the three judges that created the final MPL, as opposed to 0.5-6.5 hours for the evaluators). In contrast to the five judges used in the current study, CUE-4 used two judges to merge their problem lists (Molich & Dumas, in press).

To begin the process, the author and one judge independently created a list of problems by reviewing 14 reports from pilot participants and participants in the Long et al. (2005) study; the author then merged these two lists. Three additional judges then independently reviewed the 44 reports from the current study, making a list of matches (instances where an evaluator mentioned a problem in the MPL) and problems to add to or remove from the MPL. The judges did not know which reports were from practitioners vs. students. The three judges then met and reconciled their lists; the author was part of these discussions but did not have a vote in the final decision. The final MPL included 41 usability problems.

Evaluating Problem Identification

Two measures of problem identification were used. *Thoroughness* is a measure of how many problems each evaluator finds, or the percent of the total problems as calculated using Equation 1.

$$\begin{array}{c}\textit{Thoroughness}\\ \text{of problem}\\ \text{identification}\end{array} = \frac{\text{\# found by this evaluator}}{\text{\# in master problem list}} \qquad (1)$$

Reliability is the degree to which evaluators tend to find the same problems. A common measure of reliability in problem identification is what Hertzum and Jacobsen (2003) call *any-two agreement*, which is the percent overlap in problem sets from a pair of evaluators, averaged across all pairs of evaluators. This was calculated using Equation 2. High reliability is not generally associated with either good or bad evaluations, but reporting reliability gives a broader picture of the problem sets collected.

$$\begin{array}{c}\textit{Reliability}\\ \text{of problem}\\ \text{identification}\end{array} = \frac{\displaystyle\sum_{i=2}^{n}\sum_{j=1}^{i-1}\frac{|P_i \cap P_j|}{|P_i \cup P_j|}}{n(n-1)/2} \quad \begin{array}{c}\text{where } n \text{ is the number of}\\ \text{evaluators and}\\ P_i \text{ is the set of problems}\\ \text{found by evaluator } i\end{array} \qquad (2)$$

Evaluating Problem Descriptions

The three judges that created the final MPL reviewed each usability report and rated the degree to which the problem descriptions adhered to the following six guidelines:

1. **Describe a solution to the problem**, providing alternatives and tradeoffs. Be specific enough to be helpful without dictating a solution, guessing, or jumping to conclusions. Supplement with pictures, screen capture, usability design principles and/or previous research. (*Solutions*)
2. **Be clear and precise while avoiding wordiness and jargon.** Define terms that you use. Be concrete, not vague. Be practical, not theoretical. Use descriptions that non-HCI will appreciate. Avoid so much detail that no one will want to read the description. (*Clarity/Jargon*)
3. **Describe the cause of the problem**, including context such as the interaction architecture and the user's task.

Describe the main usability issue involved in the problem. Avoid guessing about the problem cause or user's thoughts. (*Problem Cause*)

4. **Support your findings with data** such as: how many users experienced the problem and how often; task attempts, time and success/failure; critical incident descriptions; and other objective data, both quantitative and qualitative. Provide traceability of the problem to observed data. (*Backing Data*)
5. **Describe the impact and severity of the problem**, including business effects (support costs, time loss, etc.), impact on the user's task and importance of the task. Describe how often the problem will occur, and system components that are affected or involved. (*Impact/Severity*)
6. **Describe observed user actions**, including specific examples from the study, such as the user's navigation flow through the system, user's subjective reactions, screen shots and task success/failure. Mention whether the problem was user-reported or experimenter observed. (*User Actions*)

These guidelines are part of a set of ten guidelines developed in a series of previous studies (Capra, 2006; Capra & Smith-Jackson, 2006). They represent the five guidelines rated most important by 74 usability practitioners, plus *Describe a Solution*, about which there was mixed opinion as to whether it is a good or bad idea for a usability report. For each guideline, the judges were given the text of the guideline and a statement of the format "According to this guideline, this report is **clear and precise**." The judges rated the report using a six-point Likert-type scale (strongly disagree to strongly agree). The scale points were assigned values of –2.5 to 2.5 and averaged across the three judges. Across all ratings, the Pearson correlations between pairs of judges were .39, .52 and .46.

RESULTS

The 44 evaluators submitted 409 usability problem descriptions and 91 positive findings. Practitioners and students were compared initially on 12 dependent measures: comments written (total, severe, critical, minor, positive, good idea, bug, other), number of tables/images, total words, words per comment, and hours spent on the evaluation. Using a single multivariate analysis of variance (MANOVA) there was not a significant difference between students and practitioners, $F(11, 32) = 1.03$, $p = .44$.

An analysis of variance (ANOVA) comparing practitioners and students found no significant difference in thoroughness across all problems (minor and severe), $F(1, 42) = 0.31$, $p = .58$ (Table 1). Reliability (any-two agreement) was .37 for both practitioners and students (Table 2). A formal ANOVA was not conducted because reliability is measured on pairs of evaluators, rather than individual evaluators. Identification of severe problems and thoroughness/reliability for severe problems is too complex an issue to address in this paper, but differences between the two groups were small (Capra, 2006).

Table 1. Problem Identification Thoroughness

	Thoroughness	
	M	*SD*
Practitioners (n=21)	.22	.08
Students (n=23)	.22	.10

Table 2. Problem Identification Reliability

	Reliability	
	M	*SD*
Practitioners (n=21*20)	.37	.12
Students (n=23*22)	.37	.13

Differences in ratings for practitioners and students were tested using a 2x6x3 mixed-factor ANOVA, with evaluator group (practitioner and student) as a between-subject factor, and guideline and judge as within-subject factors (differences among judges is reported in Capra, 2006). Using an alpha level of .01, there was a main effect due to evaluator group, $F(1, 42) = 7.27$, $p = .01$. There was also a significant interaction between evaluator group and guideline, $F(2, 210) = 3.13$, $p = .01$. Post-hoc tests indicate significant differences for *Impact/Severity*, *Solutions*, and *Backing Data*, but not for *User Actions*, *Problem Cause*, and *Clarity/Jargon*, as marked in Figure 2.

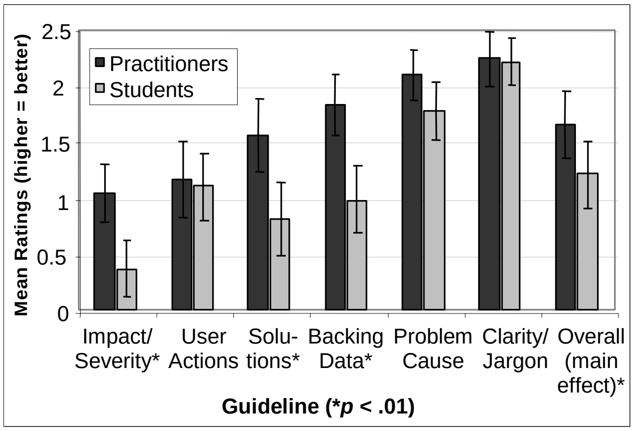

Figure 2. Mean ratings by guideline, evaluator group

In this study, evaluator group (practitioner, student) was used as an indicator of a good evaluation, with the assumption that practitioners write better reports than students. Many other possible indicators of a good evaluation were measured: experience, thoroughness, validity of the problem set, and hours spent on the evaluations. Correlations were computed between these measures and scores for the six guidelines; a conservative alpha level of .01 was used to protect against false positives due to the number of correlations computed (72). None of the correlations were significant; ratings for the guidelines differ between practitioners and students but guideline ratings were not related to any other indicators of a good evaluation.

It is possible that individual evaluators have differing opinions about the importance of each of the guidelines, and that some of the guidelines are more difficult to follow; both of these factors could affect the way evaluators write their reports. At the end of the study, each evaluator was asked to rate each of the guidelines for the following adjectives: helpful, relevant, required and difficult. Did evaluators' opinions about the guidelines affect their behavior, i.e. the guideline scores? The guideline scores were correlated with evaluators' ratings of the guidelines for each of the four adjectives. Only the correlations with *Provide a Solution* were significant ($|r|$ = .51-.67, p = .00-.02); this was the only guideline for which evaluator opinion was correlated with evaluator behavior.

DISCUSSION

The expected outcome of this study was that practitioners would be better at describing usability problems, receiving higher scores for the six guidelines, and would also be better at identifying usability problems, having higher thoroughness than the students. There was a difference in overall ratings across all six guidelines, with practitioners receiving better scores than the students. Practitioners also received higher scores for three of the individual guidelines: *Impact/Severity*, *Solutions*, and *Backing Data*. There was, however, no difference between the two groups in terms of thoroughness, and no correlation between thoroughness and the guideline ratings. Thoroughness of both groups was rather low, although within the range of previous studies of usability testing (e.g. .22-.52 in Hertzum & Jacobsen, 2003).

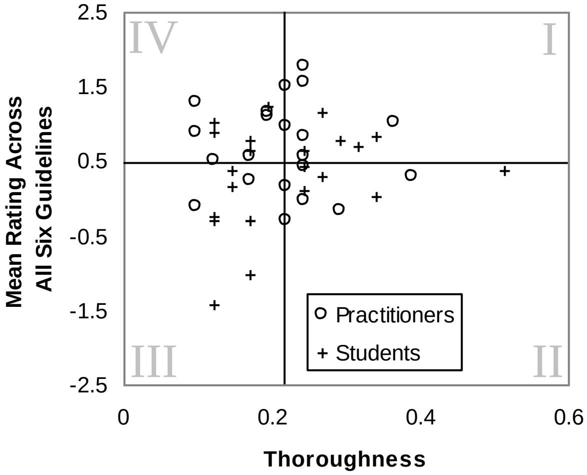

Figure 3. Comparison of thoroughness and mean guideline ratings

Figure 3 shows a scatter plot comparing thoroughness in finding severe problems and mean rating across five of the guidelines (excluding *Solutions*). The quadrant lines represent median values. As an example of the lack of correlation between the two measures, the student at the right edge of the graph received the highest score for thoroughness but only average scores for following the guidelines. Similarly, the practitioner at the top edge of the graph had the highest scores for following the guidelines but only average thoroughness. The best reports are likely those found in quadrant I, reports that received high marks for both finding severe problems and describing the problems. The weakest reports are likely those in quadrant III, reports that received low marks for both thoroughness and following the guidelines.

Consider the two example problem descriptions in Figure 4. Both describe the same problem, that the user expected the search box to support complex searches, such as logical (and, or) operations. The first example is from a practitioner with high marks for *Backing Data* and in quadrant I. This practitioner describes users' actions, how many users experienced the problem, and discusses user expectations based on other common search tools. The second example is from a student with low marks for *Backing Data* and in quadrant III. This student states that the search is missing functionality, but provides no additional information.

Example 1: Practitioner with high score for *Backing Data*
overall: 1.7; thoroughness: .24; quadrant: I

Two of the users expected to be able to filter the search by entering more than one item (the names of both actors) into the search box. One might argue that this requirement – multiple intersecting set search – is provided for elsewhere, to which the unbiased observer might reply:

- Users have come to expect just this behavior in all search boxes; many will try here and be disappointed.
- The user unfamiliar with the "joint venture" [search box] hunted for something like this, and saw no clue or direction to it, only stumbling on it by accident. Nobody ever misses the search box.
- Google serves one page to solve this problem. The site served, on the average, 5. I suppose that the marketing department could say that these users were looking at ads along the way, but I doubt it sorely. All that was accomplished was that the users – and the site's servers – got overheated.

Example 2: Student with low score for *Backing Data*
overall: -1.3 ; thoroughness: .12; quadrant: III

Search tool does not allow joint searches with logic operators. Perform Joint searches in the search tool by using the plus sign.

Figure 4. Examples of problem descriptions with high and low scores for *Backing Data*

When counting thoroughness, both of these problem descriptions would be given equal weight. However, the first description will be more useful to the product team and more likely to be fixed because it provides additional information and context. The second description may be misunderstood or dismissed because it provides little explanation or detail. The lack of association between the report ratings for following the guidelines and thoroughness measures suggests that performance in finding problems is not related to performance in describing problems. The two activities may be influenced by different factors and rely on different skills. Measuring both problem identification (i.e. thoroughness) and description (guideline ratings) gives a more complete picture of the effectiveness of a usability evaluation.

Why did both groups in this study receive low ratings for *Impact/Severity*, when severity ratings are generally considered essential companions to problem descriptions? The cause is likely to be the report template used in the study. The template included a specific field for marking the severity of the problems. The evaluators may have left this information out of their description text, knowing that there was a severity code for every problem. It is interesting that practitioners received higher scores than students for *Impact/Severity*. Perhaps the practitioners understood that communicating severity requires explanation and description, and not just a single severity code.

The studies used to develop the guidelines found that opinion about the *Solutions* guideline was very mixed (Capra & Smith-Jackson, 2006). Some practitioners feel strongly that solutions should be included together with descriptions, particularly consultants who want to provide clients with both problems and suggestions for how to fix them. Some practitioners feel equally strongly that solutions should not be included with problem descriptions, particularly in-house practitioners who want to include the entire product team in designing solutions to usability problems.

Providing a solution was, however, the one guideline whose report ratings were correlated with evaluators' opinions about the guideline. Evaluators who felt the guideline was more relevant, required and helpful were more likely to include solutions in their reports, and evaluators who felt the guideline was more difficult to follow were less likely to include solutions in their reports. In contrast, there was no relationship between opinion about the guideline and rating for following the guideline for any of the other six guidelines used in this study.

The guidelines explored in this study can be used in future studies comparing the effectiveness of usability evaluations. Measuring both problem identification and description gives a more complete picture of the evaluation than either measure alone. The guidelines may be used in training usability evaluators, either to grade reports or to select examples of good and bad problem descriptions. Usability practitioners could use the guidelines as a helpful checklist when writing usability reports, or to evaluate past reports to ensure that they are writing effective problem descriptions in their usability reports.

ACKNOWLEDGEMENTS

Portions of this research were conducted while the author was supported by the Alexander E. Walter Fellowship from the Grado Department of Industrial and Systems Engineering at Virginia Tech. Tonya Smith-Jackson and Rex Hartson assisted with the study design. Rob Capra, Laurian Hobby, Jon Howarth and Pardha Pyla assisted with data coding. Terence Andre provided the usability session recorings and reports from a previous study. Joe Dumas and Rolf Molich provided materials and suggestions from the CUE studies.

REFERENCES

Capra (2005). Factor Analysis of Card Sort Data: An Alternative to Hierarchical Cluster Analysis. In *Proceedings of the Human Factors and Ergonomics Society 49th Annual Meeting* (pp. 691-696). Santa Monica, CA: HFES.

Capra (2006). *Usability Problem Description and the Evaluator Effect in Usability Testing.* Unpublished dissertation, Virginia Tech: Blacksburg. Available via http://scholar.lib.vt.edu/theses/available/etd-03222006-201913/.

Capra & Smith-Jackson (2006). Developing Guideines for Describing Usability Problems. Technical Report #ACE/HCI-2005-002. Blacksburg: Virginia Tech, Grado Department of Industrial and Systems Engineering. Available via http://ace.ise.vt.edu/publications_technicalreports.htm

Dumas, J. S., Molich, R., & Jeffries, R. (2004). Describing usability problems: are we sending the right message? *interactions, 11*(4), 24-29.

Hertzum, M. & Jacobsen, N. E. (2003). The evaluator effect: a chilling fact about usability evaluation methods. *International Journal of Human-Computer Interaction, 15*(1), 183-204.

Jacobsen, N. E., Hertzum, M., & John, B. E. (1998). The Evaluator Effect in Usability Tests. In C.-M. Karat & A. Lund & J. Coutaz & J. Karat (Eds.), *Proceedings of the Conference on Human Factors in Computing Systems (CHI 98)* (pp. 255-256). New York: ACM.

Long, K., Styles, L., Andre, T., & Malcolm, W. (2005). Usefulness of Nonverbal Cues from Participants in Usability Testing Sessions. In G. Salvendy (Ed.) *Proceedings of the 11th International Conference on Human-Computer Interaction (HCII 2005)*. St. Louis, MO: Mira Digital.

Molich, R. (2004). *Comparative Usability Evaluation - CUE*. DialogDesign: Stenløse, Denmark. Retrieved August 8, 2004 from http://www.dialogdesign.dk/cue.html.

Molich, R. & Dumas, J.S. (in press). Comparative Usability Evaluation 4 (CUE-4). *Behaviour & Information Technology*.

Molich, R., Ede, M. R., Kaasgaard, K., & Karyukin, B. (2004). Comparative Usability Evaluation. *Behaviour & Information Technology, 23*(1), 65-74.

Molich, R., Thomsen, A. D., Karyukina, B., Schmidt, L., Ede, M., van Oel, W., et al. (1999). Comparative evaluation of usability tests. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI '98) [extended abstracts]*. New York: ACM.

Rourke, C. (2003). *CUE-4: Lessons in Best Practice for Usability Testing and Expert Evaluation*. User Vision. Retrieved December 14, 2004 from http://www.uservision.co.uk/usability_articles/usability_CUE.asp.

Skov, M. B., & Stage, J. (2005). Supporting problem identification in usability evaluations. In *Proceedings of the 19th conference of the computer-human interaction special interest group (CHISIG) of Australia on Computer-human interaction: citizens online: considerations for today and the future (OZCHI 2005)*. Narrabundah, Australia: Computer-Human Interaction Special Interest Group (CHISIG) of Australia.

Vermeeren, A. P. O. S., van Kesteren, I. E. H., & Bekker, M. M. (2003). Managing the Evaluator Effect in User Testing. In M. Rauterberg (Ed.) *Proceedings of the Human-Computer Interaction - INTERACT'03* (pp. 647-654): IOS.