Usability Problem Description and the Evaluator Effect in Usability Testing

Miranda G. Capra

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Industrial and Systems Engineering

Dr. Tonya L. Smith-Jackson, Chair
Dr. John K. Burton
Dr. H. Rex Hartson
Dr. Brian M. Kleiner
Dr. Maury A. Nussbaum

March 13, 2006
Blacksburg, Virginia

Keywords: usability testing, formative usability evaluation methods, problem descriptions, evaluator effect, thoroughness, reliability

Usability Problem Description and the Evaluator Effect in Usability Testing

Miranda G. Capra

ABSTRACT

Previous usability evaluation method (UEM) comparison studies have noted an *evaluator effect* on problem detection in heuristic evaluation, with evaluators differing in problems found and problem severity judgments. There have been few studies of the evaluator effect in usability testing (UT), task-based testing with end-users. UEM comparison studies focus on counting usability problems detected, but we also need to assess the content of usability problem descriptions (UPDs) to more fully measure evaluation effectiveness. The goals of this research were to develop UPD guidelines, explore the evaluator effect in UT, and evaluate the usefulness of the guidelines for grading UPD content.

Ten guidelines for writing UPDs were developed by consulting usability practitioners through two questionnaires and a card sort. These guidelines are (briefly): be clear and avoid jargon, describe problem severity, provide backing data, describe problem causes, describe user actions, provide a solution, consider politics and diplomacy, be professional and scientific, describe your methodology, and help the reader sympathize with the user. A fourth study compared usability reports collected from 44 evaluators, both practitioners and graduate students, watching the same 10-minute UT session recording. Three judges measured problem detection for each evaluator and graded the reports for following 6 of the UPD guidelines.

There was support for existence of an evaluator effect, even when watching pre-recorded sessions, with low to moderate individual thoroughness of problem detection across all/severe problems (22%/34%), reliability of problem detection (37%/50%) and reliability of severity judgments (57% for severe ratings). Practitioners received higher grades averaged across the 6 guidelines than students did, suggesting that the guidelines may be useful for grading reports. The grades for the guidelines were not correlated with thoroughness, suggesting that the guideline grades complement measures of problem detection.

A simulation of evaluators working in groups found a 34% increase in severe problems found by adding a second evaluator. The simulation also found that thoroughness of individual evaluators would have been overestimated if the study had included a small number of evaluators. The final recommendations are to use multiple evaluators in UT, and to assess both problem detection and description when measuring evaluation effectiveness.

ACKNOWLEDGEMENTS

I would like to thank my advisory committee, Tonya Smith-Jackson, John Burton, Rex Hartson, Brian Kleiner, and Maury Nussbaum. In particular, Tonya Smith-Jackson provided suggestions for experimental design and statistical analysis, and Rex Hartson and Tonya Smith-Jackson provided suggestions for the judging procedure used to analyze the usability problems collected from evaluators. I also owe a special thanks to Laurian Hobby, John Howarth, and Pardha Pyla, each of whom generously donated over 50 hours of their time to read and evaluate hundreds of usability problems. The final study would not have been possible without them. Thanks also to my thesis advisor, Bob Williges, who guided me when I began studying usability methods.

This dissertation would not have been possible without the support of my husband, Rob Capra, who shared the journey with me as we both completed our dissertations within a month of each other. I might not have discovered Human Factors or pursued a PhD without him. He is my partner in research, and many important ideas grew out of discussions of our research at school, at home, and over meals. He is my partner in life, and I love him with all my heart.

Many other people contributed to this research. Terence Andre shared the usability movies and reports from studies run by him and his students. Rob Capra assisted with data coding in the first and fourth studies. Suzanne Aref suggested using factor analysis to cluster the items in the card sort. Joe Dumas shared his experiences analyzing the CUE-4 reports, provided criteria for identifying descriptions that discuss the same usability problem, and helped recruit practitioners. Rolf Molich gave permission to use the report template and severity rating scales from the CUE studies. The comparison of usability diagnosis to medical diagnosis was refined through discussions with Rex Hartson and Steve Belz.

Thanks to the dozens of usability practitioners and graduate students who volunteered to participate in my studies or provided feedback as pilot participants, especially the practitioners that took extra time above and beyond the study requirements to discuss their usability practices and reporting habits. Thanks to my family and friends for years of encouragement and patience, and fellow students for support and commiseration along the way. Thanks to my former colleagues at what is now AT&T Labs for introducing me to Human Factors and sparking my interest in the field.

Thanks to IMDb, Inc. for permission to include screen shots of their website in this document. Information courtesy of:

The Internet Movie Database
(http://www.imdb.com/).
Used with permission.

Portions of this research were conducted while I was supported by the Alexander E. Walter Fellowship from the Grado Department of Industrial and Systems Engineering (2001-2004).

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATIONS

CHAPTER 1. Introduction

Formative usability evaluations are an important part of the usability life cycle, identifying usability problems present in an interface that designers should fix in the next design iteration. Figure 1.1 provides an overview of a formative usability evaluation and the usability life cycle. The output of a formative usability evaluation is a set of usability problems (Hartson, Andre, & Williges, 2003), but there are many different usability evaluation methods (UEMs). *Empirical* evaluations involve end-users. The most common empirical method is usability testing or think aloud testing, which is generally a task-based session in a usability laboratory. *Analytical* evaluations involve expert review of an interface, such as Heuristic Evaluation (Nielsen, 1994b; Nielsen & Molich, 1990).

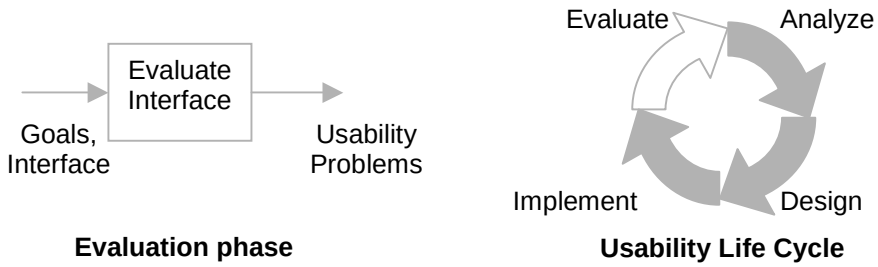


Figure 1.1 Formative Usability Evaluation and the Usability Life Cycle

Formative usability evaluation is not a reliable process. Evaluators discover different sets of usability problems depending on the usability evaluation method (UEM) used or the individual evaluator that performs an analytical evaluation (Hertzum & Jacobsen, 2003). Other factors that can affect problem detection are the number and type of users involved in usability testing (Law & Hvannberg, 2004a; Nielsen, 1994a; Spool & Schroeder, 2001; Virzi, 1990, 1992) and the number of evaluators involved in an expert review (Dumas & Sorce, 1995; Dutt, Johnson, & Johnson, 1994). Evaluators also differ in their judgment of the severity of usability problems (Hassenzahl, 2000; Hertzum & Jacobsen, 2003).

Hertzum and Jacobsen coined the term *evaluator effect* to refer to differences in problem detection and severity judgments by evaluators using the same UEM (Hertzum

& Jacobsen, 2003; Jacobsen, Hertzum, & John, 1998). Several previous studies have suggested the presence of an evaluator effect in usability testing. The Comparative Usability Evaluation (CUE) studies run by Rolf Molich have found low overlap in sets of problems discovered by professional teams evaluating the same interfaces with usability testing (Dumas, Molich, & Jeffries, 2004; Molich et al., 1998; Molich, Ede, Kaasgaard, & Karyukin, 2004; Molich et al., 1999; Rourke, 2003). Studies combining new data with data from CUE-2 (Molich et al., 2004) have found similar effects with professional teams (Kessner, Wood, Dillon, & West, 2001) and students (Skov & Stage, 2003, 2004).

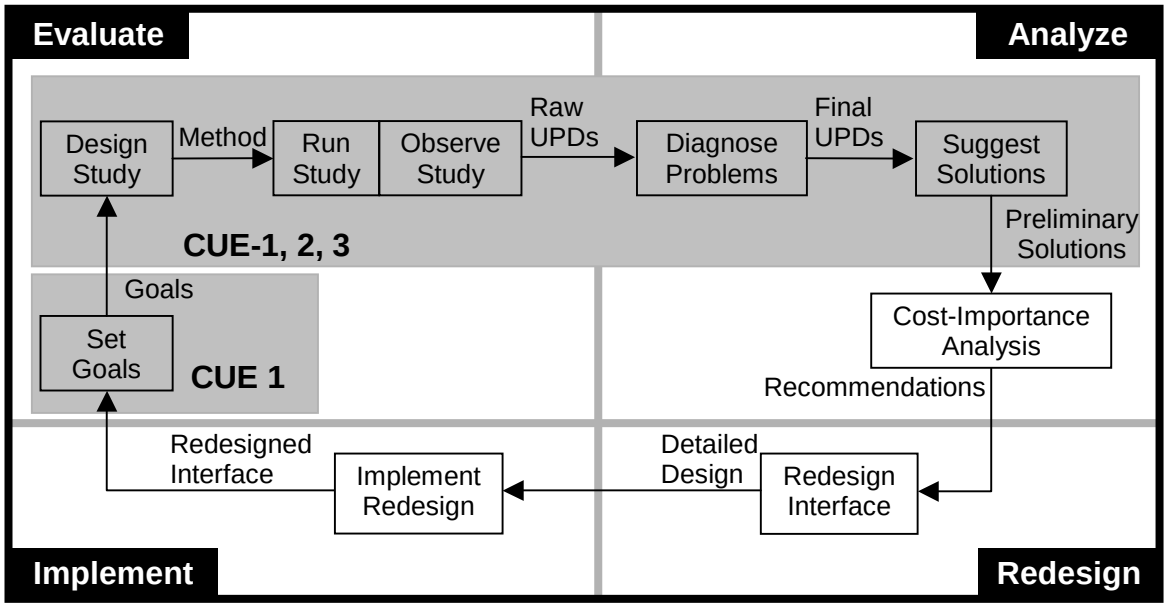


Figure 1.2 Portions of the Usability Life Cycle Studied in the CUE studies

These studies incorporated several activities relating to formative usability evaluation, as shown in Figure 1.2. Thus, the low levels of reliability could be due to differences in study design, task selection, participant selection, or observation and analysis. Several studies have examined reliability among evaluators used pre-recorded usability sessions (Jacobsen et al., 1998; Vermeeren, van Kesteren, & Bekker, 2003) or simultaneous viewing of live sessions (Palacio, Bloch, & Righi, 1993) to control for effects due to task and participant selection, but these studies used four or fewer evaluators, resulting in low statistical power. In contrast, several studies of the evaluator effect in Heuristic Evaluation have used between 30 and 77 evaluators (Hornbæk & Frøkjær, 2004a; Molich & Nielsen, 1990; Nielsen, 1992; Nielsen & Molich, 1990), with

any-two agreements ranging from 7-45%. See Appendix A for list of 57 UEM comparison studies.

Most studies comparing UEM effectiveness or measuring the evaluator effect rely on measures that involve counting usability problems identified by each evaluator, such as the any-two agreement measure of reliability (Hertzum & Jacobsen, 2003) or the thoroughness and validity of problem sets (Hartson et al., 2003). These measures are useful, but they do not give a complete measure of the effectiveness of an evaluation. Counting usability problems detected is part of the equation, but a better question to ask is how can an evaluation help to efficiently improve a product (Wixon, 2003). Communicating the results of the test (either through a written report or verbally) is an essential part of the usability testing process (Rubin, 1994). Andre, Hartson, Belz, and McCreary (2001) express the opinion that poor documentation and communication of usability problems identified diminish the effectiveness of a usability evaluation and can reduce the return on the effort invested in conducting the evaluation. Dumas, Molich and Jeffries (2004) suggest that communication style and attitude in report writing can affect recipients' acceptance of suggestions and the number of the problems recipients choose to fix. Jeffries (1994) suggests that developers may interpret poorly described problems as false alarms, causing the developers to ignore the poorly described problems and increasing the likelihood that developers will treat future problems as opinion or false alarms. A more complete measure of UEM output would assess not only the quantity of problem descriptions generated but also the quality.

There has been little formal research into usability problem description. Two articles by members of the usability community (and sharing an author) have commented on UPDs collected from UEM comparison studies (Dumas et al., 2004; Jeffries, 1994). However, the authors appear to base the articles on their expertise and personal review, rather than formal analysis. Many authors have developed structured problem reporting forms and problem classification schemes (Andre et al., 2001; Cockton, Woolrych, & Hindmarch, 2004; Hvannberg & Law, 2003; Lavery, Cockton, & Atkinson, 1997; Sutcliffe, 2000) to increase the utility of problem descriptions and thoroughness of problem diagnosis. However, these studies have not provided formal documentation of

poor problem descriptions, nor have they provided measures of problem description quality to demonstrate the effectiveness of these tools. We cannot begin to measure evaluation effectiveness in terms of description quality until we have established measures of UPD content.

1.1 Problem Statement

Formative usability evaluation is not a reliable process. There is evidence of an evaluator effect on problem detection in analytical methods and a user effect in usability testing. Usability testing had been considered the gold standard of usability evaluation, but there have been few studies of whether it is also subject to an evaluator effect. Previous studies of the evaluator effect on problem detection in usability testing have allowed different tasks and users, used a small sample size, or used student teams; we need larger studies of the evaluator effect that focus specifically on study analysis, rather than design and execution. In addition, previous studies of usability testing have focused primarily on comparing the number of problems identified by each evaluation. Counting usability problems identified by an evaluation is a necessary but not sufficient measure of evaluation effectiveness. We need measures of usability problem description (UPD) content to be able to document shortcomings in UPDs and more fully measure evaluation effectiveness.

1.2 Goals

This research had three goals: (1) develop guidelines for describing usability problems, (2) explore the evaluator effect in usability testing, and (3) evaluate the usefulness of the guidelines for judging the content of usability problem descriptions. This research should lead to a better understanding of usability problem descriptions and the evaluator affecting usability testing, and provide a basis for future studies to formally develop metrics of usability problem description quality.

Research Question 1: How do practitioners describe usability problems?

Research Question 2: Is there an evaluator effect in usability testing?

Research Question 3: How can we assess the content of UPDs?

1.3 Approach

The first phase of this study focused on developing guidelines for describing usability problems. There is little discussion of UPDs in usability articles or textbooks to form the basis of guidelines. Instead, research used an exploratory approach to develop guidelines, consulting usability practitioners about important qualities of UPDs with a series of questionnaires.

The second phase focused on the evaluator effect in usability testing. There have been 10 previous studies of the evaluator effect in usability testing. CUE-1, -2, and -4 (Dumas et al., 2004; Molich et al., 1998; Molich et al., 2004; Molich et al., 1999; Rourke, 2003) and related studies (Kessner et al., 2001; Skov & Stage, 2003, 2004) allowed evaluators to select their own tasks and users. The remaining four studies controlled these factors with pre-recorded sessions or simultaneous live viewing. Three studies used four or fewer evaluators (Jacobsen et al., 1998; Palacio et al., 1993; Vermeeren et al., 2003) or did not report the results of individual evaluators (Lesaigle & Biers, 2000). Two additional studies using pre-recorded sessions and 20 or more participants were published after Study 4 completed (Long, Styles, Andre, & Malcolm, 2005; Skov & Stage, 2005). The current research used pre-recorded usability sessions and focused on collecting usability problem descriptions, rather than complete usability reports. This provided a consistent testing scenario and a narrower focus than the CUE studies, as illustrated in Figure 1.3.

The second phase also served as a preliminary assessment of use of the guidelines to assess the content of usability reports. The guidelines from Phase I were used to evaluate the usability reports collected in Phase II. Assessment measures included the extent to which practitioners followed the guidelines, if following the guidelines was an indicator of practical usability experience, and comparing opinions about the importance of the guidelines with behavior in terms of following the guidelines when writing the usability reports. Table 1.1 summarizes the goals and approach of this research, and Table 1.2 summarizes the phases and outputs.

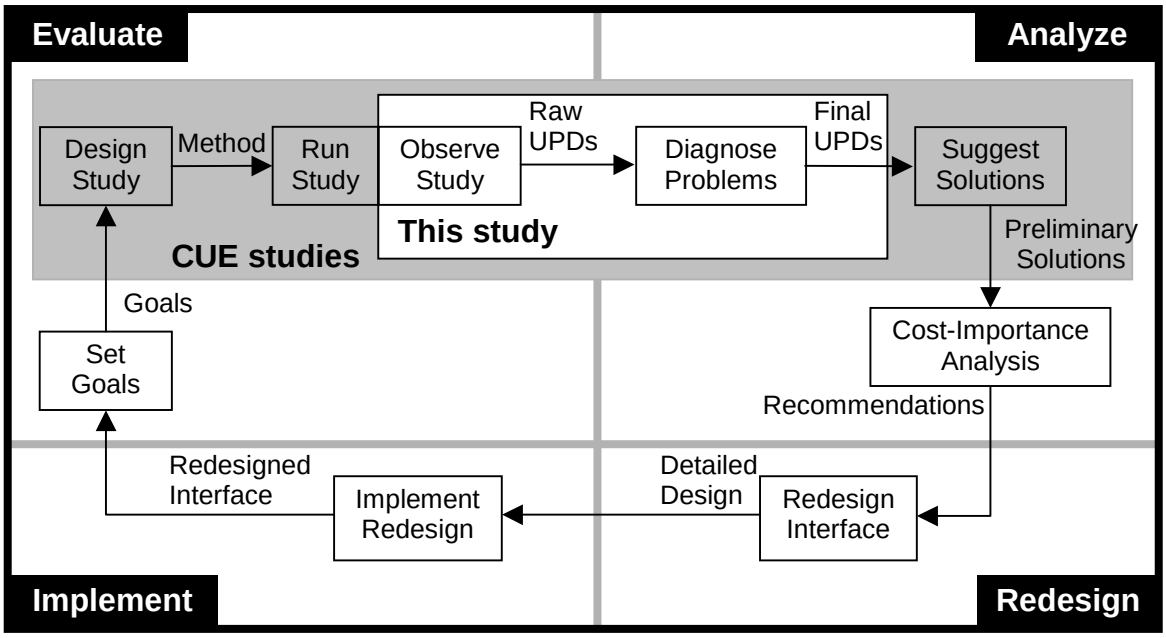


Figure 1.3 Comparison of Scope to CUE Studies

Table 1.1 Research problems, goals and approach

	Problem	**Goal**	**Approach**
1.	We need measures of usability problem description (UPD) content to be able to document shortcomings in UPDs and more fully measure evaluation effectiveness.	Develop guidelines for describing usability problems.	Consult experienced usability practitioners using a questionnaire.
2.	Previous studies of the evaluator effect on problem detection in usability testing have allowed different tasks and users, used a small sample size, or used student teams; we need larger studies of the evaluator effect that focus specifically on study analysis, rather than design and execution.	Explore the evaluator effect in usability testing.	Review UPDs collected from both practitioner and student evaluators conducting usability tests in a controlled setting.
3.	Evidence of poorly written UPDs is anecdotal. Previous studies of the evaluator effect in usability testing have focused on problem detection and severity judgments.	Evaluate the usefulness of the guidelines for judging the content of usability problem descriptions	Use the guidelines to rate usability reports collected from practitioner and student evaluators. Compare students to practitioners, and compare opinions of the guidelines to reporting behavior.

Table 1.2 Summary of Chapters, Phases and Studies

Chapter, Phase	Study	Output
CHAPTER 2. Phase I: Developing Guidelines for Describing Usability Problems	**Study 1:** Open-ended questionnaire 19 respondents	70 items to address in a UPD
	Study 2: Open card sort 8 participants	10 guidelines for UPDs
	Study 3: Questionnaire 74 respondents Rate the 10 guidelines for: *Helpful* vs. *Harmful* *Relevant* vs. *Irrelevant* *Required* vs. *Optional* *Difficult* vs. *Easy*	5 *required* guidelines Comparison of *required* vs. *difficult* guidelines
CHAPTER 3. Phase II: The Evaluator Effect in Usability Testing	**Study 4:** Evaluators • Watch pre-recorded session • Describe problems • Repeat study #3 questionnaire 21 practitioners 23 students Judges • Create master problem list (MPL) • Decide which MPL problems are "real" • Determine which MPL problems match each UPD, indicating detection of a problem • Rate each UPD for errors, vagueness • Rate each report for 5 required guidelines and *Describe a Solution*	***RQ1: How do usability practitioners describe usability problems?*** • Do practitioners follow the guidelines? • Does practitioner behavior match practitioner opinion? ***RQ2: Is there an evaluator effect in usability testing?*** • Is there an effect for problem discovery (thoroughness, validity, reliability) for either students or practitioners? • Are problem severity ratings reliable for either students or practitioners? ***RQ3: How can we assess the content of UPDs?*** • Do "good" evaluators follow the guidelines? • Are the report ratings reliable?
CHAPTER 4. Conclusion	Usability Testing Benefits from Multiple Evaluators. Is Usability Testing the "Gold Standard" of Evaluation? Measure Both Quantity and Quality.	

This page intentionally left blank.

CHAPTER 2. Phase I: Developing Guidelines for Describing Usability Problems

Formative usability evaluations are an important part of the usability life cycle, identifying usability problems present in an interface that designers should fix in the next redesign phase. This is in contrast to summative studies, where the goal is to measure usability, or how well an interface meets stated usability goals. The output of a formative usability evaluation is a set of usability problems (Hartson et al., 2003), but the usability test is not over when the last participant goes home. Communicating the results of the test (either through a written report or verbally) is an essential part of the usability testing process (Rubin, 1994). Evaluators need to understand the problems found, record them, and communicate them to the team that will redesign the interface. Figure 2.1 shows how formative usability evaluations fit into the usability life cycle.

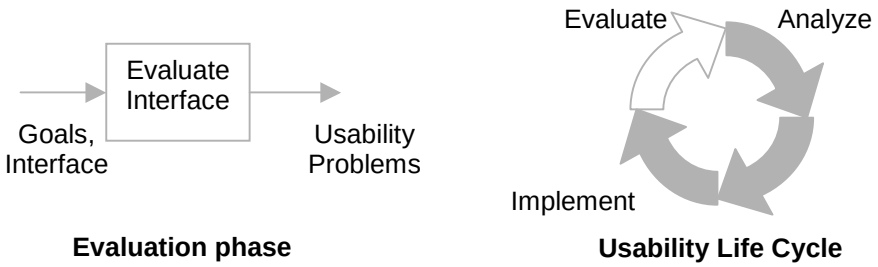


Figure 2.1 Formative Usability Evaluation and the Usability Life Cycle

Anecdotal evidence suggests that many evaluators create ineffective documentation to convey the results of formative usability evaluations. Andre, Hartson, Belz and McCreary (2001) reviewed hundreds of written usability problem descriptions contributed by professional usability laboratories for a study by Keenan, Hartson, Kafura and Schulman (1999), and describe many of them as "*ad hoc* laundry lists" (p. 108) that would require significant verbal communication to supplement the information provided in the written descriptions. Jeffries (1994) reviewed problem descriptions collected for a comparative usability study (Jeffries, Miller, Wharton, & Uyeda, 1991) and found that they often described solutions without describing the problem the solution addresses, or describe small examples of problems without pointing out the larger trend across many problems. Dumas, Molich and Jeffries (2004) reviewed 17 usability reports submitted to

the fourth of their Comparative Usability Studies series (CUE-4), and noted that reports had many qualities that would affect their acceptance by their intended audience, including being overly negative and critical, using jargon, and being too vague. Andre et al. express the opinion that poor documentation and communication of usability problems found diminishes the effectiveness of a usability evaluation and can reduce the return on the effort invested in conducting the evaluation. Dumas, Molich and Jeffries suggest that communication style and attitude in report writing can affect recipients' acceptance of suggestions and the number of the problems recipients choose to fix. Jeffries suggests that developers may interpret poorly described problems as false alarms, causing the developers to ignore the poorly described problem and increasing the likelihood that developers will treat future problems as opinion or false alarms.

Books on usability testing by researchers and practitioners in the field (Dumas & Redish, 1993; Hix & Hartson, 1993; Mayhew, 1999; Nielsen, 1993; Preece, 1994; Rubin, 1994; Stone, Jarrett, Woodroffe, & Minocha, 2005) provide guidance on observations to make during a usability testing session and the major sections to include in a usability report, but few specific recommendations about how to describe individual usability problems. Jeffries (1994) and Dumas et al. (2004) discuss the most common problems with usability reports, but do not summarize the areas that the reports covered well (they also share an author). Usability standards also provide little guidance. The Industry USability Reporting (IUSR) project of the National Institute of Standards and Technology (NIST) has a working group dedicated to creating a standard reporting format for formative usability evaluations, similar to the existing standard for summative studies, ANSI NCITS 354-2001 *Common Industry Format for Usability Test Reports* (ANSI, 2001). However, the latest report from this working group (Theofanos, Quesenbery, Snyder, Dayton, & Lewis, 2005) focuses on the entire usability report, with just a few line items for individual problem descriptions. A recent journal article (Theofanos & Quesenbery, 2005) discusses high-level issues about describing usability problems, such as including screen shots, quotes and positive comments, but also has few detailed guidelines for describing the problem cause and context.

The goal of this research was to identify important qualities of usability problem descriptions and develop guidelines for writing descriptions. Usability practitioners might benefit from guidelines for describing usability problems. Such guidelines could be particularly useful for practitioners-in-training, such as human-computer interaction (HCI) students and professionals that are new to usability. Beginner evaluators have little practical experience with the usability life cycle and the role of the usability report in ensuring that problems found through usability evaluation are fixed, and may be unaware of the information necessary to include in an effective problem description. Given the lack of existing guidelines, the approach was to consult usability practitioners through questionnaires. This research involved three studies, summarized in Table 2.1.

Table 2.1 Overview of Studies and Outputs

Study	Overview	Output
1	Open-ended questionnaire 19 respondents	70 items to address when writing usability problem descriptions
2	Card sort 8 participants	10 guidelines for describing usability problems
3	Questionnaire 74 respondents	5 *required* guidelines 6 *difficult* guidelines

2.1 Study 1: Open-Ended Questionnaire

The goal of this study was to gather an initial set of guidelines for describing usability problems.

2.1.1 Method

An open-ended questionnaire was used to gather important qualities of usability problem descriptions. The responses were coded for specific items to include in problem descriptions and suggestions for how to phrase or present descriptions.

2.1.1.1 Respondents

Nineteen respondents were recruited by sending email to the mailing list of the Usability Special Interest Group of the Society for Technical Communications, a private mailing list of usability professionals, and professional contacts. Respondents were required to have at least one of the following two requirements: five years of usability experience, or have conducted 10 usability evaluations. The respondents had 3-20 years of experience (M = 10.16, SD = 4.49). Eighteen of the participants had conducted between 10 and 1,000 usability evaluations (M = 133.61, SD = 244.13, Median = 50), and one participant had conducted 10,000 usability evaluations. Sixteen worked in industry, and one each in university, government and military; no respondents marked more than one of these areas. Respondents who met neither of the experience criteria (years of experience or number of evaluations) were excluded. Access to the website was not restricted in any fashion. However, the experience cutoffs were not mentioned in the questionnaire, so respondents had little incentive to misrepresent their level of experience.

2.1.1.2 Questionnaire

The questionnaire was posted on a public website and consisted of three pages:

1. Demographic questions
2. Background scenario, shown in Figure 2.2 (also used in Studies 2, 3 and 4), and open-ended questions about important qualities of UPDs
3. Open-ended questions about important skills and common mistakes in usability evaluation (optional).

For each important quality, the questionnaire prompted respondents to provide a name and a description (what it is, why it is important, and how it contributes to or detracts from the usability process), and to indicate whether it was a good or bad quality. The complete questionnaire is included in Appendix B.

Imagine the following scenario:

> You, a usability practitioner, are part of a team assessing the usability of a user interface (website, software, consumer product, etc.) that you may or may not have designed. The ultimate goal of your assessment is to develop a set of improvements to the interface.
>
> You have just completed a formative usability evaluation of the interface using your favorite method (usability testing with end-users, inspection, etc.). Your current goal is to generate a list of usability problem descriptions that summarizes the data collected during the study both from user interactions and from expert observations during the test.
>
> The people reading the descriptions (e.g. usability practitioners, developers, marketing, product development, management, clients) may or may not have usability training. They may use this list for many purposes, such as deciding which problems to fix in the next product release, designing interface changes, or adding features to the software.

Given this scenario, consider the individual usability problem descriptions that you will write. What should you include in these descriptions? What should you avoid? What makes one usability problem description better or worse than another problem description?

Figure 2.2 Questionnaire Scenario Used in Studies 1-4

2.1.2 Results

Respondents described a total of 99 qualities, with individual respondents describing 3 and 11 qualities (M = 5.21, SD = 2.10); the full text of the responses is included in Appendix B. The number of words in each description ranged from 4 to 77 (M = 24.88, SD = 14.09). These qualities varied greatly in level of granularity and specificity, with some respondents providing a brief description and others providing a long paragraph of description. Some of the qualities discussed usability problem descriptions, some discussed usability reports in general, and some discussed both. Two Ph.D. candidates in HCI coded the descriptions independently, a graduate student in HCI and the experimenter. Rather than consider each submitted description as a single entity, we coded each specific item mentioned in each description. We each read the text of the descriptions, creating a list of UPD qualities and using ATLAS.ti ("ATLAS.ti," 2000) to mark text that corresponded to each quality. The experimenter then merged the two lists, identifying duplicates and highly similar items. The result yielded a detailed list of 70 very specific items to include or address in usability problem descriptions, which was the input for the card sort in Study 2. This list of 70 items is included in Table 2.4.

2.2 Study 2: Card Sort of UPD Qualities

The goal of the second study was to identify a set of high-level guidelines for describing usability problems. The set of 70 items collected in the previous study (Table 2.4), while interesting, was too long and detailed for fast comprehension or practical use. Many of the items clearly related to each other, and provided detailed suggestions related to high-level issues such as describing the problem cause or users' actions. For example, "be clear and precise," "be concise, avoid wordiness," and "use precise terminology" are all closely related. This suggests that the 70 items are specific examples relating to more abstract concepts, and that the items can be grouped into a few high-level categories. However, preliminary groupings by the first author and two other HCI students were somewhat different, indicating that the groupings were not clear and obvious. A card sort was used to group the items, with each group forming the basis for one guideline for describing usability problems.

2.2.1 Method

In Study 2, usability practitioners performed a card sort on the 70 items collected in the first study to identify high-level themes within the items.

2.2.1.1 Participants

Eight participants were recruited from the same pool as Study 1 by sending email to the mailing list of the Usability Special Interest Group of the Society for Technical Communications, a private mailing list of usability professionals, and professional contacts. Participants in Study 2 may or may not have been the same as those for Study 1, since all responses were anonymous. Respondents were required to meet at least one of the following two criteria: four years of usability experience, or have conducted 10 usability evaluations. Table 2.2 summarizes the key participant demographics.

Table 2.2 Study 2: Participant Demographics

Years of Experience	Number of Evaluations	Area of Work
3	10	Industry
4	9	University
4	20	Industry
5	9	University
9	100	Government
10	200	Industry, University
15	50	Industry
12	100	Industry, Government, University

2.2.1.2 Card Sort Materials

The card sort was posted on a public website and consisted of three pages: demographic questions, the card sort, and a comment box. The card sort page began with the same scenario used in Study 1 at the top of the page (see Figure 2.2), and then had a text box pre-filled with all 70 items. The third page had a text box to collect comments from the participants. This website is reproduced in Appendix C.

2.2.1.3 Card Sort Procedure

Participants sorted the 70 items into categories. They could create one level of subcategories, and could duplicate items that belonged in multiple categories. They gave each category and subcategory a name. If they desired, they could copy the items into another application, such as a word processor, sort the items there, and then copy them back into the web page when done.

2.2.2 Results

Participants created a total of 77 categories and subcategories, with individual participants creating between 4 and 13 categories (M = 9.63, SD = 3.42). Figure 2.4 shows the complete list of categories and subcategories. While many categories were similar, the differences among the responses required formal analysis of the card sort data

to generate the final set of categories. Factor analysis on the 70 items was used to create 10 groups and a brief summary of each group. This section describes the steps for analyzing the card sort data: measuring similarity and creating a similarity matrix, running a factor analysis and selecting the final factors, identifying the participant-supplied category names that match each factor, and interpreting each factor based on the items (cards) that loaded on each factor and the matching participant-supplied category names.

2.2.2.1 Measuring Similarity

In order to run a factor analysis, a measure of the similarities between the 70 items (cards) in the card sort was needed. This measurement began by counting, for each pair of items, the number of categories and subcategories that contained both items, and also the number of categories and subcategories that contained at least one of the items, as shown in Table 2.3. The similarity of each pair of items was calculated using Jaccard's Coefficient of Communality (Jaccard, 1912), or the Jaccard Score. The Jaccard Score (J) of two items is the ratio of the count of their intersecting categories and subcategories to the count union of their categories and subcategories, as shown in Equation 2.1. Capra (2005) described this process in detail. The Jaccard scores were then assembled into a 70x70 similarity matrix.

Table 2.3 Study 2: Counting category memberships

Category or subcategory contains item 1?	Category or subcategory contains item 2? **Yes**	**No**
Yes	a	b
No	c	d

$$J = \frac{\text{intersection}}{\text{union}} = \frac{a}{a + b + c} \qquad (2.1)$$

2.2.2.2 Factor Analysis

A factor analysis of the card sort data (using the Jaccard score similarity matrix as a covariance matrix) generated the scree plot shown in Figure 2.3, with eigenvalues listed in Appendix C. The Kaiser-Guttman criterion selects all factors with eigenvalues of at least one, which results in 18 factors (discarding the first scree), but this technique tends to select too many factors. A variation is to retain all factors with eigenvalues greater than the mean eigenvalue, which results in 10 factors. This corresponds with the suggestion of Cattell (1966) to discard both screes, retaining the first point on the scree. Eighteen guidelines is too many to easily comprehend, so 10 factors were retained, explaining 53% of the variance in the data. A factor analysis using varimax rotation resulted in the memberships shown in Table 2.4. Varimax maintains the orthogonality of the factors, and tends to result in easily interpretable factors by generating loadings that are either very high or very low, minimizing the number of items needed to explain each factor, and the number of factors needed to explain each item (Kaiser, 1958). A loading of .40 or higher was used to determine factor membership, resulting in eight items that loaded on two factors and five items that loaded on zero factors.

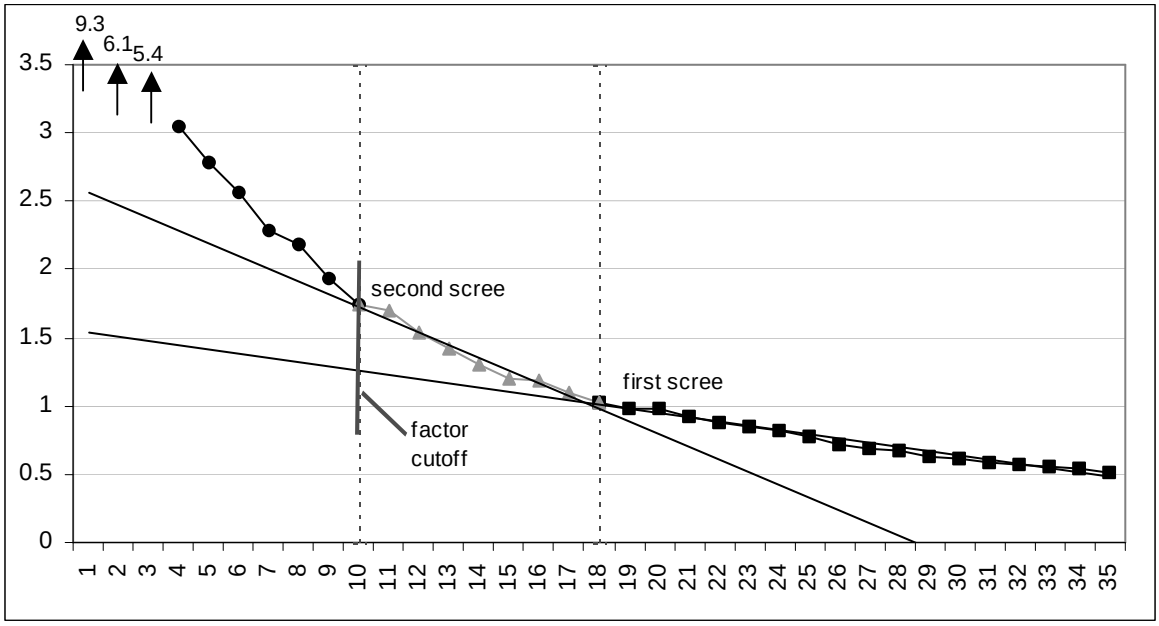


Figure 2.3 Study 2: Scree Plot for Factor Analysis

2.2.2.3 Matching Participant-Supplied Category Names to Factors

The card sort participants not only sorted the 70 items into categories and subcategories, but also supplied a name for each category and subcategory that they created, shown in Figure 2.4. Jaccard scoring (Jaccard, 1912) was used to identify the names that best matched each of the 10 factors. When scoring groups, rather than individual items, the Jaccard score is the ratio of the number of intersecting items between the two groups to the number of items in the union of the two groups. The Jaccard score was calculated between each factor and each of the participant-supplied category and subcategory names, and those with the highest Jaccard score are the best matches. Capra (2005) describes this process in detail. Table 2.4 shows the participant-supplied category names that match each factor and their Jaccard scores (multiplied by 100).

2.2.2.4 Summarizing Factors

A summary of approximately 40 words was created for each factor. The first sentence of the summary was based on the best matching participant-supplied category names for the factor, and the highest loading items. The remainder of the summary was based on the remaining items that loaded on the factor. This set of summaries became the 10 guidelines for writing UPDs used in Studies 3 and 4. These guidelines are shown next to the factor members and participant-supplied category names in Table 2.4.

Procedural Requirements:
Scientific soundness
Scientific soundness and descriptiveness to others
Repeatability, to obtain validity

Quality of recommendations -- accuracy
About the test and participants
Quality of result explanation -- thoroughness
Justification to address the problem: (to provide justification to the stakeholders that convinces them they should listen to you.)
Prioritization of findings: (to prioritize the findings for stakeholders)
Next steps

Characterization: Nature of Problem
Characterization: Severity
Elucidation
Foreseen Effects
General Style
Probable Causes
Professional Technique
Raw Data
    Qualitative
    Quantitative
Recommendation
    Design Iteration
    Political Considerations

Providing a solution
Wording
Language
Reporting a specific problem
    Background
    Description
    Explanation
    Effects
    Decision
Describing the evaluation in general
Reflecting on the larger issues of a problem
Report guidelines

General
    Be clear and precise
    Recommendations for content
Methodology
General Results
Problems
    The problems
    Severity of the Problems
    Cause
    Data To Support Findings:
Solutions
Conclusions

Style
Background
System
User-centered
Problem Descriptions - Data
Etiology
Problem descriptions
Recommendations

Terminology
Data
    Quantitative data
    Qualitative data
Problem context
Problem descriptions
Causes of the problem
Effects of the problem
Solutions to the problem

Background
General
    Precision
    Use data
    Write to your reader
Problem
    Cause
    Describe
    Impact
    Scope
Research
Suggesting solutions
Users

Figure 2.4 Study 2: Card Sort Category Names Supplied By Eight Participants

Table 2.4 Study 2: 70 Card Sort Items: Factors and Category Names*

Factor	Loading, Item Text	Jaccard Score, Category Name*	Summary
1	94 Describe a potential solution to the problem 94 Include several alternate solutions, if possible 94 Describe advantages and disadvantages of alternative solutions 86 Avoid dictating a specific solution - don't alienate other team members 77 Don't guess about a good design solution 74 Be specific enough about a solution to be helpful, without jumping to a design conclusion 54 Use pictures/captures of interface to describe a suggested solution 50 Make sure that you impart good design principles (also 8) 49 Don't jump to conclusions and settle on a specific solution 45 Mention usability practices/previous research to back your suggestions (also 8)	90 Providing a solution 75 solutions to the problem 64 Recommendations 58 suggesting solutions 47 Quality of recommendations -- accuracy 36 [Recommendation] Design Iteration	**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.
2	91 Be concise, avoid wordiness 88 Use precise terminology 85 Avoid jargon or technical terms 80 Define any terms that you use 78 Don't use vague terms and descriptions; be concrete 73 Be clear and precise 61 Be pragmatic/practical; avoid theories/jargon that non-HCI people wouldn't appreciate (also 9) 40 Avoid so much detail that no one will want to read to description (also 5)	87 Terminology 75 [General] Be clear and precise 72 Style 66 Wording 53 General Style 46 General 37 [General] precision 33 General 20 Quality of result explanation -- thoroughness	**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 people will appreciate. Avoid so much detail that no one will want to read the description.

* Table continues on the next page

Brackets [] indicate name of parent category for a subcategory

Table 2.4 Study 2: 70 Card Sort Items: Factors and Category Names (cont.)*

Factor	Loading, Item Text	Jaccard Score, Category Name*	Summary
3	87 Describe the problem's cause 87 Include the cause of the problem 68 Describe how the interaction architecture contributed to the problem 63 Describe the main usability issue involved in this problem 54 Avoid guessing about the problem cause or user thoughts 43 Mention the context of the problem, such as the user's task	66 [Problems] Cause 55 Causes of the problem 44 Etiology 42 [Problem] cause 20 Problems 18 Quality of result explanation -- thoroughness 17 Problem 17 Reporting a specific problem	**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.
4	69 Describe how many users that experienced the problem 68 Mention how often the problem occurred during testing 67 Use quantitative data to support your arguments 67 Mention the number of task attempts 52 Include time spent on the task 51 Mention whether the user succeeded or failed at the task (also 7) 47 Describe critical incidents 45 Provide traceability of problems to observed data 42 Use supporting data (also 7) 42 Include objective data from the study to support your arguments	70 [Problems] Data To Support Findings 60 [Raw Data] quantitative 53 Raw Data 47 [Reporting a specific problem] Description 45 [Data] Quantitative Data 41 Data 27 Reporting a specific problem 25 Problems	**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.

* Table continues on the next page

Brackets [] indicate name of parent category for a subcategory

Table 2.4 Study 2: 70 Card Sort Items: Factors and Category Names (cont.)*

Factor	Loading, Item Text	Jaccard Score, Category Name*	Summary
5	76 Use anecdotes to make the problem seem real, promote sympathy 72 Be evocative, help the reader understand/sympathize with what happened 64 Make sure the description is readable/understandable 62 Use user-centered descriptions rather than system-centric 54 Avoid so much detail that no one will want to read to description (also 2) 44 Make sure that you are complete in your descriptions (also 7) 43 Use a user-centric perspective 42 Include sufficient detail to understand exactly what happened (also 7) 40 Avoid just listing the heuristic violated; explain why it's a problem	64 Problem descriptions 54 Language 40 [General] write to your reader 26 General 16 Quality of result explanation -- thoroughness	**Help the reader sympathize with the user's problem** by using descriptions that are evocative and anecdotal. Make sure the description is readable and understandable. Use user-centric language rather than system-centric. Be complete while avoiding excessive detail.
6	78 Describe the impact of the problem 72 Describe the business effects - support costs, time loss, etc. 65 Describe the importance of the task the user was performing 53 Describe exactly which system components are affected/involved 56 Mention how often the problem will occur during usage 45 Describe the eventual outcome of the user's actions	71 [Problem] impact 55 [Problems] Severity of the Problems 50 [Reporting a specific problem] Effects 50 Foreseen Effects 45 Effects of the problem 21 Problem 16 Problems 13 Reporting a specific problem	**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.

* Table continues on the next page

Brackets [] indicate name of parent category for a subcategory

Table 2.4 Study 2: 70 Card Sort Items: Factors and Category Names (cont.)*

Factor	Loading, Item Text	Jaccard Score, Category Name*	Summary
7	66 Described observed behaviors 65 Use specific examples from the study 53 Describe the user's navigation flow through the system 50 Describe the users' subjective reactions 46 Use pictures/screen shots of the user interface to describe the problem 46 Use supporting data (also 4) 45 Make sure that you are complete in your descriptions (also 5) 43 Include sufficient detail to understand exactly what happened (also 5) 41 Mention whether the user succeeded or failed at the task (also 4) 40 Mention whether the problem was user reported or experimenter observed	76 [Problem] describe 47 [Reporting a specific problem] Description 35 Problem 32 Reporting a specific problem 27 [Raw Data] qualitative 19 Quality of result explanation -- thoroughness	**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.
8	60 Use only facts from the study, rather than your opinions or guesses 58 Include a category from an outside taxonomy/classification scheme 55 Avoid your own opinions or subjective statements 51 Make sure that you impart good design principles (also 1) 43 Mention usability practices/previous research to back your suggestions (also 1)	50 Research 31 Professional Technique 25 Quality of recommendations -- accuracy	**Be professional and scientific in your description.** Use only facts from the study, rather than opinions or guesses. Back your findings with sources beyond the current study, such as external classification scheme, proven usability design principles, and previous research.

* Table continues on the next page

Brackets [] indicate name of parent category for a subcategory

Table 2.4 Study 2: 70 Card Sort Items: Factors and Category Names (cont.)

Factor	Loading, Item Text	Jaccard Score, Category Name*	Summary
9	63 Avoid judging the system or decisions made by other team members 61 Avoid pointing fingers or assigning blame 54 Mention good design elements and successful user interactions 41 Be pragmatic/practical; avoid theories/jargon that non-HCI people wouldn't appreciate (also 2)	57 [Recommendation] Political Considerations 42 Report guidelines 33 Recommendation 25 General 20 general	**Consider politics and diplomacy** when writing your description. Avoid judging the system, criticizing decisions made by other team members, pointing fingers or assigning blame. Point out good design elements and successful user interactions. Be practical, avoiding theory and jargon.
10	66 Mention the testing context (laboratory, field study, inspection, etc.) 61 Describe limitations of your domain knowledge 57 Describe the user groups that could be affected during system usage 48 Describe the user groups that were affected during testing 41 Describe the breadth of components of the system involved in the problem	66 Background 60 Methodology 50 Characterization: Nature of Problem 40 Users 20 Repeatability, to obtain validity	**Describe your methodology and background.** Describe how you found this problem (field study, lab study, expert evaluation, etc.). Describe the limitations of your domain knowledge. Describe the user groups that were affected and the breadth of system components involved.
	Items that did not load on any factors		
	Avoid misleading statistics and presentation of results		
	Include definitions of severity/importance/impact to avoid confusion		
	Mention the testing context (laboratory, field study, inspection, etc.)		
	Mention usability practices/previous research to back your explanations		
	Mention whether or not this problem should be fixed		

Brackets [] indicate name of parent category for a subcategory

2.3 Study 3: Importance and Difficulty of 10 UPD Guidelines

In the previous study, 10 guidelines for describing usability problems were developed. Which of these are the most important to follow? Which are the most difficult? Given the limited time generally available to write usability reports, knowing where to prioritize and focus efforts could be helpful to a usability practitioner. Knowing which are the most difficult could be helpful in designing usability training or modifying usability methods and problem reporting templates. The goal of this third study was to identify the guidelines that were the most important, as well as the guidelines that were the most difficult.

2.3.1 Method

Participants completed a questionnaire to rate each of the 10 guidelines from the card sort for difficulty and importance. Difficulty was rated on a scale of *difficult/easy*. Importance was rated on three scales: *required/optional*, *relevant/irrelevant* and *helpful/harmful*; participants were not provided with a definition of these adjectives. All of these adjective pairs are antonyms listed in Roget's Thesaurus (*Roget's New Millennium™ Thesaurus*, 2005). The selection of *helpful/harmful* was spurred by several card sort participants who commented that they did not agree with all of the 70 items they sorted, or described some of them as "wrong." In particular, several disagreed with the suggestions to provide a solution along with the problem description. *Unhelpful* was rejected as the opposite of *helpful* in favor of *harmful*. *Unhelpful* can imply merely a lack of helpfulness and has a somewhat neutral connotation, whereas *harmful* has a negative connotation and is more strongly the opposite of *helpful*.

2.3.1.1 Respondents

Seventy-four respondents were recruited from the same pool as Studies 1 and 2 by sending email to the mailing list of the Usability Special Interest Group of the Society for Technical Communications, a private mailing list of usability professionals, and professional contacts. Respondents in Study 3 may or may not have been the same as those for Studies 1 and 2, since all responses were anonymous. Sixty respondents were

from industry, 16 from university, four from government, and four listed other sectors; seven respondents marked multiple responses. Respondents had to meet at least one of the following two criteria: five years of usability experience, or have conducted 10 usability evaluations. Respondents had 2-25 years of experience (M = 8.88, SD = 5.26, Median = 7) and had conducted 6-800 usability evaluations (M = 94.99, SD = 141.90, Median = 42.5). Table 2.5 summarizes the distribution of respondent experience, with respondents broken into quartiles for both years of experience and number of evaluations. Access to the website was not restricted in any fashion. However, the experience cutoffs were not mentioned in the questionnaire, so respondents had little incentive to misrepresent their level of experience.

Table 2.5 Study 3: Frequency Distribution of Respondent Experience

		Years of Experience				
		2 - 4	**5 - 7**	**8 - 12**	**13 - 25**	**Total**
Number of Evaluations	**6 - 18**	5	6	3	2	16
	20 - 40	11	5	3	2	21
	45 - 100	2	7	5	8	22
	120 - 800	1	1	7	6	15
	Total	19	19	18	18	74

2.3.1.2 Questionnaire

The first page of the site contained the same demographic questionnaire used in Study 1. The second page began with the same scenario used in Study 1 at the top of the page (see Figure 2.2). The page then displayed a brief version of each of the 10 guidelines developed in Study 2, consisting of the first few words of each quality description, shown in bold in Table 2.12. The questionnaire then presented the full text of each guideline and a semantic differential rating scale (Osgood, Suci, & Tannenbaum, 1957) for each adjective pair. Figure 2.5 shows a sample guideline with the four rating scales. The presentation order for each respondent was randomized for the 10 guidelines, the four adjective pairs, and the left/right placement of each pair of adjectives, but these orders were constant for an individual respondent across all 10 guidelines. The complete questionnaire is included in Appendix D.

1. 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.

When a usability practitioner writes a description of a single usability problem, addressing this quality of a problem description is generally...

required	O	O	O	O	O	O	O	optional
helpful	O	O	O	O	O	O	O	harmful
easy	O	O	O	O	O	O	O	difficult
irrelevant	O	O	O	O	O	O	O	relevant

Figure 2.5 Study 3: Rating Scale Layout

2.3.2 Results

The primary factor of interest was differences among the 10 guidelines for each of the four adjective pairs. Figure 2.6 shows the frequency counts of respondent ratings for each guideline and adjective pair. Also of interest were effects due to level of experience, both in terms of years of experience and number of evaluations conducted. Two supplemental ANOVAs were conducted to assess effects due to the presentation of the ratings scales, one for the left/right order in which each adjective was presented at the ends of the semantic differential rating scale, and one for the order that the four adjectives were presented to the respondent. Four repeated measures ANOVAs, one for each adjective pair, were performed to test for differences among the 10 guidelines and effects due to experience. Appendix D shows the SAS code for all three ANOVAs and relevant output. All ANOVAs used type three sums of squares. Post-hoc tests used least square means, Tukey multiple comparisons test and adjusted *p*-values (Tukey-Kramer for unbalanced comparisons). All tests of statistical significance used an alpha level of .05.

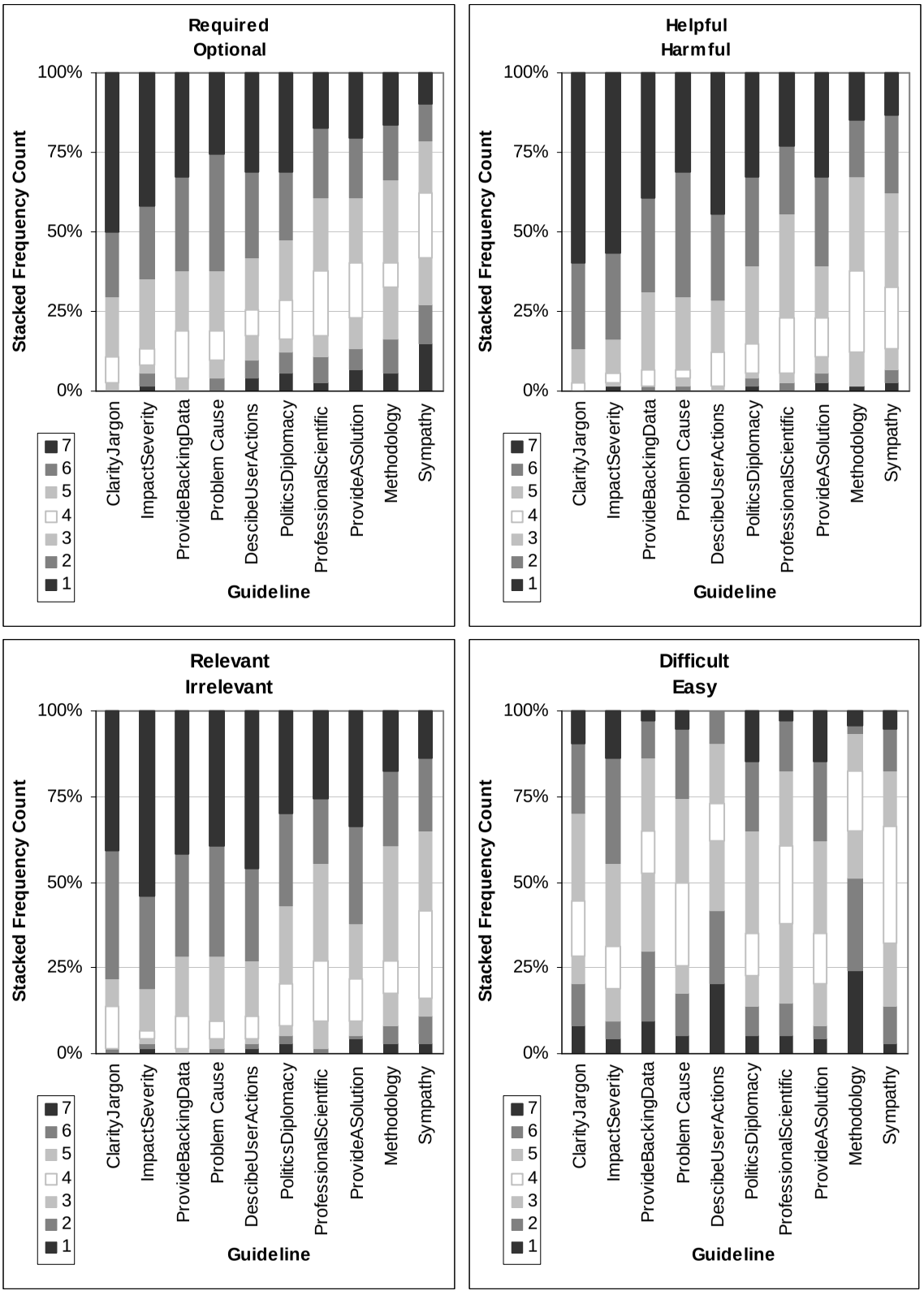


Figure 2.6 Study 3: Stacked Frequency Counts for Each Adjective, Guideline

2.3.2.1 Supplemental ANOVAs for Presentation Effects

Left-right order of each pair. Four ANOVAs, one for each of the adjective pairs, were used to assess if there were any effects due to the left-right presentation order of the adjective pairs. For example, did the respondents give a different *difficulty* rating when presented with *difficult-easy* versus *easy-difficult*. Using an alpha level of .05, there were no significant effects due to left-right presentation order for any of the four adjective pairs: $F(1, 72) = 0.01$, $p = .92$ for *difficult*; $F(1, 72) = 0.07$, $p = .80$ for *helpful*; $F(1, 72) = 2.10$, $p = .15$ for *relevant*; and $F(1, 72) = 1.08$, $p = .30$ for *required*.

Ordering of the four adjective pairs. Four separate ANOVAs, one for each of the adjective pairs, were used to assess if there were any effects due to the position of the adjective pair within the four pairs. Since adjective ordering was randomly assigned, the four adjective pairs were not equally represented at each position. Figure 2.7 displays the mean ratings and standard error for each of the four adjective pairs at each of the four positions. The numbers below each column indicate the number of participants that saw that adjective at that position (for an individual participant, the adjectives were shown in the same order for all 10 guidelines). Using an alpha level of .05, there were no significant differences for *difficult*, *helpful* or *relevant*, but there were for *required*: $F(3, 70) = 0.61$, $p = .61$ for *difficult*; $F(3, 70) = 0.76$, $p = .52$ for *helpful*; $F(3, 70) = 2.10$, $p = .11$ for *relevant*; and $F(3, 70) = 4.27$, $p < .01$ for *required*. Table 2.6 summarizes the *required* ratings for each adjective position.

Table 2.6 Study 3: Significant Order Effects for *Required*

Position	*M*	SD	Tukey Groups
1	**5.49**	1.55	A
2	**5.20**	1.60	AB
3	**4.71**	1.82	B
4	**5.35**	1.67	A

Note: Means that do not share a common letter differed significantly in post-hoc tests ($p < .05$).

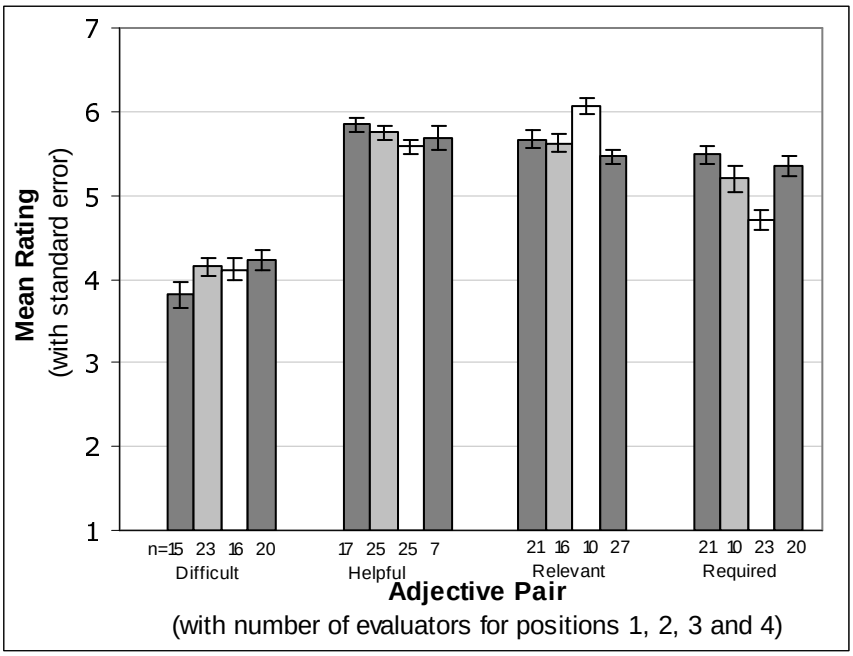


Figure 2.7 Study 3: Mean Ratings for Each Adjective at Each Position

2.3.2.2 Results for Required, Helpful *and* Relevant

Three of the adjective pairs were examined as aspects of importance: *relevant/irrelevant*, *required/optional*, and *helpful/harmful*. There were moderate correlations among the three adjective pairs, as shown in Table 2.7. This resulted in similar, but not identical, rankings across the three adjective pairs.

Table 2.7 Study 3: Correlation between *Relevant, Required, Helpful*

	Helpful	**Required**
Relevant	.75	.72
Required	.69	

$r(72)$, $p < .0001$

There were significant differences in the ratings for *required*, *helpful* and *relevant* across the 10 guidelines, $F(9, 603) = 15.61$, 15.02, and 11.36, respectively, $p < .0001$ for each. The Tukey multiple comparisons post-hoc tests identified guidelines that are not significantly different from each other. Table 2.8 summarizes the participant ratings of each of the 10 guidelines for each of the three importance adjective pairs, with the letters

indicating the similarity group(s) for each of the 10 guidelines, and a line separating group A from the rest of the items.

Table 2.8 Study 3: Contrasting Ratings for *Required, Helpful, Relevant*

	Required: 7 Optional: 1			Helpful: 7 Harmful: 1			Relevant: 7 Irrelevant: 1		
Guideline	***M***	**SD**	**Tukey Groups**	***M***	**SD**	**Tukey Groups**	***M***	**SD**	**Tukey Groups**
Clarity/Jargon	**6.07**	1.13	A	**6.43**	0.80	A	**6.03**	1.10	AB
Impact/Severity	**5.78**	1.44	AB	**6.30**	1.08	A	**6.20**	1.19	A
Backing Data	**5.72**	1.19	AB	**5.99**	1.04	AB	**6.01**	1.05	AB
Problem Cause	**5.55**	1.34	ABC	**5.89**	1.05	ABC	**5.96**	1.13	ABC
User Actions	**5.32**	1.73	ABCD	**6.03**	1.09	AB	**6.00**	1.26	ABC
Politics/Diplomacy	**5.22**	1.79	BCD	**5.68**	1.33	BC	**5.50**	1.46	BCD
Professional/Scientific	**4.88**	1.60	CD	**5.36**	1.24	CD	**5.32**	1.33	CDE
Describe a Solution	**4.76**	1.80	D	**5.51**	1.54	BCD	**5.55**	1.55	BCD
Describe Methodology	**4.55**	1.80	DE	**4.95**	1.30	D	**5.01**	1.53	DE
Evoke Sympathy	**3.85**	1.87	E	**4.96**	1.44	D	**4.77**	1.54	E

Note: Means that do not share a common letter differed significantly in post-hoc tests ($p > .05$)

2.3.2.3 Results for Difficulty Ratings

Using an alpha level of .05, there were no significant correlations between *difficulty* and each of the adjectives *required*, *helpful* and *relevant*. Table 2.9 summarizes these correlations. There were significant differences in the *difficulty* ratings for the 10 guidelines, $F(9, 603) = 18.52$, ($p < .0001$). Tukey multiple comparisons post-hoc tests resulted in groups of guidelines that were not significantly different from each other. Table 2.10 summarizes the *difficulty* ratings for the 10 guidelines.

Table 2.9 Study 3: Correlations Between *Difficult* and *Required, Helpful, Relevant*

	r(72)	*p*
Helpful	-.05	.21
Relevant	-.00	.80
Required	-.05	.20

Table 2.10 Study 3: Mean Ratings for Difficult/Easy

Guideline	Difficult: 7 Easy: 1 *M*	SD	Tukey Groups
Impact/Severity	**4.95**	1.59	A
Provide a Solution	**4.85**	1.58	A
Politics/Diplomacy	**4.73**	1.69	AB
Clarity/ Jargon	**4.38**	1.76	ABC
Problem Cause	**4.32**	1.59	ABC
Evoke Sympathy	**4.08**	1.42	BCD
Professional/Scientific	**4.01**	1.48	CD
Backing Data	**3.59**	1.63	DE
User Actions	**3.12**	1.65	E
Describe Methodology	**2.88**	1.66	E

Note: Means that do not share a common letter differed significantly in post-hoc tests ($p < .05$).

2.3.2.4 Effects of Experience

The demographic portion of the questionnaire collected two aspects of experience: years of usability experience and number of usability evaluations conducted. Individual correlations between each of these variables and ratings for each of the four adjective pairs for each of the 10 guidelines were all non-negative ($0 < r < 0.3$); Table 2.11 shows the correlations that were significant. Inspection of the data indicated that the distributions of each of the experience variables were not linear, and that the relationships between these variables and the ratings were not monotonic, suggesting that an ANOVA would be better for testing differences.

Table 2.11 Study 3: Significant Correlations with Experience

Adjective…	… for Guideline	Experience Variable	*r*(72)	*p*
Difficult	Professional/Scientific	Number of Evaluations	.26	.02*
Relevant	Evoke Sympathy	Years of Experience	.24	.04*
Required	Evoke Sympathy	Years of Experience	.27	.02*

*$p < .05$.

The participants were divided into quartiles by years of experience (2-4, 5-7, 8-12, 13-25) and number of evaluations (6-18, 20-40, 45-100, 120-800), with frequency counts for each group shown in Table 2.5. These quartiles were used as independent variables in the ANOVA shown in Appendix D to check for both main effects and interactions with the 10 guidelines. Interactions between years of experience and number of evaluations could not be tested due to small cell counts (see Table 2.5). Using an alpha level of .1, there were no significant effects due to either type of experience for any of the four adjectives.

2.4 Discussion

Table 2.12 lists the complete text of all 10 guidelines sorted from most to least *required*, and Figure 2.8 illustrates the relationship between mean ratings of *required* and *difficult* for each of the 10 guidelines. Guidelines in Quadrant I, *Required/Difficult*, include items that generate much discussion in the usability community: rating the impact or severity of the problem (*Impact/Severity)* and determining the cause of a usability problem (*Problem Cause*). *Impact/Severity* had the highest mean difficulty rating, which might contribute to unreliability in severity judgments (for a summary of studies about unreliability in severity judgments, see Hertzum & Jacobsen, 2003). Severity judgments rely greatly on evaluator experience and skill, based as they are on a combination of observation of users in the laboratory, extrapolation to effect on the user, and occurrence in the larger user population. Problem classification systems like the User Action Framework (Andre et al., 2001) have been proposed to help evaluators diagnose usability problems thoroughly and identify the root cause of usability problems, but greatly increase the skill level required of the evaluator and time commitment needed for problem analysis. Also falling into Quadrant I is being clear while avoiding jargon (*Clarity/Jargon*), one of the recurring problems observed by Dumas, Molich, and Jeffries (2004) in the CUE-4 reports. The other problems discussed in that article fall under the *Politics/Diplomacy* guideline, which was rated higher in *difficulty* and in the middle in terms of being *required*. In contrast, items Quadrant II, *Required/Easy*, are the descriptive items that are more straightforward to present: *User Actions* and *Backing Data* (justifying problems with data from the study).

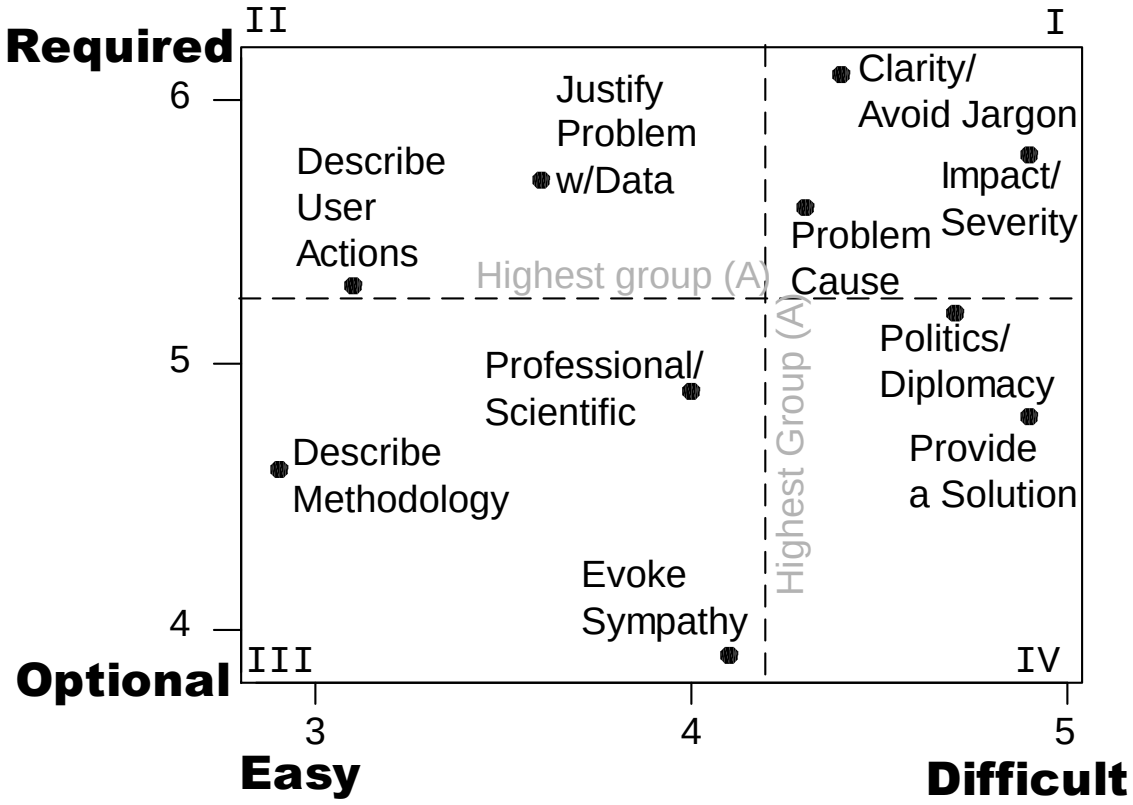


Figure 2.8 Study 3: *Required* vs. *Difficult* Guidelines

Table 2.12 Ten Guidelines for Describing Usability Problems

Guideline number (most to least *required*) and summary	
1	**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 people will appreciate. Avoid so much detail that no one will want to read the description.
2	**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.
3	**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.
4	**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.
5	**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.
6	**Consider politics and diplomacy** when writing your description. Avoid judging the system, criticizing decisions made by other team members, pointing fingers or assigning blame. Point out good design elements and successful user interactions. Be practical, avoiding theory and jargon.
7	**Be professional and scientific in your description.** Use only facts from the study, rather than opinions or guesses. Back your findings with sources beyond the current study, such as external classification scheme, proven usability design principles, and previous research.
8	**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.
9	**Describe your methodology and background.** Describe how you found this problem (field study, lab study, expert evaluation, etc.). Describe the limitations of your domain knowledge. Describe the user groups that were affected and the breadth of system components involved.
10	**Help the reader sympathize with the user's problem** by using descriptions that are evocative and anecdotal. Make sure the description is readable and understandable. Use user-centric language rather than system-centric. Be complete while avoiding excessive detail.

2.4.1 Solutions

Many questionnaire respondents commented on the *Describe a Solution* guideline, both in agreement and in disagreement. Several respondents cautioned against solutions that address individual problems, and recommended waiting to develop solutions when the entire team could consider all the problems together and brainstorm ideas as a group. Others seemed to take more of a consulting view that solutions are an essential part of a complete usability evaluation package. In order to explore these differences in opinion further, an optional comment field was added to the post-task questionnaire for Study 4, asking evaluators whether or not they included solutions with their reports and why.

2.4.2 Applications

These guidelines will be useful for usability practitioners, instructors and researchers.

- Usability practitioners can use this list to create usability problem reporting forms, to create checklists of items to address in writing problem descriptions, or to evaluate usability reports generated in usability studies. Usability groups could evaluate their work products to ensure that they are writing effective problem descriptions in usability reports.
- Usability instructors can use them to explain what should be in a usability problem description and to grade usability evaluations conducted by students. Following as many of the guidelines as possible would be an appropriate exercise for students, where thoroughness of the report is more important than the time spent on it. Usability students need to practice writing complete descriptions so that they will learn what information they need to take note of during a usability evaluation. Training of new practitioners could include more practice and review opportunities for the guidelines rated as more *difficult*.

- Usability researchers can use the guidelines to assess problem descriptions generated by different usability practitioners or usability evaluation methods. Current research in UEM effectiveness focuses on counting usability problems identified in evaluations, but more effort should be focused on the quality of descriptions as well as the quantity. This was explored further in Phase II.

The most important suggestion for using the guidelines is to *select the guidelines that fit your project and organization.* Many usability practitioners responding to the questionnaires commented that it was difficult to give a single rating to the guidelines because they write different reports for different audiences, or have used different styles of reports when working for different companies. Some organizations require scientific backing for all reports, while others prefer a more anecdotal style. Developers might prefer system-centric language because it is what they understand, but other usability practitioners might appreciate HCI jargon that precisely describes the design issues. Reports may differ from one product to another, even within the same organization. Dumas (1993) describes two different kinds of usability reports: the brief summary presented immediately after the evaluation is completed, and the more detailed report submitted several weeks later. Different guidelines are important for the different styles of reports. Review the guidelines and decide which guidelines, and individual parts of each guideline, are most appropriate for your organization, project and readers.

Some aspects of the guidelines may be equally or more important for a usability report as a whole, rather than an individual usability problem description. For example, a usability report may have a section describing tasks used in the study, and then each problem description can just reference the task number. Solutions may address multiple problems, and a separate section for solutions may be appropriate. Similarly, multiple problem descriptions can reference the same screen capture. This work complements the August 2005 report by the working group developing a standard for formative usability evaluation reports (Theofanos et al., 2005). This standard will be similar to the reporting standard for summative studies, ANSI NCITS 354-2001 *Common Industry Format for Usability Test Reports* (ANSI, 2001). When writing a report describing the results of a

formative usability evaluation, the guidelines presented in this section would be useful for writing the section describing individual usability problems.

2.5 Limitations

The respondents to the two questionnaires were recruited from groups of usability engineers and members of the Society for Technical Communication that are interested in usability. Other stakeholders in the usability process (managers, developers, marketing, product support, etc.) may have different opinions about what is important in a usability problem description. Many respondents also mentioned that they tailor their reports depending on who their audience is, and so different guidelines may require different levels of emphasis depending on the readers of the report (usability team members, consulting clients, developers, etc.). The respondents were self-selected, and so there may be a bias due to collecting responses only from people willing to respond to an email solicitation.

The card sort in Study 2 involved eight participants. Since the sort was conducted, Tullis and Wood (2004) have suggested 20-30 people as an appropriate size for a card sort. Their conclusions are based on card sorts by 168 employees of Fidelity Investments of 45 items for the company intranet. They performed simulations of card sorts using smaller groups, and found that the dendogram created by a card sort with eight participants would have a correlation of approximately .82 with the result from all 168 participants. This suggests that that the results of Study 2 should be similar to the results had more users participated in the card sort. Since the goal of Study 2 was to create high-level categories, rather than precisely place individual web pages in a website, this is an acceptable level of approximation. Study 2 used factor analysis to create soft (overlapping) clusters, whereas Tullis and Woods used hierarchical cluster analysis to create hard (distinct) clusters. Hard clusters may be more sensitive to small changes than soft clusters because items either do or do not belong to a cluster, whereas soft clusters include a continuous measure of association for each item with each group.

2.6 Conclusion

The output of Phase I was a set of 10 guidelines for describing usability problems. The 10 guidelines presented in this section are suggestions, not rules. It would be difficult and time-consuming to follow every single guideline for every single problem description. Different guidelines may be more or less important for different projects, clients, and organizations. Students may benefit from thoroughly addressing each guideline to practice observation, note taking, and problem description skills.

Phase II focused on the evaluator effect in usability problem detection, but also used these guidelines to asses both opinions about describing problems and behavior in describing problems. Both students and practitioners were involved to make possible comparison across a wider range of experience levels. It was surprising in Study 3 to find so few significant differences in opinion about the guidelines based on practitioner experience. This could be because all of these practitioners already had some practical experience, or because individual differences in opinion are stronger than any trend due to increased experience. The students in Study 4 can be compared to the practitioners to see if there are experience effects on opinion when less experienced evaluators are included. Study 4 was also used to assess whether the guidelines are useful in judging the content of usability problem descriptions, that is, whether following the guidelines is associated with better usability problem reports.

This page intentionally left blank.

CHAPTER 3. Phase II: The Evaluator Effect in Usability Testing

The primary goal of the second phase of this research was to explore the evaluator effect in usability testing. The *evaluator effect* refers to differences in problem detection and severity judgments by evaluators using the same UEM (Hertzum & Jacobsen, 2003; Jacobsen et al., 1998). This phase consisted of a single study, with many evaluators examining the same website. The following approach was used.

1. The test setting was controlled by using pre-recorded test sessions.
2. Analysis activities were limited by collecting just usability problem descriptions, without asking for an executive summary and recommendations.
3. Usability practitioners were used as evaluator participants to increase the realism of the results. Student evaluators were also included for comparison.
4. Evaluation thoroughness was measured both by counting problems detected and by assessing which problems are "real" (Hartson et al., 2003).
5. A "usability problem" was defined as a problem experienced by the user and caused by an interaction flaw.
6. Experimenter bias was reduced by using independent judges to match descriptions of the same problems and to determine which problems are real.

For this study, a usability problem was defined as a problem experienced by the user, which is caused by an interaction flaw, as illustrated in Figure 3.11. A description that only contains a suggestion for a solution implies the problem that it is fixing. Two problems, A and B, were considered the same if fixing problem A also fixes problem B **and** fixing problem B also fixes problem A. This is based on the criteria used in the analysis of usability reports collected in CUE-4 (Molich & Dumas, 2006).

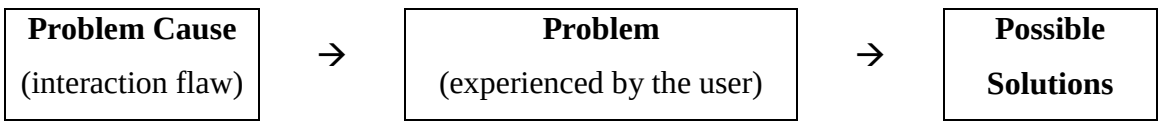


Figure 3.1 Study 4: Usability Problem Definition

Providing pre-recorded test sessions and limiting reports to UPDs resulted in a smaller scope for the study than the CUE studies, as shown in Figure 3.2. It also decreased time commitment for the study participants, resulting in increased study participation and statistical power. The usability reports collected in this phase were also used to examine the usefulness of the guidelines developed in Phase I for judging UPD content.

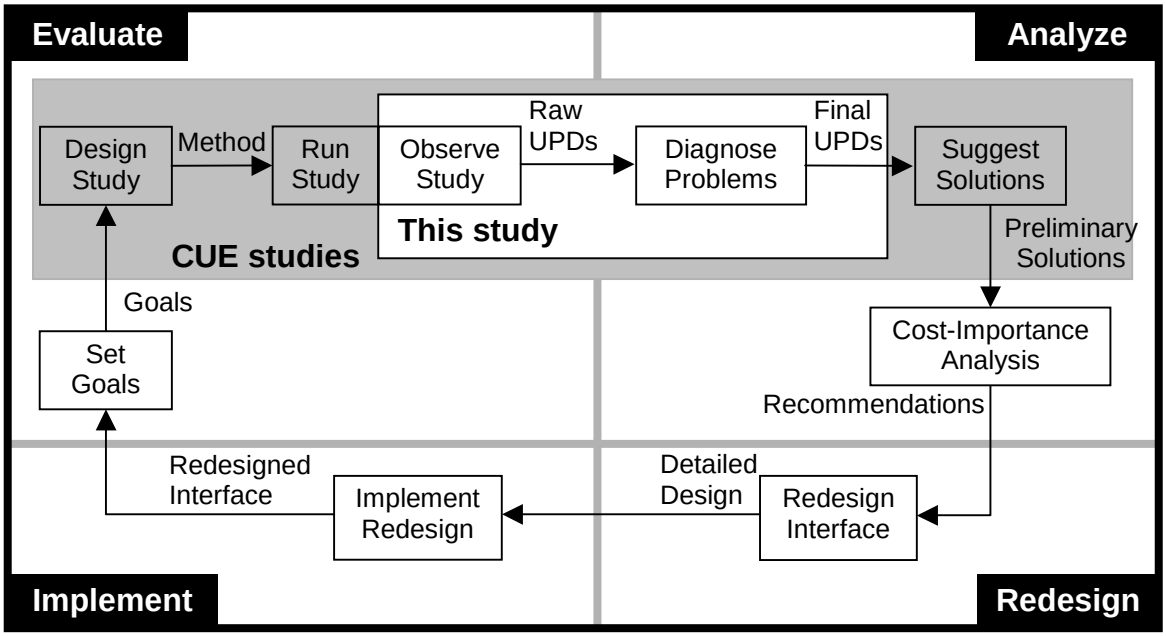


Figure 3.2 Study 4: Comparison of Scope to CUE Studies

3.1 Background

This section defines terms used in this document, summarizes previous studies in this research area, and explains the approach for this phase of the research.

3.1.1 Formative Usability Evaluation

Summative usability evaluations, as defined in ANSI NCITS 354-2001 *Common Industry Format for Usability Test Reports,* are used to measure how well a product meets its stated usability goals, and are generally conducted on completed products (ANSI, 2001). In contrast, *formative* evaluations are used to evaluate works-in-progress as part of the iterative design process (ANSI, 2001). The goal of a formative evaluation is

to identify user interaction problems so that they can be fixed in the next design iteration. Thus, the output of a formative evaluation is a list of usability problems (Hartson et al., 2003), as shown in Figure 3.3. Formative evaluations rely heavily on qualitative data, such as critical incident descriptions and interviews. In contrast, summative evaluations rely on quantitative metrics of effectiveness, efficiency, and satisfaction (ANSI, 2001; ISO, 1998).

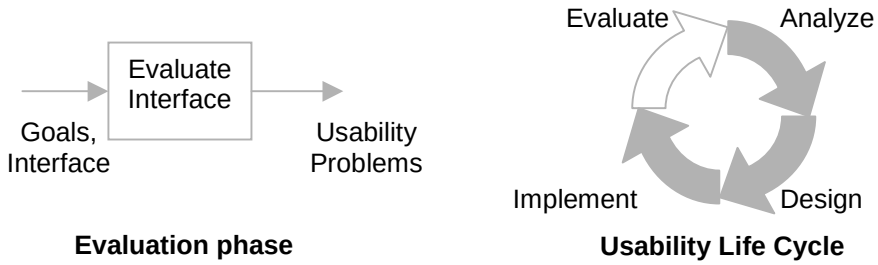


Figure 3.3 Formative Usability Evaluation and the Usability Life Cycle

There are many different formative usability evaluation methods (UEMs), and they fall into two groups: analytical and empirical. *Analytical* methods are used to predict usability problems that users may encounter. These include techniques that involve expert review such as Heuristic Evaluation (Nielsen, 1994b; Nielsen & Molich, 1990) and Cognitive Walkthrough (Lewis, Polson, Wharton, & Rieman, 1990; Wharton, Rieman, Lewis, & Polson, 1994). They also include automated techniques such as Bobby ("Bobby 5.0," 2005), a tool to identify universal access issues in websites. *Empirical* methods involve users and are generally referred to as usability testing. These are usually task-based evaluations conducted in a usability laboratory, and are commonly called think-aloud testing due to the use of verbal protocol during testing. Performance testing has also been used to describe usability testing without verbal protocol.

3.1.2 Previous Studies of Evaluator Effect in Usability Testing

There are many studies comparing UEMs; Appendix A provides a list of 57 such studies. Several UEM comparison studies observed an *evaluator effect* in analytical methods (Hertzum & Jacobsen, 2003). Different evaluators find different usability problems, even when using the same analytical UEM. Six studies have studied this effect

in analytical methods using at least 30 experts per UEM, as shown in Table 3.1, with any-two agreements ranging from 7-45%. The low reliability of expert review techniques can be leveraged by combining the results from several individual evaluators, resulting in higher thoroughness for the group as a whole (Hartson et al., 2003). For example, Dumas and Sorce (1995) found that problem identification by an individual expert conducting a review diminishes over time spent on the evaluation, and so it would be more efficient to have several experts conduct shorter evaluations than to have one expert conduct a long evaluation.

Usability testing has been called the "gold standard" (Nickerson & Landauer, 1997, p. 17) of usability evaluations for its thoroughness in finding usability problems and focus on the user's experience with the interface. It has been used as a benchmark or baseline to assess other formative evaluation methods, particularly analytical methods (Cuomo & Bowen, 1994; Desurvire, Lawrence, & Atwood, 1991; Desurvire, Kondziela, & Atwood, 1992; Law & Hvannberg, 2004b; Mankoff, Fait, & Tran, 2005; Nielsen, 1994b). Yet several authors have suggested that usability testing may not be appropriate to use as a yardstick for other techniques because it is not perfect, being subject to limitations of the laboratory environment, study design flaws, and evaluator biases (Hartson et al., 2003; Jacobsen et al., 1998; Newman, 1998).

Table 3.1 Large Studies Comparing Evaluators Performing Heuristic Evaluation

Study	Evaluation Method	#, type of evaluators		Reliability	Thorough-ness
Hornbæk & Frøkjær (2004b)	HE	43 1st-year CS undergraduates		.07	.03[a]
	MOT	44 1st-year CS undergraduates		.09	.03[a]
Molich & Nielsen (1990), Nielsen & Molich (1990), Nielsen (1992), Nielsen (1994b)	HE	77 readers of Computerworld	(Mantel)	.45[b]	.38
		37 CS students in UI course	(Teledata)	—	.51
		34 CS students in UI course	(Savings)	.26[b]	.26
		34 CS students in UI course	(Transport)	—	.20
		31 CS students, no training	(Banking)	—	.22
		19 usability specialists	(Banking)	.33[b]	.41
		14 usability & IVR "double experts"	(Banking)	—	.60

Note: This table only includes studies with at least 30 evaluators per UEM/interface. See Appendix A for a complete list of studies. MOT=Metaphors of human thinking. A dash (—) indicates that the measure could not be calculated from published data. Reliability is any-two agreement. [a]Calculated based on data published in the article. [b]As reported by Hertzum & Jacobsen (2003).

Table 3.2 Studies Comparing Evaluators Performing Usability Tests

Study	Same tasks?	#, type of evaluators	Reliability	Thorough-ness
Dumas, Molich & Jeffries (2004), Rourke (2003): CUE-4	No	9 professional teams	—	—
Kessner, Wood, Dillon & West (2001) same system as CUE-2	No	6 professional teams 6 professional teams (reanalyzed from CUE-2)	— —	.36[a] —
Molich et al. (1998): CUE-1[c]	No	3 professional teams	.06[a]	.36[a]
Molich et al. (2004): CUE-2 [c]	No	6 industry/university teams	.07[b]	.22[b]
Skov & Stage (2003; 2004) same system as CUE-2	No	36 teams 1st year students (in 2000) 27 teams 1st year students (in 2001) 8 professional teams (from CUE-2)	— — —	— — —
Jacobsen, Hertzum & John (1998)	Yes (video)	2 experienced evaluators 2 beginner evaluators	.42[b]	.52
Lesaigle & Biers (2000)	Yes (video)	5 usability professionals (x 3 different types of video; between-subject study)	—	—
Long, Styles, Andre & Malcolm (2005)	Yes (video)	12 usability students 12 usability students	— —	.33[a] .34[a]
Palacio, Bloch, & Righi (1993)	Yes (live)	4 temps with 3 mos. experience	—	—
Skov & Stage (2005)	Yes (video)	14 undergraduate students 2 usability experts	—	.27[a]
Vermeeren, Kesteren & Bekker (2003)	Yes (video)	2 unspecified individuals, Study 1 2 unspecified individuals, Study 2 2 unspecified individuals, Study 3	.59[a] .64 .75[a]	.78[a] .82[a] .87[a]

Note: Reliability is any-two agreement. A dash (—) indicates that the measure could not be calculated from published data. Excludes Wright and Monk (1991) with 13 teams of just-trained evaluators using the same tasks and observing a single (differing) user because the evaluators wrote individual reports and team results for problem detection were based on the aggregate results of the individual team members, rather than a true group effort. [a]Calculated based on data published in the article. [b]As reported by Hertzum & Jacobsen (2003). [c]Other teams participated, but only these teams used UT exclusively.

How appropriate is it to use usability testing as a benchmark for other UEMs? Is there an evaluator effect in usability testing, similar to the effect in analytical methods? A few studies have examined multiple evaluators performing usability testing on the same interface, summarized in Table 3.2. CUE-1, -2, and -4 (Dumas et al., 2004; Molich et al., 1998; Molich et al., 2004; Rourke, 2003) and studies that re-analyzed data from CUE-2 (Kessner et al., 2001; Skov & Stage, 2003, 2004) gave the usability testing teams the freedom to use their own facilities and testing methods, and to choose their own tasks and participants. It is impossible to separate the user effect (differences in problem detection due to the number and type of user participants; Law & Hvannberg, 2004a; Nielsen, 1994a; Spool & Schroeder, 2001; Virzi, 1990, 1992) from the evaluator effect in these studies, and to separate the evaluator effect due to study design from the evaluator effect due to study observation and analysis. Table 3.2 table excludes a study by

Study 4 differed from the CUE studies in the use of use of pre-recorded usability sessions. While this decreased the realism of the test setting, or external validity, it helped focus the scope of this research (see Figure 3.2 Study 4: Comparison of Scope to CUE Studies) and increased internal validity. It had the added benefit of greatly reducing time commitment for the evaluators. Teams participating in the CUE-2 study spent 16-218 person-hours on the study, and 11 teams that had expressed interest in participating changed their minds after hearing about the time commitment required (Molich et al., 2004). Simultaneous live viewing of usability sessions (Palacio et al., 1993) has a similar effect to pre-recorded sessions, but is limited in the number of evaluators that can participate and requires coordination among all evaluators.

Four previous studies used recordings or simultaneous live viewing to control for effects due to user and task. However, one of these studies did not report individual evaluator reliability (Lesaigle & Biers, 2000), and three used four or fewer evaluators (Jacobsen et al., 1998; Palacio et al., 1993; Vermeeren et al., 2003). Two additional studies using pre-recorded sessions were published after Study 4 was completed. Skov and Stage (2005) studied 14 just-trained undergraduate students, but did not report reliability (thoroughness was 27%). Long, Styles, Andre & Malcolm (2005) examined different formats for session recordings, using Cohen's κ (Cohen, 1960) to measure

reliability in problem detection among evaluators. The study used 24 usability students watching the same video as Study 4 (half with the addition of the users' faces in the corner of the video). They found evaluator reliability to be low, with $\kappa = .25$ and .36 for their two experimental groups (thoroughness was 33-34%).

While Long et al. (2005) provides an interesting comparison of recording formats, it is difficult to draw conclusions about evaluator reliability from it because of the use of student evaluators. The same could be said of Skov and Stage (2005), had they reported reliability for their just-trained students. Studies have shown that experts are better than students at finding usability problems in Cognitive Walkthroughs (Desurvire et al., 1992), Heuristic Evaluation (Connell & Hammond, 1999; Desurvire et al., 1992; Nielsen, 1992), PAVE (Desurvire & Thomas, 1993) and usability testing (Skov & Stage, 2004). Relying on just-trained participants can be useful when testing techniques designed specifically for novice evaluators, or to control for effects due to experience with techniques, but is not representative of usability practice by experienced professionals. Usability evaluation has been called an art, although whether evaluation is an art or science is still debated within the usability community (Spool & Schaffer, 2005), and user interface design has been called a craft, requiring skill and experience to apply the basic tools of the field to create and test interfaces (Waloszek, 2003). This is not to say that students always conduct poor evaluations – many usability professionals attending a CHI2000 tutorial were unable to identify the two student reports collected in the CUE-2 study (Molich et al., 2004). However, care should be taken in using studies of usability students to draw conclusions about practitioner performance. Appendix A lists 57 UEM comparison and evaluator effect studies, including the type of evaluators used in each study.

Several studies have compared novice, expert and double-expert (both domain and usability experts) performance in heuristic evaluation (Baker, Greenberg, & Gutwin, 2002; Connell & Hammond, 1999; Desurvire, 1994; Desurvire & Thomas, 1993; Desurvire et al., 1992; Nielsen, 1992) and cognitive walkthroughs (Mack & Montaniz, 1994) with mixed results about whether experts find more problems than novices do. Jacobsen, Hertzum and John (1998) compared two novice and two experienced evaluators watching identical session tapes and found that the experts spent more time

watching the tapes and found more problems (Hertzum & Jacobsen, 2003), but point out that the sample size was four evaluators. Skov and Stage (2004) compared 36 first-year students with eight professional teams reanalyzed from CUE-2 and found that the students reported fewer problems than the professionals did, but evaluators in this study picked their own tasks and participants. It is likely that experienced evaluators find more problems than novices do, but more studies are needed to confirm this effect.

Usability testing is highly regarded as a UEM, but there are few comparison studies of how thorough and reliable it is. We need a better understanding of the evaluator effect in usability testing with studies that involve a greater number of experienced evaluators, exert more experimental control over the study setting (increasing internal validity), and involve more evaluators to compensate for individual variability. The Phase II study of this research was specifically designed to examine the evaluator effect in usability testing. It included both experience practitioners and graduate students to examine the range of outputs generated by evaluators with a range of usability experience.

3.1.3 Assessing Problem Description in Addition to Problem Detection

UEM comparison studies and studies of the evaluator effect in usability testing have focused on usability problem detection, severity judgments, and number of users to include in a usability test. Problem detection is an important part of a usability evaluation, but the usability test is not over when the last participant goes home. Communicating the results of the test (either through a written report or verbally) is an essential part of the usability testing process (Rubin, 1994). Evaluators need to understand the problems found, record them, and communicate them to the team that will redesign the interface.

Anecdotal evidence suggests that many evaluators create ineffective documentation to convey the results of formative usability evaluations. Andre, Hartson, Belz and McCreary (2001) reviewed hundreds of written usability problem descriptions collected by Keenan, Hartson, Kafura, & Schulman (1999) from professional usability laboratories, and describe many of them as "*ad hoc* laundry lists" (p. 108) that would

require significant verbal communication to supplement the information provided in the written descriptions. Jeffries (1994) reviewed the problems collected for Jeffries, Miller, Wharton and Uyeda (1991), and found the issues listed in Figure 3.4, although many of these issues are with the overall report or the quality of the suggested solutions, rather than the problem descriptions themselves. Dumas, Molich and Jeffries (2004) reviewed 17 usability reports submitted to the fourth of their Comparative Usability Evaluation series (CUE-4), and suggested four areas of improvement for usability reports, also listed in Figure 3.4. Andre et al. express the opinion that poor documentation and communication of usability problems found diminishes the effectiveness of a usability evaluation and can reduce the return on the effort invested in conducting the evaluation. Dumas, Molich and Jeffries suggest that communication style and attitude in report writing can affect recipients' acceptance of suggestions and the number of problems recipients choose to fix. Jeffries suggests that developers may interpret poorly described problems as false alarms, causing the developers to ignore the poorly described problem and increasing the likelihood that developers will treat future problems as opinion or false alarms.

Criticisms by Jeffries (1994)

- Describes the solution without describing the problem being solved. Solution lacks justification, and problem may go unsolved if this particular solution is rejected.
- Fails to consider trade-offs in design decisions, optimizing for one task, user group, or design guideline at the expense of others.
- Describes small examples of problems without pointing out the larger trend across multiple problems, leading to a solution that is too narrow.
- Evaluator misunderstands interface, resulting in an incorrect description of why a problem occurred and how to fix it.
- Multiple evaluators provide conflicting comments or recommendations.
- Evaluator allows personal biases for interaction styles to overly influence recommendations.

Suggestions by Dumas, Molich and Jeffries (2004, p. 24)

- Emphasize the positive
- Express your annoyance tactfully
- Avoid usability jargon
- Be as specific as you can

Figure 3.4 Published Criticisms of Usability Problem Descriptions

A recent study by Skov and Stage (2004) to determine the effectiveness of basic usability engineering training compared eight usability reports from CUE-2 to five usability reports written by student evaluators evaluating the same interface. The practitioners received higher scores on 9 of 17 variables used to rate the reports: test procedure conduction, test description, clarity of the problem list, executive summary, clarity of report, number of problems, practical relevance of the results, and conclusion. The variables used to grade the report were developed for the study. They concluded that the students were able to effectively select tasks and describe problems found, but that students were less effective at finding problems and describing them in a way that would be relevant to software developers. Skov and Stage's study suggests that experience plays a role in finding problems and writing problem descriptions, which deserves to be examined in a study with a larger sample size.

Several problem classification schemes and structured reporting formats have been proposed (e.g. Andre et al., 2001; Cockton et al., 2004; Hvannberg & Law, 2003; Lavery et al., 1997; Sutcliffe, Ryan, Doubleday, & Springett, 2000) to improve the quality or utility of usability problem descriptions without providing measures of these values. A possible reason for this lack of measures is a lack of standards or guidelines for describing usability problems. Existing national and international usability standards focus on the summative evaluation measures of effectiveness, efficiency and satisfaction (ANSI, 2001; ISO, 1999a) or usability design process (ISO, 1999b). Workshops on reporting the results of formative usability tests were organized by NIST in October 2004 and UPA in July 2005, but the workshop reports (Quesenbery, 2005; Theofanos et al., 2005) focus on the formative usability test report as a whole more than individual usability problem descriptions, and the standard is still under development. Books on usability testing by usability practitioners and researchers (Dumas & Redish, 1993; Hix & Hartson, 1993; Mayhew, 1999; Nielsen, 1993; Preece, 1994; Rubin, 1994; Stone et al., 2005) provide guidance on observations to make during a usability testing session and the major sections to include in a usability report, but few specific recommendations about how to describe individual usability problems. Observations during a study are close in form to raw UPDs, but final UPDs reflect further diagnosis of the problems and are more

polished. Figure 3.2 Study 4: Comparison of Scope to CUE Studies illustrates the relationship between raw/final UPDs and the formative evaluation cycle.

Without a way to evaluate the content of UPDs, any measurement or comparison of evaluation effectiveness is incomplete. Measuring evaluation effectiveness is important for comparing UEMs, testing individual evaluators or usability laboratories, and developing training for usability evaluators. Phase II of this research explored the suitability of the UPD guidelines in Phase I for evaluating usability reports, providing a basis for future studies to formally develop metrics of usability problem description quality.

3.1.4 The Evaluator Effect in Severity Judgments

Hertzum and Jacobsen (2003) describe a second type of evaluator effect, that of differences in severity judgments. They provide measurements of agreement in severity ratings calculated from data for three studies (Hertzum & Jacobsen, 1999; Jacobsen et al., 1998; Nielsen, 1994b), and report any-two agreements ranging from .20-.28, and Spearman correlations from .23-.31. Problem severity is one factor considered in a cost-importance analysis (Hix & Hartson, 1993), and so different severity judgments could result in selection of a different set of problems to fix. Phase II of this research explored reliability in severity judgments by comparing severity ratings for sets of evaluators that identified the same problem.

3.1.5 Further Design Considerations in the Phase II Study

A design consideration for this research was the type of session recording to use, and specifically whether or not to include the face of the users in the usability test in the session recording. Including the face provides more information about users' emotions and opinions than just think-aloud comments, but also obscures a portion of the computer screen, does not preserve anonymity of the users, and requires evaluators to divide attention between three channels of information: think aloud (auditory), screen capture (visual), and user's face (visual). The study by Long et al. (2005) compared two different versions of a digital usability session movie made with Morae. One version had screen

capture and think aloud, the other also had a video of the user's face in the bottom-right corner. Long et al. found that there was not a significant difference in the number of problems identified by each group, indicating that omitting the users' faces does not impair problem detection. That study was published after the current study was completed, but preliminary results were made available to the author while designing the current study. Lesaigle and Biers (2000) reported similar results for usability professionals watching videotapes, finding no differences in problem detection between groups. Lesaigle and Biers found that severity ratings were higher for the group with the face in the video, but this should not be an issue if all evaluators in the current study watch the same video. The video used in the current study was a modified version of the no-face video studied by Long et al.

Another consideration was the number of users to include in the session video. Previous research indicates that 4-5 users in usability testing are enough to find the most serious problems (Nielsen & Landauer, 1993; Virzi, 1990). Research that is more recent has indicated that these results may be true for extremely simple tasks and systems, but more users are needed for complex systems and diverse user populations (Spool & Schroeder, 2001), or to ensure finding a high percentage of problems in the system (Faulkner, 2003; Woolrych & Cockton, 2001). Hundreds of users may be necessary to find severe but rare problems in systems with highly diverse user populations, such as elections (Bailey, 2000, November). The website tested in this study is very large, although the study focused on a single task of moderate complexity. The movie included four users, which is unlikely to be enough to find all usability problems affecting the task. However, the goal of the current study was not for all evaluators to find all of the problems with the target interface, but rather to compare the number problems they identified based on identical session recordings. This does limit external validity and applicability of the results, since most usability studies do try to find most of the problems present in an interface. The small scope of the study is consistent with the point of view expressed by Andrew Monk in his "Experiments Are for Small Questions" section of the "Commentary on Damaged Merchandise" article (Monk, 1998), that formal experiments can be useful for answering very specific questions. The goal of this

particular study was to use a more controlled setting than the CUE study in order to focus on the observation and description portion of the usability evaluation process and increase internal validity.

Some UEM comparison studies that compare reliability in expert reviews with reliability in usability testing match number of expert evaluators with number of usability testing participants (Beer, Anodenko, & Sears, 1997; Fu, Salvendy, & Turley, 1998; Karat, Campbell, & Fiegel, 1992). This may be appropriate if the experimenter is specifically comparing experts to users. However, if the experimenter is comparing usability testing to other UEMs, number of expert evaluators cannot be equated with number of users in a usability test. A better choice to ensure statistical conclusion validity is to match number of evaluators performing expert reviews with number of evaluators conducting usability tests when determining sample size (Gray & Salzman, 1998). Hertzum and Jacobsen (2003) examined the contributions of both users and evaluators, and suggest that the number of problems found will approximate C x SQRT(number of users x number of evaluators). The current study held constant the number of users and examined the contributions of individual evaluators.

Several authors have suggested the use of structured reporting forms or classification schemes (Andre et al., 2001; Cockton et al., 2004; Hvannberg & Law, 2003; Lavery et al., 1997; Sutcliffe, 2000) to increase the thoroughness and utility of usability problem descriptions, or to facilitate matching usability problems found through the study. However, using such reporting and diagnostic tools changes the observation and analysis processes. They have the potential to affect not only the way problems are described, but also which problems are identified and noted. They also require evaluator training and can greatly increase the time spent on an evaluation. An open reporting format, similar to that used in the CUE studies, was chosen to avoid interference with the evaluator's typical observation and reporting practices. This increased the complexity of the process of matching UPDs to problems in the master list, but was compensated for by having three judges perform the matching process independently and using consensus among at least two out of three judges to determine the final list of usability problems detected by each evaluator.

The CUE studies gathered more than just usability problem descriptions in the usability reports. They specifically asked for an executive summary, and three lists of important items: three most important positive findings, three most important usability problems, and three most important recommendations. Many teams included other details in the body and appendices of their reports, such as usability metrics (e.g., task success/failure, completion times, questionnaire responses). However, creation of such items is time-consuming and only peripherally related to the area of interest in this study. One pilot participant in this study described spending half of the study time writing the front matter (executive summary and top-three lists). These items were excluded from the current study to maintain the focus on usability problem descriptions and reduce time commitment for study participants.

The eventual goal of a formative evaluation is to improve the user interface, and several authors have suggested that UEM effectiveness should be assessed by downstream measures of increased usability in the redesigned interface. However, evaluating the usability of an interface and its redesign is complicated and can be costly, and there have been few studies of downstream usability; Bailey (1993), John and Marks (1997), and Law (2004) are three examples of this type of study. Further, downstream measures of increased usability are dependent upon not only the quality of the initial evaluation but also the quality of the redesign, and the usability retest is subject to the same reliability issues as other usability evaluations. Restricting the output measurements to the list of usability problems maintains the small scope desired for this study.

3.1.6 Comparisons to Medical Diagnosis

The existence of an evaluator effect in studies using pre-recorded usability sessions suggests differences in evaluations due to the individual evaluator. Biases in diagnostic decision-making have been studied extensively in medical diagnosis, which is similar to usability diagnosis. This section describes medical decision-making and draws an analogy between medical and usability diagnosis. Medical diagnosis is an ill-defined problem setting, with an imprecise or unspecified starting state, means available for solving the problem, and goal state (Gilhooly, 1990). As in other naturalistic decision-

making settings, the act of diagnosing or describing the problem is often more difficult than solving the problem or recommending a course of treatment (Durso & Gronlund, 1999).

A common model of medical decision-making is *hypothetico-deductive* reasoning. The model was first published by Elstein, Shulman and Sprafka (1978), with Elstein and Bordage (1979) refining the model and coining the name "hypothetico-deductive." The essence of the model is that diagnosticians form hypotheses based on review of early data, use these hypotheses to predict further findings that would be present if the hypotheses were true, and use these predictions to guide their acquisition of additional data, deducing the accuracy of each hypothesis (Elstein, 1994). The model includes four stages: data collection/cue acquisition, hypothesis generation, cue interpretation, and hypothesis evaluation. Hypothesis generation helps translate an unstructured, unmanageable problem into a manageable one by creating a small number of concrete alternatives that the diagnostician can systematically test; otherwise, the diagnostic process could overwhelm working memory (Elstein, 1994). An experienced practitioner creates a structured problem space by quickly assessing relevant and extraneous information, and then matching the current situation with internal schema built from previous domain experience (Patel, Kaufman, & Magder, 1996).

During diagnosis, the medical practitioner uses hypotheses to guide data collection from the patient, from tests, from referring or consulting practitioners, and from her own observations. Practitioners need to know what questions to ask, what signs and symptoms distinguish between hypothesized diagnoses, what tests to order, and so on. They refine their conception of the problem as they gather more information, develop hypotheses, and identify limiting factors on their solution (Kurfiss, 1988). Medical practitioners need to be careful when listening to patient descriptions of symptoms because such descriptions may not always be accurate, or the practitioner may misunderstand what the patient is trying to describe (Griffin, Schwartz, & Sofronoff, 1998). Similarly, psychoanalysts need to be careful to distinguish between facts, beliefs and opinion, and need to be able to judge the quality and applicability of published research (Gambrill, 1990). Medical practitioners need to know when to reject a

hypothesis based on conflicting evidence, and when to accept a hypothesis despite conflicting evidence. Knowing how to cope with anomalies and conflicting evidence is essential in medical reasoning (Patel et al., 1996).

Usability diagnosis is similar to medical diagnosis. The usability practitioner identifies potential usability problems and speculates about possible causes for the problems. She seeks additional information by examining the interface, observing users, consulting guidelines, and comparing to experiences on previous projects. She then designs a solution based on the conjectured problem cause. A usability evaluation typically generates large amounts of data in the form of user comments, expert observations, automatically recorded information, screen capture, etc. The usability practitioner has to decide which sources to consult, and locate information that is relevant to the current problem among the available data. The practitioner also has to resolve conflicting evidence from different guidelines and different users.

The need to reconcile or merge evidence from multiple users is a point where the usability/medical analogy breaks down, at least in terms of a traditional one-doctor/one-patient relationship. In usability, the "patient" is the interface, and the practitioner generally consults multiple users to gain a more complete understanding of the interface. A closer analogy might be an epidemiologist studying an outbreak of a contagious disease. The epidemiologist studies the environment and the many people in it to understand how the disease is transmitted, why some people catch the disease but others do not, and why different people show different symptoms for the same disease. The usability practitioner studies the interface and the users interacting with it to understand why some people experience usability problems but others do not, and why different users have different reactions to the same interaction flaw.

Examination of sources of errors in medical decision-making suggests that medical practitioners are subject to many cognitive biases that can affect the diagnostic outcome. Croskerry lists 32 *cognitive dispositions to respond*, which are biases in interpreting information and making decisions (Croskerry, 2002, 2003). Many of these are based on cognitive heuristics. While heuristics can help the diagnostician make

decisions quickly, they can also impair judgment, leading to diagnostic errors. For example, in psychoanalysis, the diagnostician needs to be careful of prejudgments and biases such as race and gender bias, labeling and stereotypes, availability heuristic, and confirmatory hypothesis testing (Garb, 1998). Usability practitioners can also experience cognitive dispositions to respond. They can affect the problems the practitioner notices, the causes the practitioner ascribes to problems, and the solutions the practitioner suggests. Based on the similarities between medical diagnosis and usability diagnosis, it is expected that there will be an evaluator effect, even when watching pre-recorded usability sessions.

3.2 Research Questions and Hypothesis

This phase examined each of the three research questions and tested the following research hypotheses.

Research Question 1: How do usability practitioners describe usability problems?

This was addressed in an exploratory fashion in Phase I, which gathered practitioners' opinions about how usability problems should be described to develop 10 guidelines and rate four aspects of each guideline: *difficult*, *required*, *relevant*, and *helpful*. Study 3 found no differences in opinion due to level of experience (either number of evaluations or years of experience). There may have been a floor effect, since all of the respondents had some practical usability experience (five years or 10 evaluations). Study 4 focused on evaluator behavior in describing usability problems, and considered the following issues.

- Did evaluators follow the guidelines rated as *required*?
- What information did evaluators include with the descriptions: solutions, images or screen shots, comments about facilitator interventions?
- Does evaluator behavior in describing problems match evaluator opinion about what should be described?

Hypothesis 1a: Practitioners in Study 3 vs. Study 4. The practitioners in Study 4 will have the same opinions about the 10 guidelines as the practitioners in Study 3.

Hypothesis 1b: Practitioners in Study 3 vs. Students in Study 4. The practitioners in Study 3 will have the same opinions about the 10 guidelines as the students in Study 4. There were no significant differences due to experience in Study 3, and so none are expected here.

Hypothesis 1c: Following the Guidelines. It is expected that the Study 4 practitioners will follow the five guidelines rated most *required* by the Study 3 practitioners, and that *Describe a Solution* will show a split in behavior similar to the split in opinion found in Study 3.

Hypothesis 1d: Opinion vs. Behavior. Practitioner reporting behaviors will be consistent with their opinions, and an individual evaluator will follow most closely the guidelines that the evaluator gave the highest ratings for *required*. If there are differences between opinion and behavior, it will most likely occur with the guidelines that are rated as *difficult*.

Research Question 2: Is there an evaluator effect in usability testing?

The CUE studies (Dumas et al., 2004; Molich et al., 1998; Molich et al., 2004; Molich et al., 1999; Rourke, 2003) and studies that combined CUE data with further data collection (Kessner et al., 2001; Skov & Stage, 2003, 2004) have found an evaluator effect in problem detection in usability testing, but their realistic design (higher external validity) makes it impossible to separate effects due to differences in study design, task selection, participant selection, and observation and analysis (lower internal validity). Four previous studies found an evaluator effect with evaluators observing identical usability testing sessions, but it is difficult to draw large-scale conclusions from these studies due to small number of evaluators (Jacobsen et al., 1998; Palacio et al., 1993; Vermeeren et al., 2003) or lack of data about individual evaluator performance (Lesaigle & Biers, 2000). Several studies have also found an evaluator effect in severity judgments (Hertzum & Jacobsen, 2003). Hertzum and Jacobsen suggest that usability testing, like

heuristic evaluation, may benefit from multiple evaluators, but qualify their recommendations because they are based on a study with four evaluators.

Hypothesis 2a: Problem Discovery. Thoroughness and reliability of problem detection will be similar to heuristic evaluation, with multiple evaluators required to find the most severe usability problems.

Hypothesis 2b: Problem Severity. Practitioners will differ in their severity ratings of usability problems.

Research Question 3: How Can We Assess the Content of UPDs?

This study included a preliminary exploration of the usefulness of the guidelines for grading usability reports in terms of the content of usability problem descriptions. If the guidelines can be used to grade reports, future studies of the guidelines could refine them and study their use in grading reports more formally. Other aspects of the reports were noted, such as numbers of words used to describe the problems, and the inclusion of screen shots and solutions. Assessing the way usability problem descriptions are written would complement existing measures of evaluation effectiveness such as thoroughness, validity and reliability.

Hypothesis 3a: Good evaluators follow the guidelines. Following the guidelines will be associated with the following indicators of a "good" usability report: thoroughness of the problem set, validity of the problem set, number of errors in the report, hours spent on the evaluation, and author (practitioner vs. student).

Hypothesis 3b: Rating reliability. The ratings of the judges will be reliable, in terms of association (considering the same underlying traits in assigning a rating) and bias (tendency to give overall higher or lower ratings).

3.3 Method

Evaluators watched the same recording of a usability evaluation session. Each evaluator wrote a report summarizing their comments on the task being evaluated. The

reports were then graded by three judges. The study was a study of usability studies, and so the participants in this study were the *evaluators*, or the people that wrote usability reports, unlike a traditional usability study where the participants are the *end-users* of the software being tested. The study also differed from a traditional usability evaluation in that the evaluators were distinct from the *facilitator* or *moderator* (the person sitting in the room with the end-users during the testing session).

3.3.1 Participants

The usability evaluations analyzed in this study were conducted by 44 participants, 21 *practitioner evaluators*, and 23 *student evaluators.* All evaluators were required to be fluent in English. Evaluators either indicated that English was their first language, or answered “strongly agree” or “agree” on a 6-point Likert-type scale (Likert, 1932) to the following question: “I speak English as well as someone that only speaks English.” Thirty-eight participants were native English speakers, four answered, “strongly agree,” and two answered, “agree." Thirty-six participants were from the United States of America, and eight participants were from India (4), the United Kingdom (2), Canada, and Denmark. One practitioner that experienced technical problems submitting the final report was dropped from the study.

The practitioner evaluators were recruited by sending email to several HCI and usability mailing lists, local chapters of the Usability Professionals’ Association (UPA), professional contacts, Virginia Tech graduate students or in HCI, and by word of mouth. Thirteen practitioners were from industry, two from university, two from government, and one from both government and industry. Practitioners were required to meet at least one of the following two requirements: five years of usability experience, or have 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). Practitioners were not informed of the experience requirements at the time of recruitment, and so had little incentive to misrepresent their level of experience. One practitioner who did not meet the experience requirements was dropped from the study.

Student evaluators were recruited from mailing lists of graduate students in HCI at both Virginia Tech and Georgia Tech, lists of students that had taken Usability Engineering at Virginia Tech, and through personal contacts. Students were required to have completed a graduate-level course in usability engineering. Twenty-one students had completed usability engineering at Virginia Tech and two at 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).

Before the start of the study, usability reports were collected from seven pilot participants: five graduate students and two practitioners. Seven additional pilot reports collected from usability practitioners and graduate students for Long et al. (2005) were provided by an external usability researcher

Participant Materials

Participants either were mailed a packet or downloaded a packet consisting of three items: a digital movie of a usability session, the Morae movie player for PCs, and a usability report template with the study instructions. Participants receiving the study packet through the US Postal Service received a printout of the instructions and a CD with the instructions, movie and movie player. Participants who downloaded the CD contents were given strict instructions to print out the instruction document so that all participants would have the same printed instructions in front of them while working on the study. Participant documents are included in Appendix E.

3.3.1.1 Usability Session Recording

The movie consisted of four participants using the Internet Movie Database (imdb.com) and performing the task shown in Figure 3.5. This task has been used by Rex Hartson as an example for his Usability Engineering class at Virginia Tech and for a usability training video. A movie website was chosen because most people have some domain knowledge of this area. The movie was created by Terence Andre for a different comparative study (Long et al., 2005). A movie from Andre's study was chosen because it had been viewed by many HCI students and professionals and was known to contain

numerous usability problems, despite its short length. Andre made the session recording using Morae. The recording captured the entire computer desktop, mouse and cursor movement, keyboard-typing and mouse-click noises, user think-aloud comments, and comments by the study facilitator. The movie did not include video of the users' faces. The digital movie was made available in both .avi (for PCs) and QuickTime .mov (for Macs and people unable to view the .avi file). The .avi file did not use compression; the .mov file was compressed slightly, affecting colors in full-color ads and photos, but not text quality.

Task: Name all the movies that both Owen Wilson and Luke Wilson (the actor from Old School) have appeared in together.

Answer: The Wendell Baker Story, Rushmore, The Royal Tenenbaums, Bottle Rocket, and Around the World in 80 Days

Figure 3.5 Study 4: User Task for Usability Session Recording

The initial movie from Long et al. (2005) was approximately 12.5 minutes long. Three of the participants took roughly two minutes each to complete the task. The other participant took over six minutes to complete the task, with over five minutes spent searching for the feature tested in the task (a search box located at the very bottom of the actors' homepages), and then roughly one minute using the feature. After watching the participant search for the feature for several minutes, it is clear that the participant is having a difficult time finding the feature. Approximately 1.5 minutes were cut from this section of the movie because they did not add information about the design of the task. The cut was made in such a way that evaluators should not have noticed it. The final movie was approximately 11 minutes long. Figure 3.6 – Figure 3.9 show sample screen shots from the movie. Please contact the author for a copy of the movie.

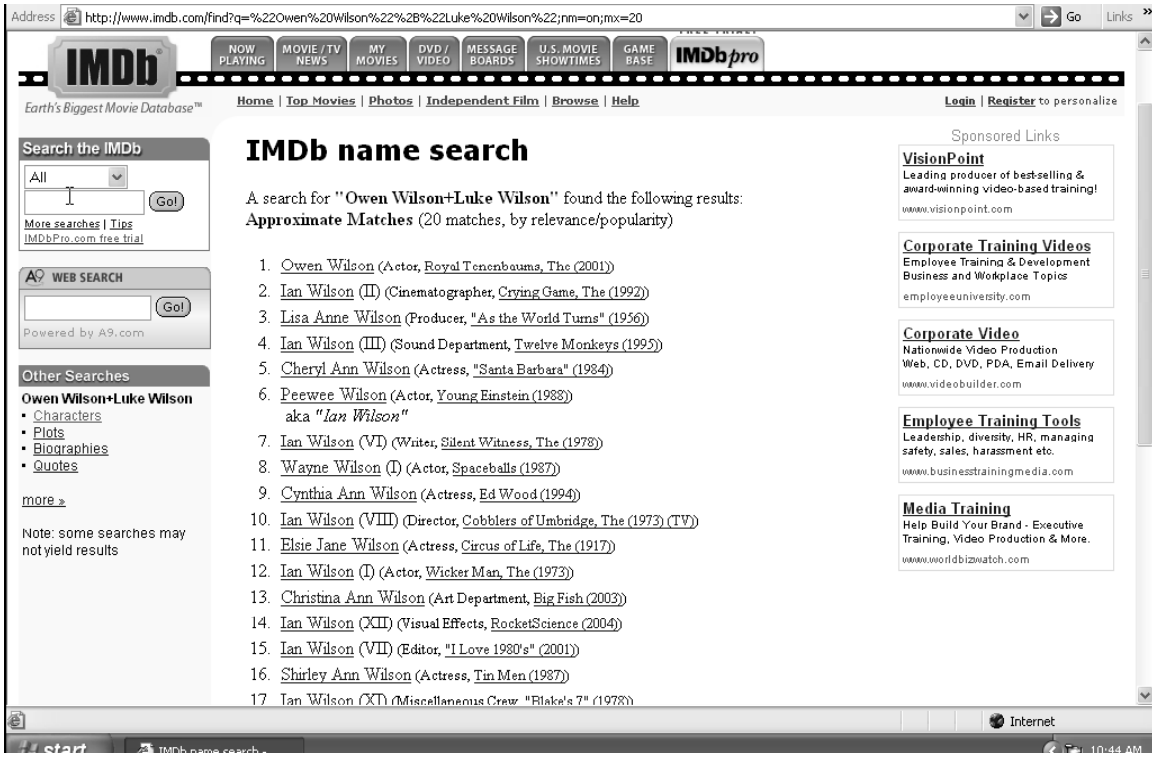


Figure 3.6 Study 4: IMDb Screen Shots, Search Results for Top-Left Box

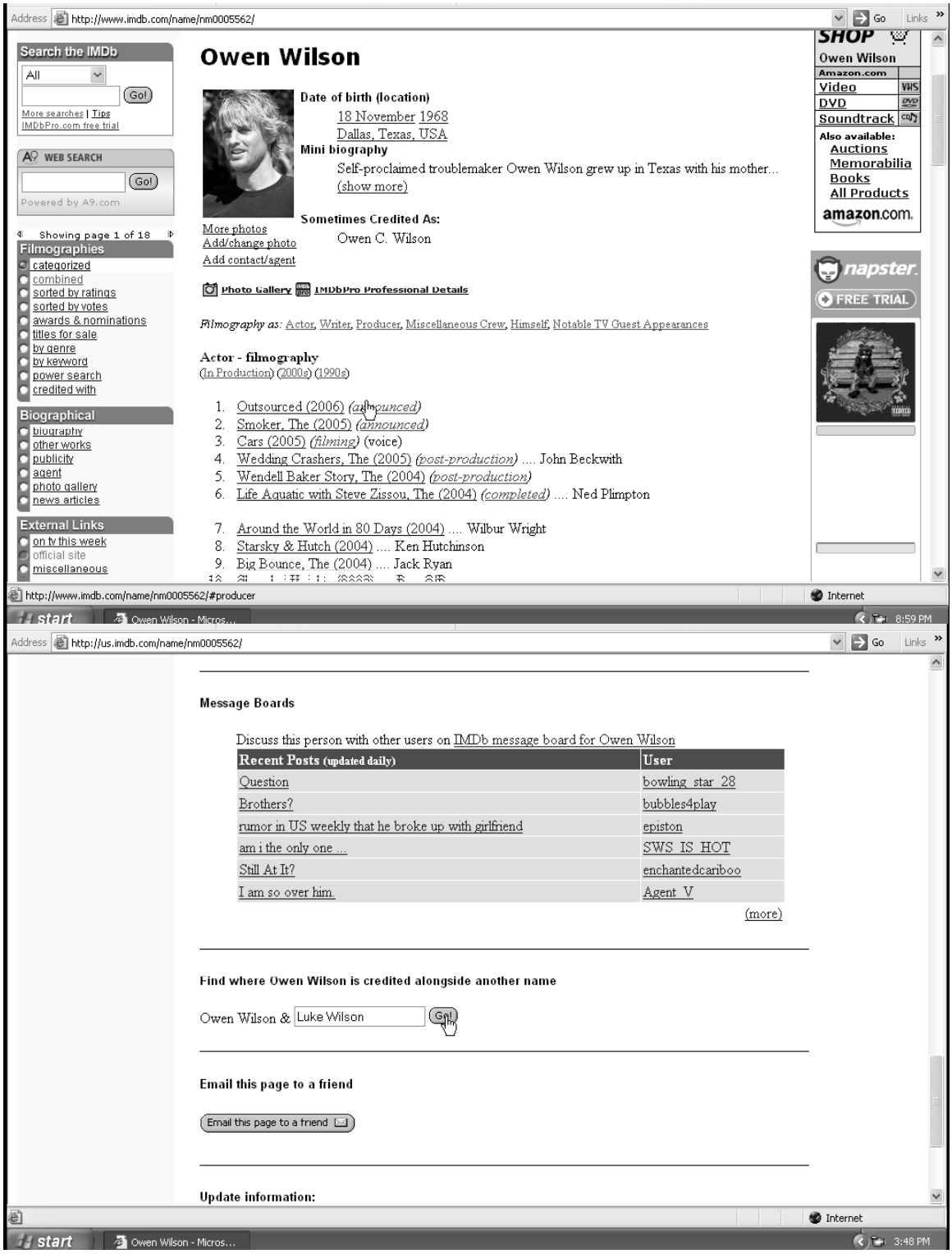


Figure 3.7 Study 4: IMDb Screen Shots, Owen Wilson Page

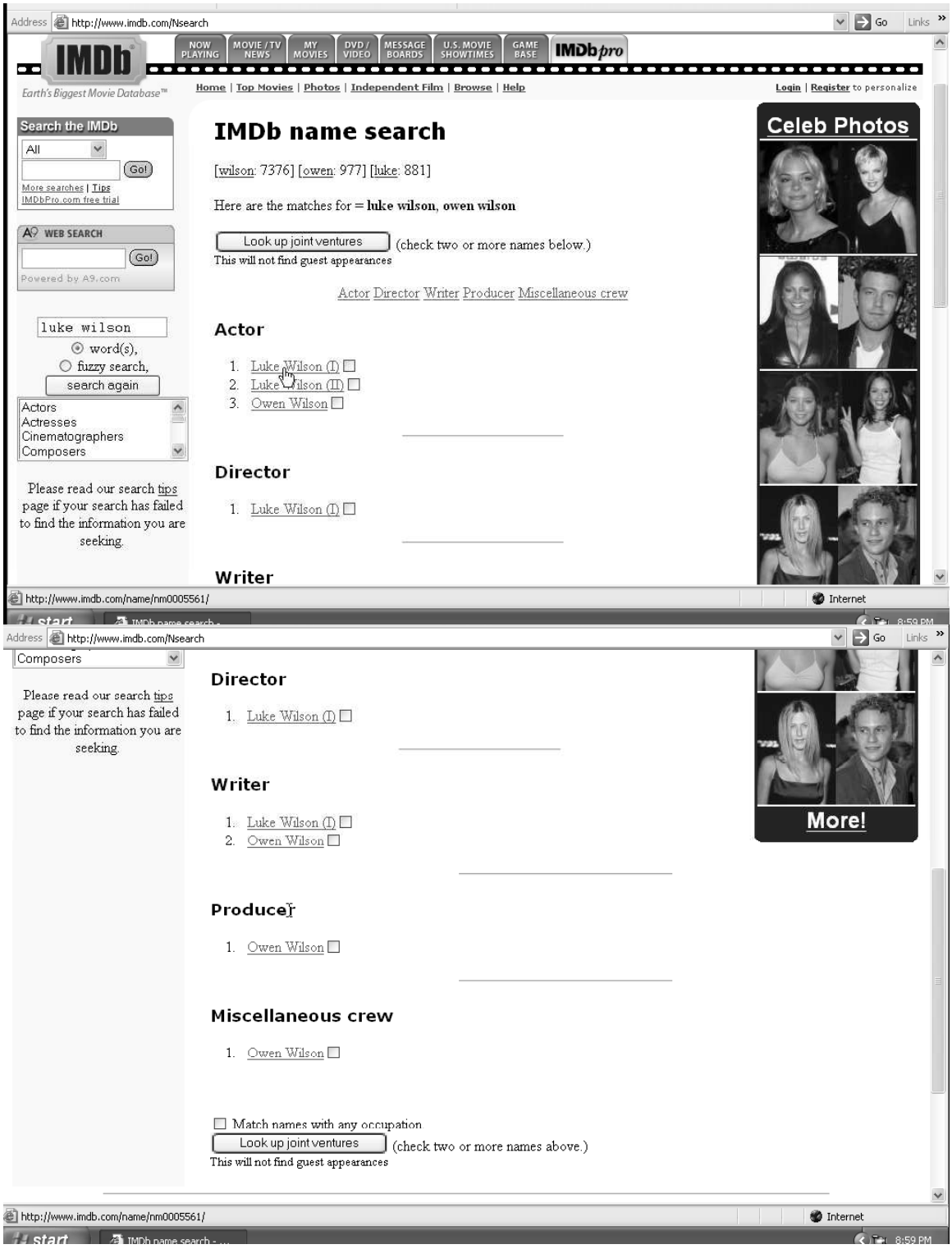


Figure 3.8 Study 4: IMDb Screen Shots, Joint Ventures Search Form

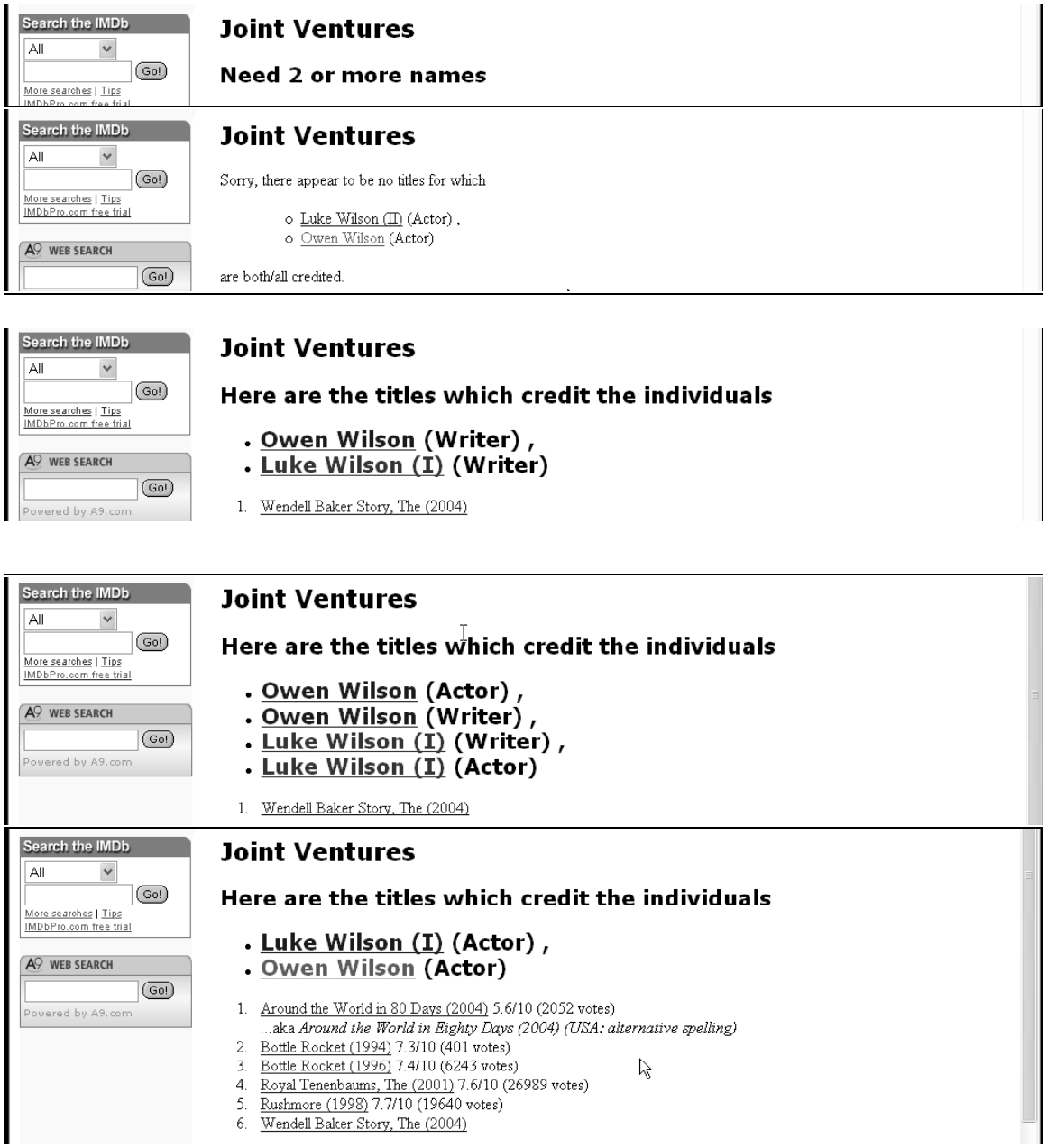


Figure 3.9 Study 4: IMDb Screen Shots, Final Results Page and Error Messages

3.3.1.2 Usability Report Template

Participants wrote comments about the usability session in the usability report template. The report template consisted of three pages.

1. Study overview, instructions for downloading and playing the movie.
2. Evaluation scenario, the task the users performed, and guidelines for watching the movie and writing the report.
3. Comment categorization scheme and two blank comment forms.

The document was made available in both MS Word and RTF formats. Figure 3.10 shows the blank template for each comment. The full text of the instructions and usability report template are included in Appendix E.

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 3.10 Study 4: Usability Comment Template with Comment Category Codes

The initial report template was modeled on the report template used in CUE-4 (Molich, 2004). Use of a session movie resulted in identical methods and test scripts for all evaluators, so these sections were eliminated. Four additional changes were made, as summarized in Table 3.3. Table 3.4 shows the comment categorization scheme used in this study, including the both original single-letter comment codes used in the CUE-4 study and the two-letter comment codes used in the current study.

Table 3.3 Study 4: Report Template Changes Made After Pilot Testing

Change	Reason(s)
The Executive Summary section (three most important positive findings, areas for improvement and recommendations) was eliminated.	1. One practitioner pilot participant spent half of the study time writing and re-writing this section, stating that it is the only section clients typically read. Nonetheless, it was not essential to the current study. 2. This reduced the scope of the experiment to concentrate more specifically on observation and diagnosis (see Figure 3.2 Study 4: Comparison of Scope to CUE Studies).
A specific field for a brief title or description of each comment was removed, leaving just the body of the description.	1. The CUE-4 analysis team found that these titles were often terse, jargon-filled, and difficult to understand (J.S. Dumas, personal communication, May 18, 2005). 2. Use of the word "brief" might imply to the evaluator that only a brief description was wanted, even in the longer description section. 3. Judges during the analysis part of the study might have difficulty giving a single rating to a two-part problem description, and judges might weight the title and description sections differently.
The following sentence was added to the instructions about writing descriptions: "If you put images in your own reports you may include them with this description."	One practitioner pilot participant did not realize that inserting images was permitted, and commented that a typical report would have screen shots. The sentence was added to indicate that adding images was possible, while trying to avoid prompting evaluators that do not typically include images or screen shots to add them just for this study.
The single-letter category codes from CUE-4 were replaced by two-letter codes.	Several pilot participants commented that the single-letter codes were difficult to remember.

Table 3.4 Study 4: CUE-4 Codes and Current Comment Categorization Scheme

CUE-4 Code	New Code	Category	Description
C	PF	Positive finding	This approach is recommendable and should be preserved
P	MP	Minor problem	Caused test participants to hesitate for a few seconds
Q	SP	Serious problem	Delayed test participants in their use of the website for 1 to 5 minutes, but eventually they were able to continue. Caused occasional "catastrophes"
R	CP	Critical problem	Caused frequent catastrophes. A catastrophe is a situation where the website "wins" over the test participant, i.e. a situation where the test participant cannot solve a reasonable task or where the website annoys the test participant considerably
A	GI	Good Idea	A suggestion from a test participant that could lead to a significant improvement of the user experience.
T	B	Bug	The website works in a way that's clearly not in accordance with the design specification. This includes spelling errors, dead links, scripting errors, etc

3.3.1.3 Post-Task Questionnaire

The post-task questionnaire consisted of four sections.

1. Form to upload the completed usability report
2. The importance/difficulty questionnaire from Study 3
3. A demographic questionnaire
4. A page of open-ended questions about this study and the evaluator's typical usability practice

The statement framing the importance/difficulty semantic differential rating scales was changed from "When a usability practitioner writes a description of a single usability problem, addressing this quality of a problem description is generally..." to "When I write a description of a single usability problem, I consider addressing this quality of a problem description..." The phrasing in Phase I was designed to assess prescriptive opinions about what practitioners should do in general, whereas the phrasing in this questionnaire was designed assess personal opinion about what that particular evaluator thought was important in her own work. The complete text of the post-task questionnaire is included in Appendix E.

3.3.2 Participant Procedure

All evaluator participants received a randomly generated four-letter participant identification code in the cover letter for their study packets. The evaluators then did the following.

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

The upload form was disabled after the participant began the post-task questionnaire so that the evaluator could not change the report after reading the questionnaire. Participants could revise all three pages of the post-task questionnaire until

they indicated that they were completely finished, at which point the participant was locked out of the questionnaire, and a notification email was sent to the experimenter.

3.4 Results

The 44 evaluators wrote 512 comments, including 409 usability problem descriptions, 91 positive findings, and 12 extra comments (executive summary, redesign suggestions, participant action log, etc. from 10 practitioners and two students). Table 3.5 presents a summary of comments submitted by each group in the study, practitioner and student evaluators. One comment may describe multiple usability problems, and one usability problem may be described in multiple comments. Using a single multivariate analysis of variance (MANOVA) and an alpha level of .10, there was not a significant difference between students and practitioners across the 12 dependent measures, $F(11, 32) = 1.03$, $p = .44$. The SAS code and relevant output are included in Appendix H.

Table 3.5 Study 4: Summary of Reports and Comments Collected

	Practitioners			**Students**		
	M	**SD**	**Range**	**M**	**SD**	**Range**
Comments	**9.71**	6.30	3–30	**12.87**	13.19	4–53
Critical	**0.62**	0.86	0–3	**0.87**	1.06	0–4
Serious	**2.81**	2.02	0–8	**2.61**	2.15	0–8
Minor	**3.81**	3.67	0–16	**6.17**	7.06	1–32
Positive	**1.29**	1.38	0–4	**2.78**	5.64	0–26
Good Idea	**0.52**	1.40	0–5	**0.22**	0.42	0–1
Bug	**0.43**	0.81	0–3	**0.13**	0.34	0–1
Other	**0.24**	0.54	0–2	**0.09**	0.42	0–2
Words*	**578.05**	245.27	179–1063	**626.09**	534.20	162–2543
Words per comment*	**73.20**	47.93	31.36–212.60	**56.76**	22.38	16.92–90.11
Images or tables	**1.14**	1.59	0–5	**1.48**	3.88	0–16
Hours spent on evaluation	**2.02**	1.52	0.5 – 6.5	**1.73**	0.88	0.5 – 4.0

*Excluding extra comments and tables

This section first describes the procedure for matching similar UPDs and measuring problem detection. It then presents the analyses for each of the research questions and hypotheses. Table 3.6 explains the terms used to describe the people involved in this study.

Table 3.6 Study 4: Groups of People Involved in the Study

Group	Description
4 Users	Performed the task in the usability session movie.
1 Facilitator	Managed the usability session and interacted with the users.
58 Evaluators	Watched the usability sessions and wrote reports; *participants* of this study.
14 Practice	5 pilot and 2 dropped participants from this study, and 7 pilot participants from Long et al. (2005); about half practitioners and half students
21 Practitioners	Evaluators with practical experience
23 Students	Students that have completed a graduate course in Usability Engineering
1 Coder	Helped create the initial version of the master problem list.
3 Judges	Coded and rated the collected usability reports, created the final version of the master problem list, decided which problems in the master list are real.

3.4.1 Measuring Problem Detection

According to Bastien and Scapin (1995), there are three primary measures of problem detection in formative usability evaluation. Briefly, they are:

- **Thoroughness:** How many problems did each evaluator find?
- **Validity:** How many of the problems found are truly problems?
- **Reliability:** Did different evaluators find the same problems?

All of these measures rely on having a list of all the usability problems present in an interface and a list of all the usability problems identified by the evaluators. It is also necessary to know which problems are truly problems and which problems are false alarms, such as problems predicted by an evaluator that actual users would not find problematic. This section begins with a description of how the list of all problems was developed, which will be referred to as the master problem list (MPL), and how the MPL

was separated into real problems and false alarms. It then describes the calculations for thoroughness, validity and reliability.

3.4.1.1 Creating the Master Problem List

The master problem list used in this study was the complete list of all problems found in the usability session movie, including both real problems and false alarms. The process of taking all of the problems found by all evaluations and creating a master list is frequently called "de-duplicating" because it involves identifying or merging similar problem descriptions written by different evaluators or experienced by different users. This matching process can be difficult because different evaluators may report problems at different levels of granularity, with one evaluator combining into a single description a set of problems reported individually by a different evaluator. For this study, a usability problem was defined as a problem experienced by the user, which is caused by an interaction flaw, as illustrated in Figure 3.11. A description that only contains a suggestion for a solution implies the problem that it is fixing. Two problems, A and B, were considered the same if fixing problem A also fixes problem B **and** fixing problem B also fixes problem A. This is based on the criteria used in the analysis of usability reports collected in CUE-4 (Molich & Dumas, 2006).

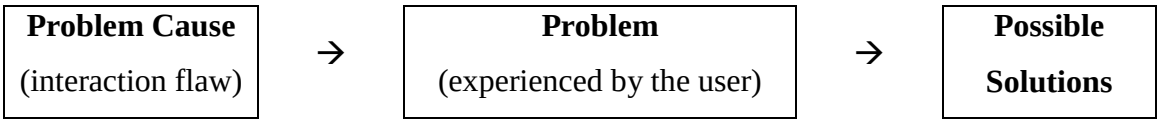


Figure 3.11 Study 4: Usability Problem Definition

To reduce experimenter bias in creating the MPL and matching evaluator UPDs to problems in the MPL, four graduate students in HCI assisted with this process: one coder and three judges. The entire process is explained in Table 3.7 and illustrated in Figure 3.12. The judges reviewed entire reports at a time, with severity ratings (except for "positive" and "good idea") removed, and with no indication of whether the evaluator was a practitioner or a student. The reports were separated into three groups. The order of the three groups was rotated for each judge, and report order within a group was randomized. The complete set of materials given to each judge is included in Appendix F.

The MPL was created using the 14 practice reports and 44 study participant reports. It is possible that there are additional usability problems in the movie, however there should not be many since the list was based on reports written by 58 different evaluators. The final list included 41 usability problems plus other items that the judges coded but were not specifically usability problems: three instances in the session movie where the facilitator prompted the user, a possible error in the testing protocol, and aspects of the interaction that evaluators misunderstood or misreported. Appendix G contains the complete list of usability problems found in the movie with screen shots for each problem.

Table 3.7 Study 4: Steps for Creating the Master Problem List

Step		Hours
1.	The **experimenter** and **coder** independently reviewed fourteen practice reports and created a list of usability problems mentioned. • Five pilot participants • Two participants dropped from the study • Seven reports contributed by an external researcher	5 ea.
2.	The **experimenter** combined the lists into a first version of the MPL	10
3.	The **three judges** independently reviewed two practice reports, matching UPDs to problems in the MPL, and then met individually with the **experimenter** to discuss the process.	5 ea.
4.	The **three judges** independently reviewed each UPD in each usability report from study participants. For each UPD, the judges did the following. • The judge identified the problems in the MPL discussed in the UPD. • If the UPD described a problem not already in the MPL, the judge created a new problem. • If the UPD was too vague to understand, the judge marked it as vague. • If the UPD contained only complimentary/neutral statements, the judge marked it as positive. For each suggestion by a judge (matching a UPD to a problem in the master list, marking as positive, or marking as vague), the suggestion was accepted if at least two of the three judges made the same suggestion.	20-40 ea.
5.	The **three judges** met to reconcile the problems they found that were not in the master problem list. A two-out-of-three vote of the judges was sufficient to add a problem to the MPL, delete a problem from the MPL, or merge two problems in the MPL. The **experimenter** participated in the discussion, but did not have a vote.	4
6.	The **experimenter** reviewed the results of the meeting in the previous step and suggested three changes to the MPL to the judges (two deletions, one merging). The **judges** approved the three changes via email voting.	5
7.	The **experimenter** made a list of all matches, positive ratings, and vague ratings suggested by a single judge. The experimenter reviewed each of these suggestions and decided which to keep and which to discard. The experimenter then resolved any conflicts (e.g. a problem rated both positive and vague), resulting in the final lists of matches between UPDs and problems in the MPL, positive UPDs and vague UPDs.	20
TOTAL		~170

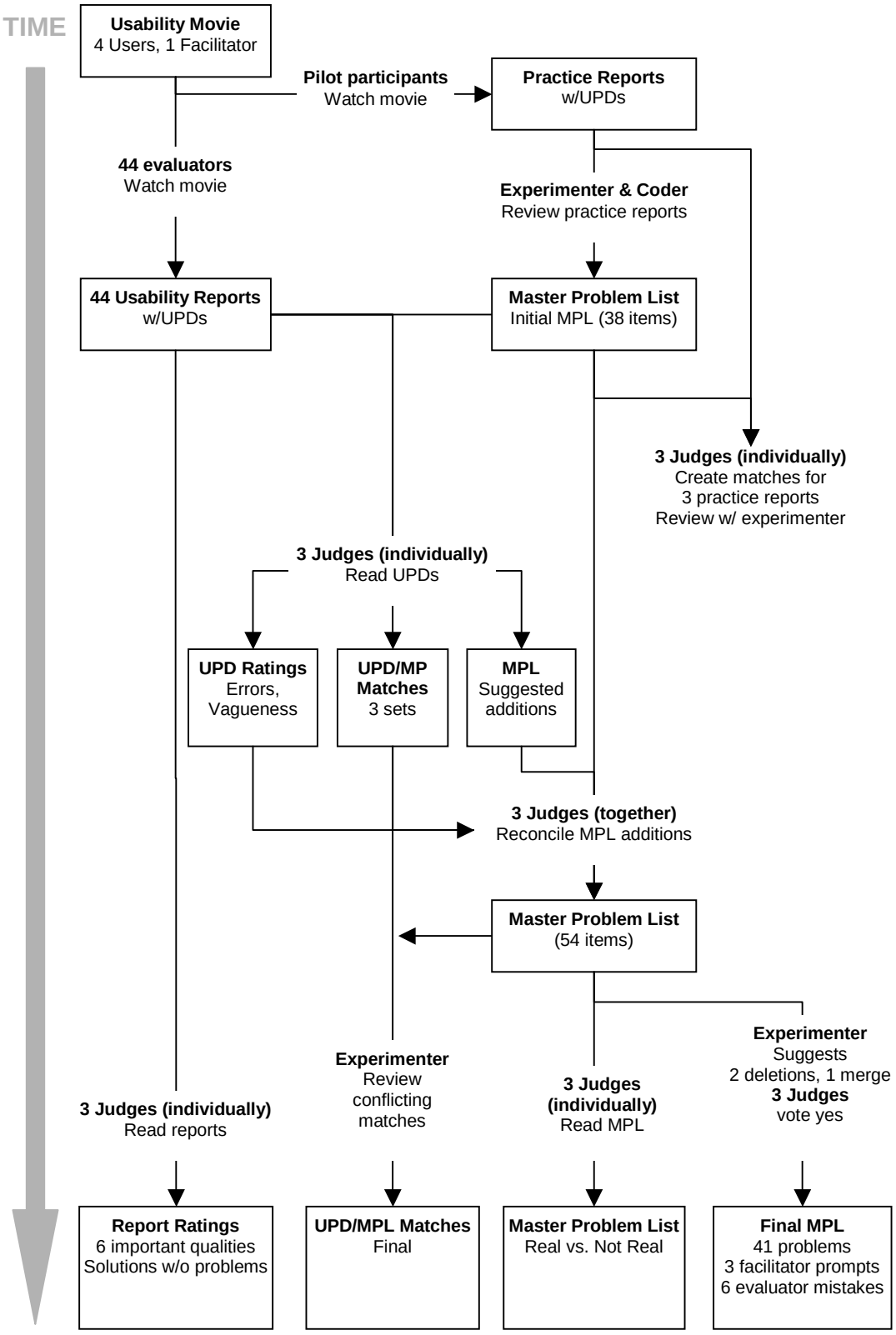


Figure 3.12 Study 4: Usability Report Coding Process

3.4.1.2 Judging Which Problems are "Real"

A "real" usability problem has two attributes: real users would encounter it in actual usage (as opposed to usage in the laboratory), and it has a significant impact on usability (Hartson et al., 2003). All formative usability evaluations have the potential to detect not only real problems but also false alarms. According to Hartson et al., this detection process can be characterized using terms from Signal Detection Theory, as explained in Table 3.8. False alarms can be problems predicted by an evaluator that would never be experienced by real users or trivial problems that do not have a significant impact on usability.

Table 3.8 Types of Problem Detection

Flaw Detected	**Interaction Flaw**	
by Evaluator?	**Real**	**Not Real**
Yes	Correct Hit	False Alarm
No	Miss	Correct Rejection

When counting usability problems detected during an evaluation, it is important to create a complete list of usability problems and to separate real problems from the false alarms. How should this benchmark list be created? Usability testing is sometimes used to create a benchmark problem set for comparing other UEMs, particularly analytical methods such as heuristic evaluation (Cuomo & Bowen, 1994; Desurvire et al., 1991; Desurvire et al., 1992; Law & Hvannberg, 2004b; Mankoff et al., 2005; Nielsen, 1994b). Problems found by both techniques are considered correct hits, problems found by usability testing but missed by the analytical evaluation are considered misses, and problems found by the analytical evaluation but not confirmed through usability testing are considered false alarms (or false positives). Usability testing is subject to its own biases, such as task selection (Cordes, 2001). However, any actual criteria for identifying real problems is, at best, an approximation of the ultimate criterion that matches the theoretical definition, as illustrated in Figure 3.13.

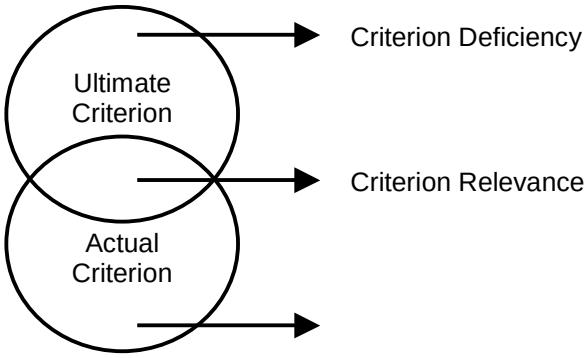


Figure 3.13 Ultimate vs. Actual Criterion

When evaluating usability testing itself, using the unfiltered union of all problems found by the usability test(s) is not appropriate; it results in all problems being considered real and no false alarms, meaning that all problem sets will be 100% valid. Hartson et al. (2003) suggest expert judgment to determine which problems are real. For example, experts can rate the severity of each problem, and the set of severe problems can be used as an approximation of the set of real problems. For this study, the three judges reviewed the final set of problems in the MPL (see Figure 3.12 for a diagram of the entire coding process) and rated the severity of each problem. The judges used three categories of severity, shown in Table 3.9. This is a subset of the categories used by the evaluators in Study 4 (see Table 3.4), and adapted from the scale used in CUE-2 (Molich et al., 2004). Problems rated as *critical* or *serious* by at least two of the three judges were considered real, and problems rated minor by at least two of the three judges were considered not real. The complete set of judgments is included in Appendix H.

Table 3.9 Study 4: Problem Severity Categories for Judges

Realness	Code	Category	Description
Not Real	MP	Minor problem	Caused test participants to hesitate for a few seconds
Real	SP	Serious problem	Delayed test participants in their use of the website for 1 to 5 minutes, but eventually they were able to continue. Caused occasional "catastrophes"
Real	CP	Critical problem	Caused frequent catastrophes. A catastrophe is a situation where the website "wins" over the test participant, i.e. a situation where the test participant cannot solve a reasonable task or where the website annoys the test participant considerably

Three aspects of the reliability of real/not real judgments between judges were assessed. Each judge was compared to each other judge (rater to rater) and to the final set of real problems (rater to group). The first aspect of reliability assessed was bias, which is the tendency of raters to give higher or lower ratings, or to have differences in rating thresholds. This was calculated using the McNemar change test with exact *p*-values, as summarized in Table 3.10. Using an alpha level of .05, judge B had a higher threshold for realness than judge C, $p = .02$, but none of the other differences were significant. The SAS code and relevant output are included in Appendix H.

Table 3.10 Study 4: Judge Bias for Master Problem Severity Ratings

$\chi^2(1)$ *p*	**B**	**C**	**Final**
A	**1.14** .29	**3.77** .09	**0.33** 1.00
B		**5.76** .02*	**2.27** .23
C			**3.60** .11

*$p < .05$.

The second aspect of reliability assessed was association, or the tendency of raters to give real or not real ratings to the same items. This was calculated using the tetrachoric correlation coefficient (r^*) for dichotomous variables with underlying continuous traits, as summarized in Table 3.11. The asymptotic standard error (ASE) from SAS was used to calculate *p*-values in Excel. Using an alpha level of .05, all correlation pairs were significant except B and C. Judge A had an almost perfect correlation ($r^* = .97$) with the final severity ratings.

Table 3.11 Study 4: Judge Association for Master Problem Severity Ratings

r* *p*	**B**	**C**	**Final**
A	**.44** .02*	**.62** .0003**	**.97** <.0001**
B		**.07** .39	**.68** <.0001**
C			**.78** <.0001**

* $p < .05$. ** $p < .001$.

The third aspect of reliability assessed was distribution, or the degree to which each judge used each category. This was assessed using visual inspection of the distributions of the judges' ratings, shown in Figure 3.14. Judge B appeared to use the extremes, *minor* and *critical*, more often, with the high use of minor ratings resulting in judgments of real for fewer problems. Judge C used fewer minor ratings, judging more problems real than judges A and B. The usage of minor ratings by judge A fell between judges B and C and was similar to the final ratings (see Appendix H).

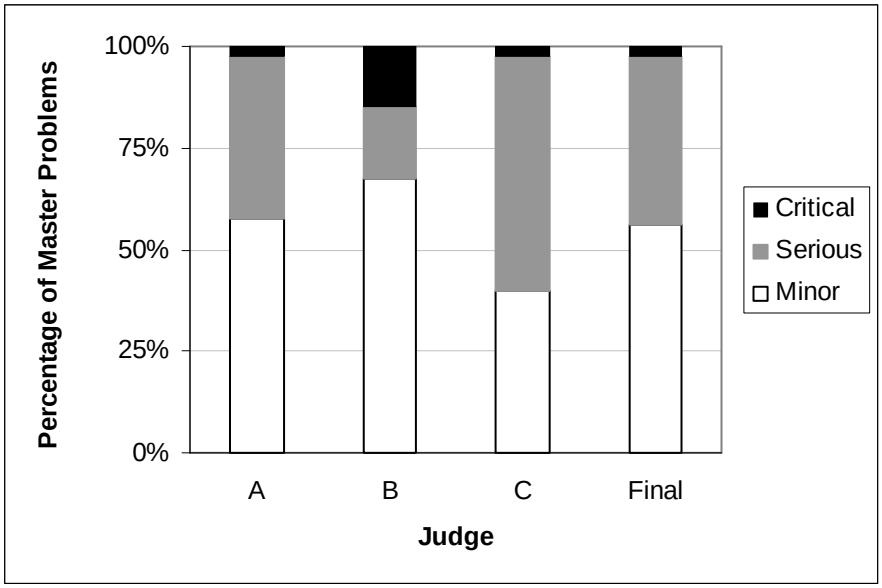


Figure 3.14 Study 4: Judge Distribution for Master Problem Severity Ratings

In summary, all three judges agreed on seventeen of the forty-one master problems. Judges B and C disagreed the most often, having the lowest association and a significantly different threshold for realness. Agreement between judges B and C accounted for three of the final ratings. Agreement with judge A accounted for the remaining 21 problems, split almost equally between judge B and judge C. The individual ratings for each judge are included in Appendix H.

Table 3.12 lists the master problems with the final decision of real and not real. **Bold** and an asterisk (*) indicate a problem rated severe by the judges. Problems la-lp were added by the judges to the initial list developed by the experimenter and coder. Appendix G contains a description of each usability problem in the master list. This table is based on problem reporting at any severity level. However, did an evaluator truly find a severe problem if the evaluator gave it a minor severity rating? This issue is discussed in detail in the following section.

Table 3.12 Study 4: Problem Detection Counts for Real/Not Real Problems

Master Problem	Real?	Master Problem	Real?	Master Problem*	Real?
aa		**ea**	**real**	**la**	**real**
ab	**real**	eb		lc	
ac	**real**	ec		lf	
ad		ed		**lg**	**real**
ae		**fa**	**real**	lh	
af		**fb**	**real**	li	
ag	**real**	**fc**	**real**	lj	
ah		**fd**	**real**	ll	
ba	**real**	fe		**lo**	**real**
bc		ff		**lp**	**real**
bd	**real**	ga		* Master list problems in this column were added by the judges.	
be		gb			
ca	**real**	gc			
da	**real**	ha			
db	**real**	**hb**	**real**		
		hc			

3.4.1.3 Determining when an Evaluator "Finds" a Problem

Determining whether an evaluator "found" a problem is not straightforward. You have to first determine whether the evaluator detected the problem. In the current study,

detection was determined by the judges through the process of matching master list problems to the problem descriptions, as described in the previous section; a problem was considered detected if it was discussed in a problem description, even if it was one of several problems discussed in the description. However, is it enough for an evaluator to detect the problem, or should the evaluator also have to rate problem as severe? Hertzum and Jacobsen (2003), in their meta-analysis of 11 studies of the evaluator effect, relied merely on detection of usability problems for calculating problem detection for severe problems. However, according to the operational definition of realness used in this study, minor problems are not real. It might follow that if an evaluator rated a problem as minor, the evaluator has decided that the problem is not real, and so did not truly "find" the problem. From a practical standpoint, problems rated as minor may be less likely to be fixed in system redesign, and so ensuring that a severe problem is fixed may rely on not only detecting it but also identifying it as severe.

The judges were told before rating problem severity that a rating of *serious* or *critical* indicated that the judge believed the problem was real. Since judges and evaluators in this study used the same definitions of *minor*, *serious* and *critical* problems, the same criterion can be used for the evaluators. However, the severity rating process for judges and evaluators was not identical. The evaluators did not know that the severity rating would be used to determine realness, and the evaluators had additional categories: bug, positive finding, and good idea. Thus, while the judges' intents in declaring a problem real or not real is clear, deciding evaluators' intents requires some interpretation of their submitted reports. Table 3.13 shows the interpretations of evaluator intent for each problem category used in determining problem severity ratings by evaluators.

Table 3.13 Study 4: Interpretation of Evaluator Intent for Single Problems

Problem Severity Rating	Severity Interpretation
Critical	Severe
Serious	Severe
Minor	Minor
Bug	Minor
Good Idea	Minor
Positive	Not found
(Problem not found)	Not found
(Mentioned in executive summary)	Not found

Problems that evaluators mentioned in UPDs categorized as a "positive findings" were not counted as "found" (seven UPDs submitted by six evaluators). A report recipient might not realize that a positive finding also describes a problem that needs to be fixed. Four of these problems were mentioned in another UPD with a *minor/serious/critical* rating, and three of these problems were mentioned only in a comment categorized as positive.

Thirteen evaluators submitted text in addition to the problem descriptions, such as an executive summary, log of participant actions, or redesign suggestion (the last row in Table 3.13). Seven of these evaluators mentioned problems in these extra texts. These 45 problems[1] were ignored. Of these problems, 34 were counted as found because they were also mentioned in a separate UPD with a minor or severe interpretation, but 11 were only discussed in extra text (1-3 problems for each of six evaluators).

Using the severity interpretations in Table 3.13, with minor, bug and positive interpreted as minor, and with *serious* and *critical* interpreted as severe, there were 378 UPDs with either a minor or severe rating. Determining an evaluator's intent is further

[1] This problem list is based on review by two judges; the third judge did not review these extra comments.

complicated by the complexity of the matching process, in that there is not a simple 1:1 mapping between problems in the MPL and UPDs submitted by an evaluator. There can be 1:n, n:1 and n:n mappings between problems in the MPL and UPDs submitted by an evaluator, as outlined in Table 3.14. One hundred and fifty one (151) UPDs of the 373 total UPDs were a 1:1 mapping between problems in the MPL and the UPD. Interpretation of these UPDs is straightforward: assign the severity rating of the UPD to its MPL problem. Also straightforward to interpret are the 70 UPDs that describe a single MPL problem, where the problem is described multiple times (1:n) but always at the same severity level: assign the severity rating of the UPDs to their MPL problem. It is possible that the evaluator might assign the problem a higher severity rating based on its multiple occurrences during the session, but this seems less likely. There were 56 UPDs where the MPL problem was described in multiple UPDs (1:n) with different severity ratings. In this case, it seems reasonable that the intended severity of the problem is the most severe rating given by the evaluator. It is possible that an evaluator would decide that the likelihood of other users having a severe reaction to a problem is low and so would give the overall problem a low severity rating, but this seems less likely.

Less clear is the interpretation of a single UPD that combines multiple problems that are not described in any other UPDs, an n:1 mapping between problems and UPDs. For the 25 of these UPDs where the evaluator gave a minor rating, it likely that all of the problems mentioned are minor; presumably, any problems mentioned in the description that the evaluator intended to rate as severe would be reported in a separate UPD with a severe rating. Most difficult to interpret are the 24 n:1 UPDs with a severe rating. Is the severe rating because all the individual problems are severe, one of the individual problems is severe, or because the individual problems combine to cause a severe problem? Interpretations that favor severe ratings will increase thoroughness but decrease validity; interpretations that favor minor ratings will decrease thoroughness but increase validity. These problems were interpreted as severe in this study.

Table 3.14 Study 4: Interpretation of Evaluator Intent for Multiple Problems

Problem:UPD mapping	Description	Interpretation of Intent	Number of UPDs	Number of Matches
UPD describes a single problem, and…				
1:1	…this evaluator does not describe the problem in any other UPD	Use the severity rating	151	151
1:n	… this evaluator describes the problem in other UPDs (1:n or n:n UPDs), all with the same severity rating	Use the severity rating	70	43
1:n	… this evaluator describes the problem in other UPDs (1:n or n:n UPDs) and gave these UPDs different severity ratings.	Use the highest severity rating assigned by the evaluator in a 1:n UPD.	56	27
UPD describes multiple problems, and…				
n:1	… this evaluator does not describe any of the problems in any other UPD	Use this severity rating for each individual problem	49	116
n:n	… there are some problems that this evaluator does not describe in any other UPD	Use this severity rating for each individual problem		
	… there are some problems that this evaluator describes in one or more 1:n UPDs	Use the severity rating from the problem description where the problem is reported alone.	52	56
	… there are some problems that this evaluator never describes in a UPD that describes just that problem	Use the highest severity rating assigned by the evaluator		
TOTAL			378 UPDs	393 matches

Note: a "match" is a problem found by a specific evaluator. An evaluator describing one problem multiple times is one match, and two evaluators describing the same problem is two matches. The last column indicates the number of matches whose severity rating was determined by the UPDs in each row.

Finally, there are UPDs that describe multiple MPL problems, and each problem is also described in other UPDs, an n:n mapping. For the 37 problems that were also described in a UPD where they were the only problem described, the severity rating from the 1:n UPD was assigned. The 20 problems that were only described in minor n:n UPDs were assigned a minor rating. Problems described in severe n:n UPDs have the same difficulty of interpretation as severe n:1 UPDs. Is the group of problems severe because one or more of the problems mentioned is severe, or because the group combined to cause a serious problem? The 21 problems only described in severe n:n UPDs and the nine problems described in both minor and severe n:n UPDs were assigned a severe rating.

Table 3.15 summarizes the number of minor and severe severity ratings given to each problem in the master list by each group of evaluators, using the interpretations of evaluator intent from Table 3.13 and Table 3.14. **Bold** and an asterisk (*) indicate a problem rated severe by the judges. Problems la-lp were added by the judges to the initial master problem list created by the experimenter and the coder.

Table 3.15 Study 4: Problem Detection Counts by Interpreted Severity Rating

MP	Practitioners			Students			All		
	Minor	Severe	All	Minor	Severe	All	Minor	Severe	All
aa	4		4	2		2	6		6
ab*	**8**	**11**	**19**	**13**	**8**	**21**	**21**	**19**	**40**
ac*	**7**	**3**	**10**	**7**	**3**	**10**	**14**	**6**	**20**
ad	--	--	--	1		1	1		1
ae	4	1	5	3		3	7	1	8
af	1		1	--	--	--	1		1
ag*	**1**		**1**	**2**	**1**	**3**	**3**	**1**	**4**
ah	--	--	--	1		1	1		1
ba*	**3**	**17**	**20**	**4**	**17**	**21**	**7**	**34**	**41**
bc	2	1	3	--	--	--	2	1	3
bd*	**1**	**1**	**2**	**2**	**3**	**5**	**3**	**4**	**7**
be	1	1	2	2	3	5	3	4	7
ca*	**1**		**1**	**1**		**1**	**2**		**2**
da*	**1**	**7**	**8**		**4**	**4**	**1**	**11**	**12**
db*		**1**	**1**		**1**	**1**		**2**	**2**
ea*	**2**	**13**	**15**	**5**	**5**	**10**	**7**	**18**	**25**
eb	1		1	1		1	2		2
ec	1	1	2	1		1	2	1	3
ed	2		2	2		2	4		4
fa*	**6**	**12**	**18**	**13**	**7**	**20**	**19**	**19**	**38**
fb*	**3**	**3**	**6**		**4**	**4**	**3**	**7**	**10**
fc*	**4**	**11**	**15**	**11**	**4**	**15**	**15**	**15**	**30**
fd*	**6**	**4**	**10**	**6**	**6**	**12**	**12**	**10**	**22**
fe	3		3	6	3	9	9	3	12
ff	2	6	8	8	3	11	10	9	19
ga	1	1	2		1	1	1	2	3
gb	1		1	4	2	6	5	2	7
gc	7	6	13	14	5	19	21	11	32
ha	2	1	3	1		1	3	1	4
hb*	**--**	**--**	**--**		**2**	**2**		**2**	**2**
hc	5		5	1	1	2	6	1	7
la*	**2**	**2**	**4**		**1**	**1**	**2**	**3**	**5**
lc	1	1	2	--	--	--	1	1	2
lf	4		4	--	--	--	4		4
lg*	**--**	**--**	**--**		**1**	**1**		**1**	**1**
lh	--	--	--	1		1	1		1
li	--	--	--	1		1	1		1
lk	--	--	--	--	--	--	--	--	--
ll	--	--	--	1		1	1		1
lo*	**--**	**--**	**--**	**1**		**1**	**1**		**1**
lp*	**--**	**--**	**--**	**1**		**1**	**1**		**1**

*** rated severe by the judges**

3.4.1.4 Thoroughness

Thoroughness is a measure of the number of problems detected out of the set of entire problems. Sears (1997) provides this theoretical definition of thoroughness:

$$\text{Thoroughness}_{\textit{Theoretical}} = \frac{\text{number of real problems found}}{\text{number of real problems that exist}} \quad (3.1)$$

Some studies measure thoroughness based on all problems found, both real and not real, as shown in Equation 3.2. This definition counts false alarms as correct hits, making all problem sets 100% valid. A closer approximation of the theoretical definition is to consider only the subset of problems deemed real by some criterion (Hartson et al., 2003), which in this study is expert judgment of problem severity. If detecting a real problem, regardless of the severity rating given, is sufficient for finding it, then an appropriate equation for thoroughness would be Equation 3.3. If finding a real problem involved not only detecting a problem but also rating it severe (using the interpretations of evaluator intent discussed in the previous section), then an appropriate equation for thoroughness would be Equation 3.4. All three of these measures of thoroughness were calculated for this study. The first two measures are important to be able to compare the results of this study to previous studies, and the third measure is consistent with the definition of realness used in the current study.

$$\text{Thoroughness}_{\textit{All Problems}} = \frac{\text{number of problems found by this evaluation}}{\text{number of problems in master problem list}} \quad (3.2)$$

$$\text{Thoroughness}_{\textit{Severe Problems}} = \frac{\text{number of severe problems found by this evaluation}}{\text{number of severe problems in master problem list}} \quad (3.3)$$

$$\text{Thoroughness}_{\textit{Severe Problems Marked as Severe}} = \frac{\text{number of severe problems found by this evaluation and marked as severe by this evaluator}}{\text{number of severe problems in master problem list}} \quad (3.4)$$

3.4.1.5 Validity

Validity is a measure of the degree to which an evaluation results in real problems versus false alarms. Sears (1997) provided this theoretical definition of validity:

$$\text{Validity}_{\textit{Theoretical}} = \frac{\text{number of real problems found}}{\text{number issues identified as problems}} \quad (3.5)$$

Using terms from Signal Detection Theory, this equation can be rephrased as Equation 3.6.

$$\text{Validity}_{\textit{SDT terms}} = \frac{\text{correct hits}}{\text{correct hits + false alarms}} \quad (3.6)$$

Using the union of all problems found by all evaluators as the set of "real" problems poses a problem for measuring validity. If only problems are considered in calculations, and if all problems are considered "real," then all problem sets are 100% valid (Hartson et al., 2003). However, evaluators will sometimes describe a problem that truly is not a problem. For example, Desurvire, Kondziela and Atwood (1992) found that 55% of problems reported by non-experts (untrained end-users of the system) were not possible in the interface, and were due to misinterpretation of system functions. These not-possible problems can be considered false alarms, in which case validity can be measured using Equation 3.7.

$$\text{Validity}_{\textit{Possible Problems}} = \frac{\text{number of problems in master problem list found}}{\text{number of problems listed}} \quad (3.7)$$

Sears (1997) suggests that if the experimenter is only interested in severe problems, validity could be measured as the ratio of severe problems to all problems as in Equation 3.8. This matches the "severe" thoroughness measure (Equation 3.3), which measures how many severe problems were found regardless of the severity rating assigned.

$$\text{Validity}_{\textit{Severe vs. Minor \& Severe}} = \frac{\text{number of severe problems in master problem list found}}{\text{number of all problems found}} \quad (3.8)$$

In this study, "real" has been defined as problems rated severe by the judges. If a severe rating is required to consider a problem found, corresponding to the "severe marked severe" thoroughness measure (Equation 3.4), then an appropriate validity measure would be in Equation 3.9. This equation compares the severe problems that the evaluators also rated severe (identified as real that are real) to the number of problems, minor and severe, that the evaluators rated severe (identified as real).

$$\text{Validity}_{\textit{Severe Problems Marked as Severe}} = \frac{\text{number of problems in master problem list identified as severe that are severe}}{\text{number of problems identified as severe}} \quad (3.9)$$

3.4.1.6 Reliability

Reliability in problem detection is a measure of the degree to which evaluators tend to find the same problems. High reliability indicates that the evaluation can be repeated to produce similar results. Low reliability indicates that the results of several individual evaluations can be combined for higher overall thoroughness (Hartson et al., 2003). Reliability can also be used to describe problem severity ratings, but this section focuses on problem detection.

A common measure of reliability in usability problem detection is what Hertzum and Jacobsen (2003) call *any-two agreement*, defined in Equation 3.10. It is a measure of the overlap between evaluators, and is equivalent to the mean Jaccard score (Equation 2.1) across all pairs of evaluators. Jaccard scoring is just one way of measuring similarity; many other scoring techniques exist (see Everitt, 1974; Kaufman & Rousseeuw, 1990; Lorr, 1983). It does not take into account the relationship of an individual's problem set to the entire set of problems, just to other individuals' problem sets. For example, it is insensitive to problems that none of the evaluators detected; evaluators receive a high score as long as they find the same problems and miss the same problems. Reliability can be calculated using the same three interpretations as thoroughness: using all problems

found, using only severe problems found, and using only severe problem found and marked as severe.

$$\text{Reliability}_{\textit{Any-two agreement}} = \frac{\sum_{i=2}^{n}\sum_{j=1}^{i-1}\frac{\left|P_i \cap P_j\right|}{\left|P_i \cup P_j\right|}}{n(n-1)/2} \quad \text{where } n \text{ is the number of evaluators and } P_i \text{ is the set of problems found by evaluator } i \qquad (3.10)$$

Hartson et al. (2003) suggest kappa as a measure of reliability in problem detection, and Long, Styles, Andre & Malcolm (2005) used it in a study very similar to the current study. However, by using trait prevalence (both positive and negative) to adjust the measure, kappa is dependent on marginal rates, making kappa values difficult to compare across studies (Cook, 1998). Kappa conflates association of judgment (tendency to rate the same items positively or negatively) with judgment bias (differences in drawing the line between positive and negative ratings). In the context of the current study, this means that kappa would depend on both agreement and thoroughness, making it difficult to compare kappa across different samples, such as all problems and severe problems. Measures of association only, such as tetrachoric correlation (or polychoric, for multiple ordered categories), are preferable to kappa (Hutchinson, 1993). For the current study, any-two agreement was used to measure level of agreement, and thoroughness was measured separately.

3.4.1.7 Overlap

Several studies comparing multiple evaluators report results in terms of evaluator overlap, reporting number of *evaluators per problem*, as in Table 3.20 (Molich et al., 1998; Molich et al., 2004; Nielsen, 1992). This is in contrast to the thoroughness measure used in the current study, which is expressed in terms of number of *problems per evaluator*. However, overlap is equivalent to thoroughness when expressed as a grand mean and converted to a percentage. To see how these are equivalent, begin with the total problem detection count, or the total number of problems detected across all evaluators and all problems, including between-evaluator duplicates. Total problem detection can be calculated from the set of problems found by each evaluator, as shown in Equation 3.11.

$$\textit{Total Problem Detection (T)} = \sum_{i=1}^{n} \left| P_i \right| \quad \text{where } n \text{ is the number of evaluators and } P_i \text{ is the set of problems found by evaluator } i \qquad (3.11)$$

Total problem detection can also be calculated from the number of evaluators detecting each problem, as shown in Equation 3.12.

$$\textit{Total Problem Detection (T)} = \sum_{i=1}^{p} \left| E_i \right| \quad \text{where } p \text{ is the number of problems and } E_i \text{ is the set of evaluators that found problem } i \qquad (3.12)$$

Mean overlap (*T/p*, or evaluators per problem), when expressed as a percentage of evaluators in the study (*n*), can be transformed into mean thoroughness (*T/n*, or problems per evaluator), when expressed as a percentage of problems in the study (*p*), as shown in Equation 3.13.

$$\underset{\text{(as a \% of evaluators)}}{\textit{Mean Overlap}} = \frac{T/p}{n} = \frac{T}{np} = \frac{T/n}{p} = \underset{\text{(as a \% of problems)}}{\textit{Mean Throughness}} \qquad (3.13)$$

The current study focuses on differences among evaluators rather than problems, and so calculations will generally be expressed in terms of thoroughness. However, the equivalence of these two measures allows comparison of the results of the current study to other studies that reported problem detection in terms of evaluators per problem, rather than problems per evaluator. This transformation only allows informal comparisons of means – the equivalence does not transfer to standard deviations and standard errors, precluding formal *t*- or *F*-tests of sample differences.

3.4.2 RQ1: How Do Practitioners Describe Usability Problems?

3.4.2.1 Hypothesis 1a: Practitioners in Study 3 vs. Study 4

Four 2x10 mixed-factor ANOVAs were performed to test for differences in opinions of the guidelines for each of the four adjectives (*difficult*, *helpful*, *relevant*, and *required*). Study (S) was a between-subjects factor with two levels (Study 3, Study 4), and Guideline (G) was a within-subject factor with 10 levels, one for each of the 10 guidelines from Study 2. A preliminary MANOVA across all four adjectives using an

alpha level of .01 indicated that both main effects and the interaction were significant. The results of the ANOVAs are summarized in Table 3.16. The results of the individual ANOVAs indicated that the main effect of Study was not significant for any of the four adjectives, but the main effect of Guideline was significant for all four adjectives. The Study x Guideline interaction was significant for *difficult* only. With an alpha level of .05, simple effects tests for each guideline across both Study groups were significant for *Politics/Diplomacy* only, $F(1, 837) = 11.36$, $p < .001$, with Study 4 practitioners having a lower *difficulty* rating ($M = 3.38$, $SD = 1.83$, $n = 21$) than Study 3 practitioners ($M = 4.73$, $SD = 1.69$, $n = 74$). Figure 3.15 illustrates the mean ratings for both groups. The SAS code and relevant output are included in Appendix H.

Table 3.16 Study 4: Guideline Ratings, Study 3 vs. Study 4 Practitioners

		Difficult		**Required**		**Relevant**		**Helpful**	
Effect	*df*	*F*	*p*	*F*	*p*	*F*	*p*	*F*	*p*
Between subjects									
Study (S)	1	0.89	.35	0.00	.98	0.82	.37	0.84	.53
S within-group error	93								
Within subjects									
Guideline (G)	9	12.31**	<.0001	12.61**	<.0001	10.86**	<.0001	13.01**	<.0001
G x S	9	2.38*	.011	0.35	.96	0.93	.50	1.23	.27
G x *S* within-group error	837								

S = subjects. * $p < .05$. ** $p < .0001$

3.4.2.2 Hypothesis 1b: Study 3 Practitioners vs. Study 4 Students

Four 2x10 mixed-factor ANOVAs were performed to test for differences in opinions of the guidelines for each of the four adjectives (*difficult*, *helpful*, *relevant*, and *required*). Study (S) was a between-subjects factor with two levels (Study 3, Study 4), and Guideline (G) was a within-subject factor with 10 levels, one for each of the 10 guidelines from Study 2. A preliminary MANOVA across all four adjectives using an alpha level of .01 indicated that both main effects and the interaction were significant.

The main effect due to Study was not significant for any of the four adjectives, $p > .05$. The main effect due to Guideline was significant for all four adjectives, $p < .0001$. The interaction between Study and Guideline was significant for *difficult* and *helpful* but not *relevant* and *required*. Post-hoc tests using effect slices by guideline found that students gave a higher *difficulty* rating for *Describe Methodology*, and lower *helpfulness* ratings for *Politics/Diplomacy* and *Professional/Scientific*, as shown in Figure 3.16. The difference in *helpful* ratings for *Impact/Severity* approach significance, $p = .054$. The SAS code and relevant output are included in Appendix H.

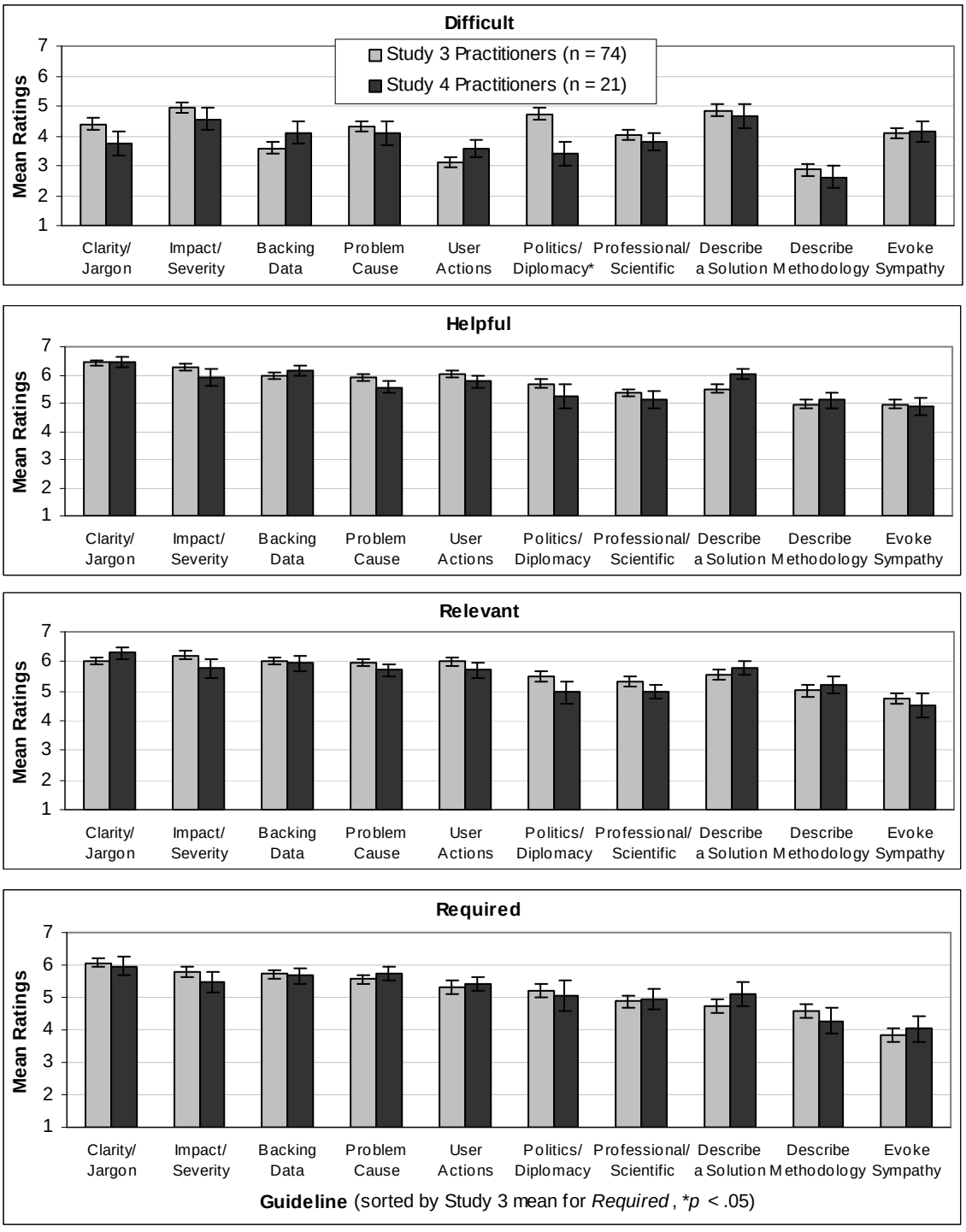


Figure 3.15 Study 4: Mean Guideline Ratings, Study 3 vs. Study 4 Practitioners

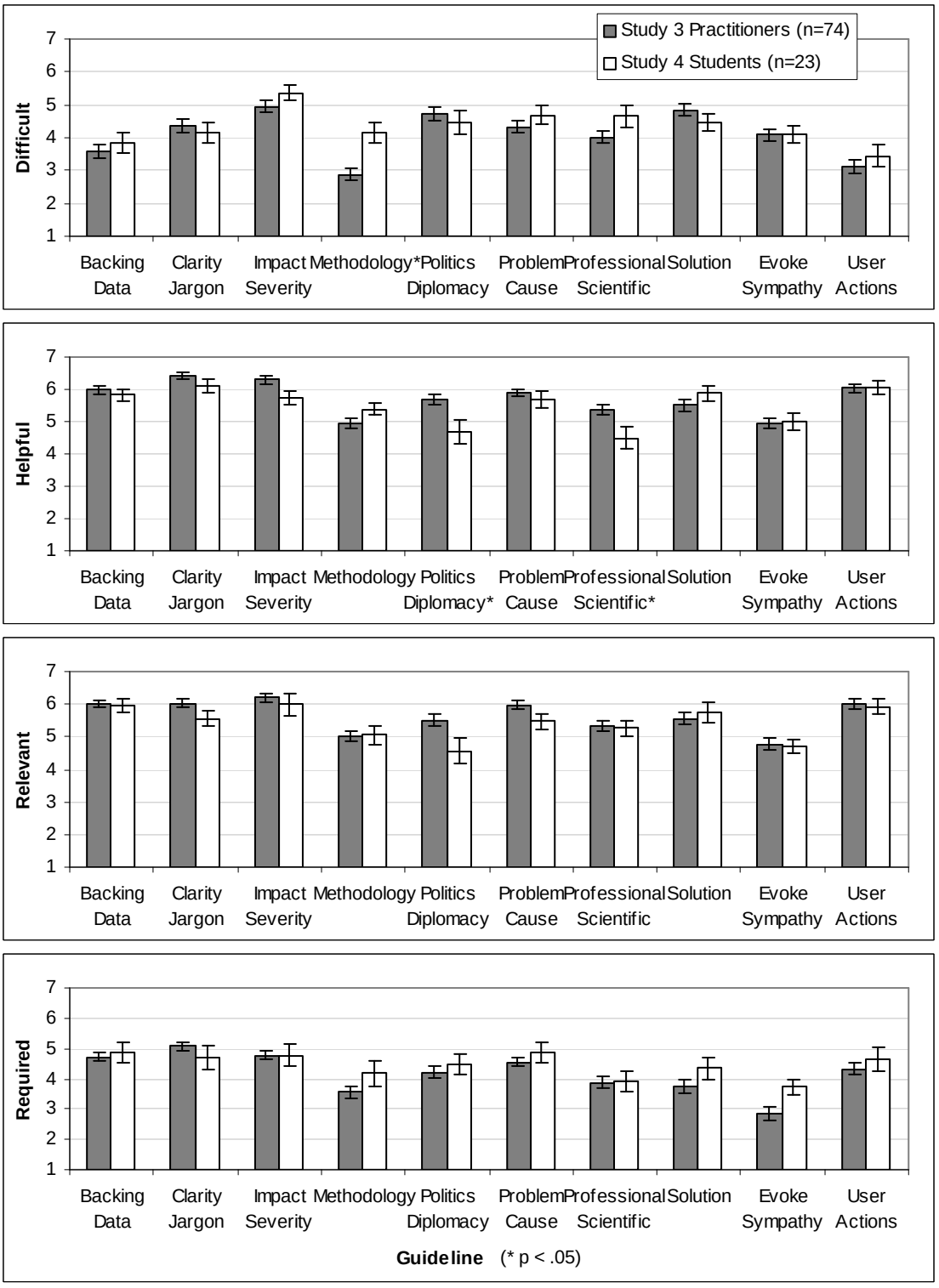


Figure 3.16 Mean Guideline Ratings, Study 3 Practitioners vs. Study 4 Students

3.4.2.3 Hypothesis 1c: Following the Guidelines

Three judges rated each of the reports for following six guidelines (*Clarity/Jargon, Backing Data, Impact/Severity, Problem Cause, User Actions,* and *Describe a Solution*) using a six-point Likert-type scale. These six guidelines represented the five most *required* guidelines and *Describe a Solution*, a topic about which there was divergent opinion in the previous studies. The rating sheet is included in Appendix E. The six-point scale was translated into a numerical value from –2.5 to 2.5, and the three ratings were summed, resulting in a theoretical range of –7.5 to 7.5 and an actual range from –6.5 to 6.5. Figure 3.17 summarizes the ratings for the practitioners in Study 4, with the box-and-whiskers representing quartiles, and the "+" signs representing the mean rating for each guideline. The dotted lines are the boundaries of a neutral rating (somewhat disagree to somewhat agree that the report follows the guideline, -1.5 to 1.5, inclusive).

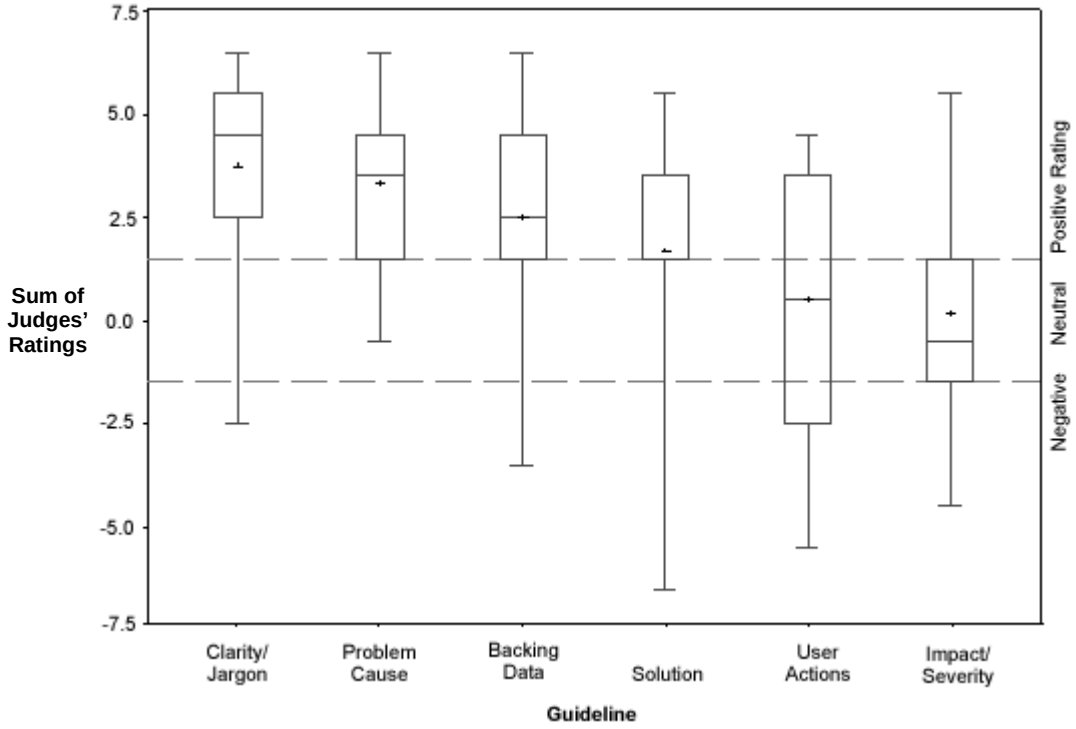


Figure 3.17 Study 4: Distribution of Practitioner Report Ratings for Guidelines

Figure 3.18 shows the frequency counts for positive, neutral and negative ratings. The only guideline that had more than 3 out of 21 negative ratings was describing using actions. The three negative ratings for *Describe a Solution* were the most negative (−6.5, −6.5 and −5.5) of all the guidelines except *User Actions*. The lowest rating for the other four guidelines (*Problem Cause, Clarity/Jargon, Backing Data* and *Impact/Severity*) was −4.5. *Impact/Severity* had few positive ratings, but that the judges' copies of the reports did not include the severity codes submitted for each problem.

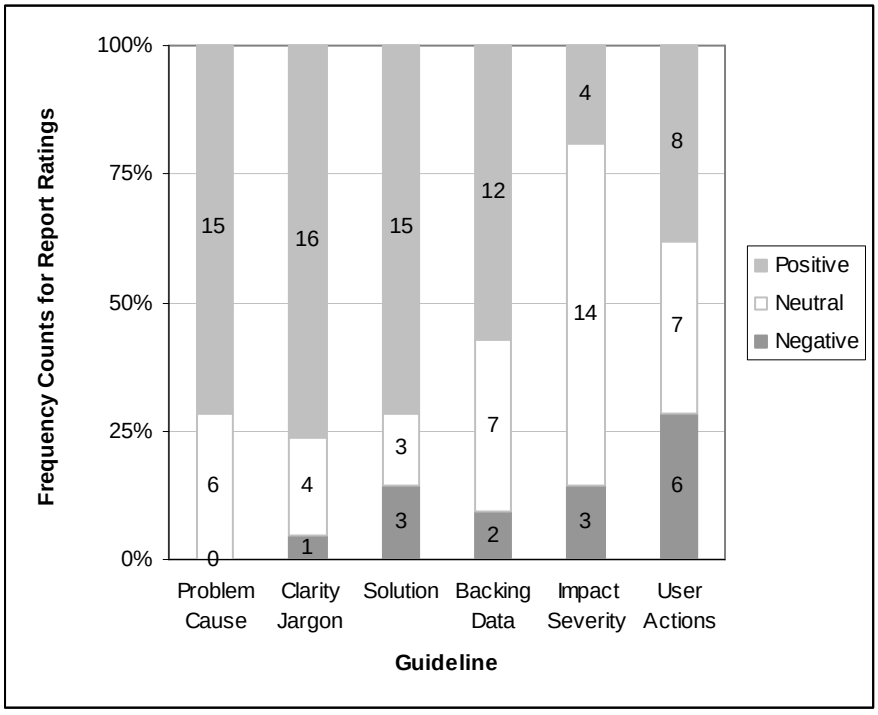


Figure 3.18 Study 4: Frequency Counts of Practitioner Report Ratings

3.4.2.4 Hypothesis 1d: Opinion vs. Behavior

Pearson correlations were calculated between evaluators' behavior (judges' ratings of how well each report followed each guideline) and opinion (ratings the evaluators gave each guideline for each adjective). Using an alpha level of .05, there were significant correlations for *Describe a Solution*, with the judges' rating of the evaluators' reports correlated with the evaluator's opinions on all four adjectives as shown in Table 3.17. There were no significant correlations between evaluator behavior and opinion for

any of the other guidelines. The SAS code and relevant output are included in Appendix H.

Table 3.17 Study 4: Opinion/Behavior Correlations for *Describe a Solution*

Adjective	*r*(19)	*p*
Difficult	-.51	.02*
Helpful	.58	.001*
Relevant	.57	.01*
Required	.67	.01*

* $p < .05$

3.4.3 RQ2: Is there an evaluator effect in usability testing?

Two aspects of the evaluator effect were assessed: problem discovery and severity ratings. Problem discovery was assessed in terms of thoroughness and reliability, and severity ratings were assessed in terms of the three aspects of reliability: bias, association and distribution.

3.4.3.1 Hypothesis 2a: Problem Discovery

Thoroughness. Three different measures of thoroughness were calculated, using Equations 3.2 through 3.4, with results summarized in Table 3.18 and illustrated in Figure 3.19. A 2x3 mixed-factor ANOVA was performed to test for effects due to type of evaluator (practitioners vs. students) and thoroughness measure (Equations 3.2 through 3.4). Using an alpha level of .05, there was not a significant effect due to type of evaluator, $F(1, 42) = 0.31$, $p = .58$. There was a significant effect due to thoroughness measure, $F(2, 84) = 61.47$, $p < .0001$, with "all" and "severe marked severe" significantly different from "severe" ($p < .0001$) but not each other ($p = .16$). There was not a significant interaction between type of evaluator and thoroughness measure, $F(2, 84) = 1.14$, $p = .33$. The SAS code and relevant output are included in Appendix H.

Table 3.18 Study 4: Problem Detection Thoroughness

Thoroughness Measure	Equation	# problems	Practitioners (n = 21)		Students (n = 23)	
			M	SD	**M**	SD
All	3.2	41	$.22_a$	.08	$.22_a$	.10
Severe	3.3	18	$.34_b$	.12	$.33_b$	.12
Severe marked severe	3.4	18	$.21_a$	.12	$.17_a$	.11

Note: means that do not share a common letter differed significantly ($p < .05$).

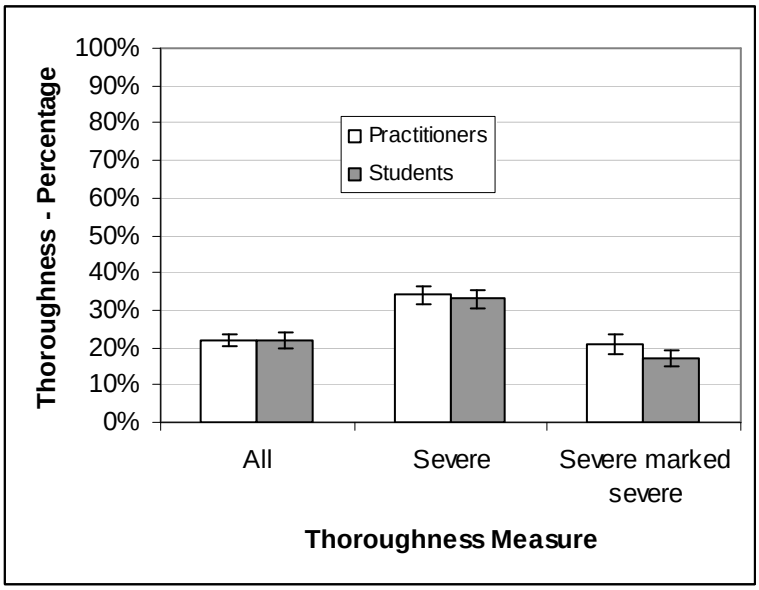


Figure 3.19 Study 4: Problem Detection Throughness

Reliability. Reliability was calculated using any-two agreement (Equation 3.10), or the average Jaccard score (Equation 2.1), across all pairs of evaluators. Table 3.19 summarizes the reliability for practitioners and students using each equation, and Figure 3.20 illustrates the reliability. Standard error is very small (<0 .1%) since reliability is calculated on the number of pairs, which is on the order of the square of the number of evaluators in each group, and so this graph does not include error bars. A 3x3 between-subject ANOVA was performed to test for effects due to type of evaluator (practitioners, students, between groups) and reliability measure (all problems, severe problems, and severe problems marked as severe) using the Jaccard scores for each pair of evaluators.

For each group of evaluators, only the reliability measures within the group were used. Reliability could not be treated as a within-subject factor because reliability is measured on pairs of evaluators rather than individual evaluators. Using an alpha level of .01, both main effects and the interaction were significant: type of evaluator, $F(2, 5667) = 39.58$, $p < .0001$, reliability measure, $F(2, 5667) = 617.07$, $p < .0001$, and the interaction between the two, $F(4, 5667) = 14.84$, $p < .0001$. The SAS code and relevant output are included in Appendix H.

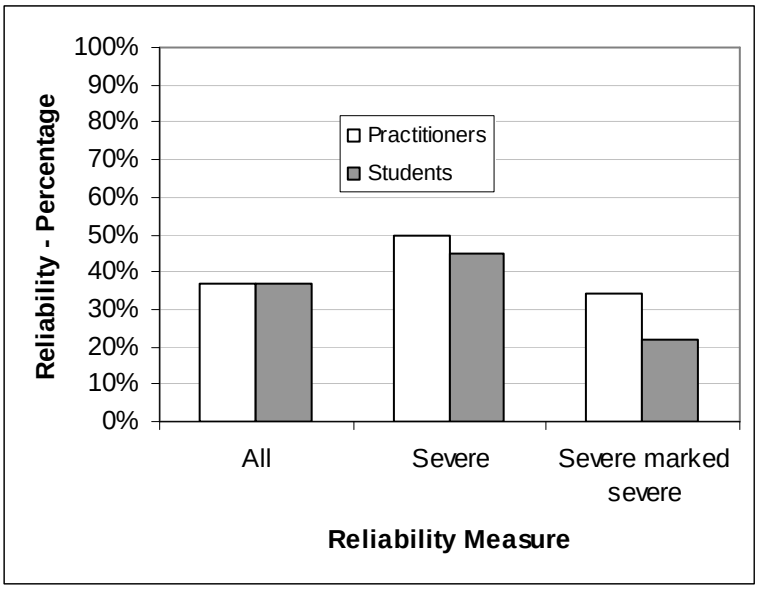


Figure 3.20 Study 4: Problem Detection Reliability – Any-Two Agreement

Table 3.19 Study 4: Problem Detection Reliability – Any-Two Agreement

Reliability Measure	# problems	Practitioners (n = 21*20)		Students (n = 23*22)		Everyone (n = 44*43)	
		M	SD	**M**	SD	**M**	SD
All	41	**.37$_a$**	.12	**.37$_a$**	.13	**.37**	.12
Severe	18	**.50$_b$**	.17	**.45$_c$**	.16	**.48**	.16
Severe marked severe	18	**.34$_a$**	.18	**.22$_d$**	.23	**.27**	.21

Note: Means for practitioners and students (but not for everyone) that do not share the same subscript differ significantly according to Tukey-adjusted post-hoc tests ($p < .0001$). * $p < 0.01$. * $p < .0001$.

Table 3.20 Study 4: Overlap in Reporting Problems

	All 41 Problems			18 Severe Problems			18 Severe Marked Severe			23 Minor Problems		
TOTAL	P	S	**A**	P	S	**A**	P	S	**A**	P	S	**A**
42-44 evaluators			**0**			**0**			**0**			**0**
40-41 evaluators			**2**			**2**			**0**			**0**
30-39 evaluators			**3**			**2**			**1**			**0**
20-29 evaluators	1	3	**3**	1	3	**3**	0	0	**0**	0	0	**1**
10-19 evaluators	7	6	**4**	6	4	**2**	5	1	**6**	0	1	**1**
5-9 evaluators	5	4	**7**	2	1	**2**	1	4	**2**	2	2	**5**
4 evaluators	3	2	**4**	1	2	**1**	1	3	**1**	3	1	**2**
3 evaluators	3	2	**3**	0	1	**0**	2	2	**1**	1	1	**2**
2 evaluators	6	4	**5**	1	1	**3**	1	1	**2**	4	3	**3**
1 evaluator (no overlap)	6	15	**9**	3	6	**3**	2	4	**2**	7	9	**8**
0 evaluators (not found)	10	5	**1**	4	0	**0**	6	3	**3**	6	6	**1**

Note: P = Practitioner, S = Student, A= All

3.4.3.2 Hypothesis 2b: Problem Severity

The reliability of severity judgments of the evaluators was assessed for just two levels of severity: severe (combining both the *serious* and *critical* categories) and minor. Reliability was measured using any-two agreement (Equation 3.10), which was previously used by Hertzum and Jacobsen (2003). Any-two agreement is equivalent to the mean Jaccard score (Equation 2.1) across all pairs of evaluators. In the current context, the Jaccard score of a pair of evaluators for severe problems is the ratio of problems they both considered severe (intersection of severe problems) to the number of problems at least one considered severe (union of severe problems). A similar score can be calculated for minor problems.

Table 3.21 shows the results for calculating any-two agreement across all pairs of evaluators, for all evaluators, just the practitioners, and just the students, for both severe ratings and minor ratings. For these analyses, severity ratings of bug or good idea were ignored – only comments specifically rated *minor*, *serious* or *critical* were considered.

This eliminated 17 out of 203 detections of problems previously interpreted as minor, and had no effect on the 189 detections interpreted as severe. The Jaccard score for a pair of evaluators is based on overlapping problem sets. Due to the low thoroughness and reliability in problem detection, many evaluators had no overlapping problems. The Jaccard score for a pair of evaluators was only calculated if the pair had overlapping problems, and at least one of the pair had rated at least one of those problems at the severity level of interest. The third line in each row of Table 3.21 shows the pair count used to calculate any-two agreement, out of 946 potential pairs across all evaluators, 210 for practitioners, and 253 for students. Figure 3.21 illustrates the reliability ratings.

Table 3.21 Study 4: Any-Two Agreement for Evaluators' Severity Judgments

M SD (Pair count)	**All** (946)	**Practitioners** (210)	**Students** (253)
Severe Ratings	**.46** .31 (933)	**.57** .28 (210)	**.39** .34 (245)
Minor Ratings	**.29** .33 (875)	**.20** .30 (186)	**.38** .35 (238)

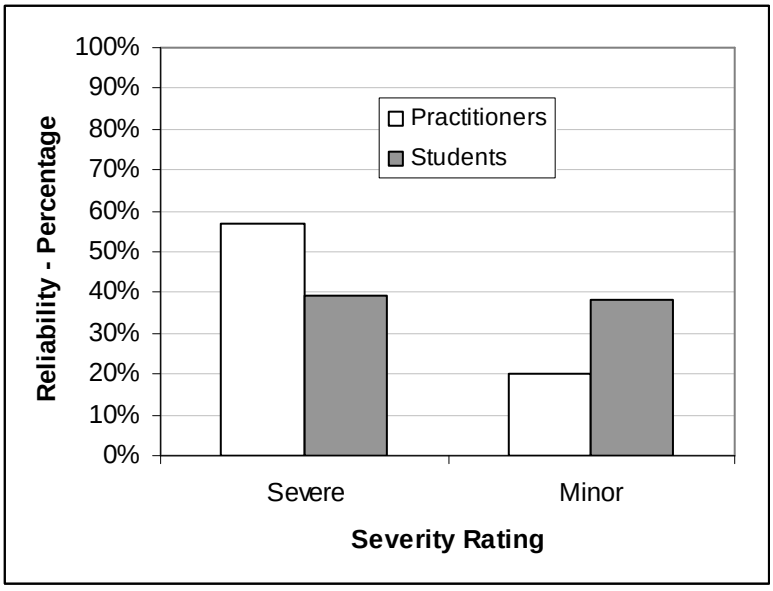


Figure 3.21 Study 4: Any-Two Agreement for Evaluators' Severity Judgments

Hertzum and Jacobsen (2003) also used Spearman's rank-order correlation for several studies where every evaluator rated every problem. This statistic could not be calculated for the current student because it requires paired ratings of all problems. In the current student, evaluators rated the severity of only the problems that they themselves found (an average of 22% of the complete problem set).

For the 12 problems that were detected by at least 10 evaluators, a chi-square goodness of fit test for equal proportions was used to determine if evaluators' ratings were equally divided between minor and severe. The calculations were then repeated separately for students and practitioners for any problem found by at least 10 evaluators in the sub-group. Table 3.22 summarizes the results. Using an alpha level of .05, ratings were divided unequally for 3 of the 12 problems (ba, da and ea, except for ea by students), all rated severe by the judges.

Table 3.22 Study 4: Equal Proportions of Evaluators' Severity Ratings

Master Problem	Group	All	Minor	Severe	$\chi^2(1)$	*p*
ab*	All	40	21	19	0.10	.87
	Practitioners	19	8	11	0.47	.65
	Students	21	13	8	1.19	.38
ac*	All	20	14	6	3.20	.12
	Practitioners	10	7	3	1.60	.34
	Students	10	7	3	1.60	.34
ba*	All	41	7	34	17.78	<.0001***
	Practitioners	20	3	17	9.80	.003***
	Students	21	4	17	8.05	.007***
da*	All	12	1	11	8.33	.006***
ea*	All	25	7	18	4.84	.04**
	Practitioners	15	2	13	8.07	.007***
	Students	10	5	5	0.00	1.00
fb*	All	10	3	7	1.60	.34
fc*	All	30	15	15	0.00	1.00
	Practitioners	15	4	11	3.27	.12
	Students	15	11	4	3.27	.12
fd*	All	22	12	10	0.18	.83
	Practitioners	10	6	4	0.40	.75
	Students	12	6	6	0.00	1.00
fe	All	12	9	3	3.00	.15
ff	All	19	10	9	0.05	1.00
	Students	11	8	3	2.27	.23
gc	All	32	21	11	3.13	.11
	Practitioners	13	7	6	0.08	1.00
	Students	19	14	5	4.26	.06

* Judged severe by the judges ** $p < .05$. *** $p < .01$.

3.4.4 RQ3: How can we assess the content of UPDs?

3.4.4.1 Hypothesis 3a: Good evaluators follow the guidelines

It is difficult to determine which evaluators are "good" evaluators based on the brief evaluations performed for the current study. However, several indicators of better evaluations were collected during this study:

- Type of evaluator, student or practitioner
- Level of experience: years of experience, number of evaluations
- Thoroughness (all problems, severe problems, severe marked severe)
- Validity (severe vs. minor, severe marked severe vs. marked severe)
- Hours spent on the evaluation

An analysis of variance (ANOVA) was used to test whether there were significant differences between report ratings for students and practitioners, and correlations were used to test for relationships with the remaining variables.

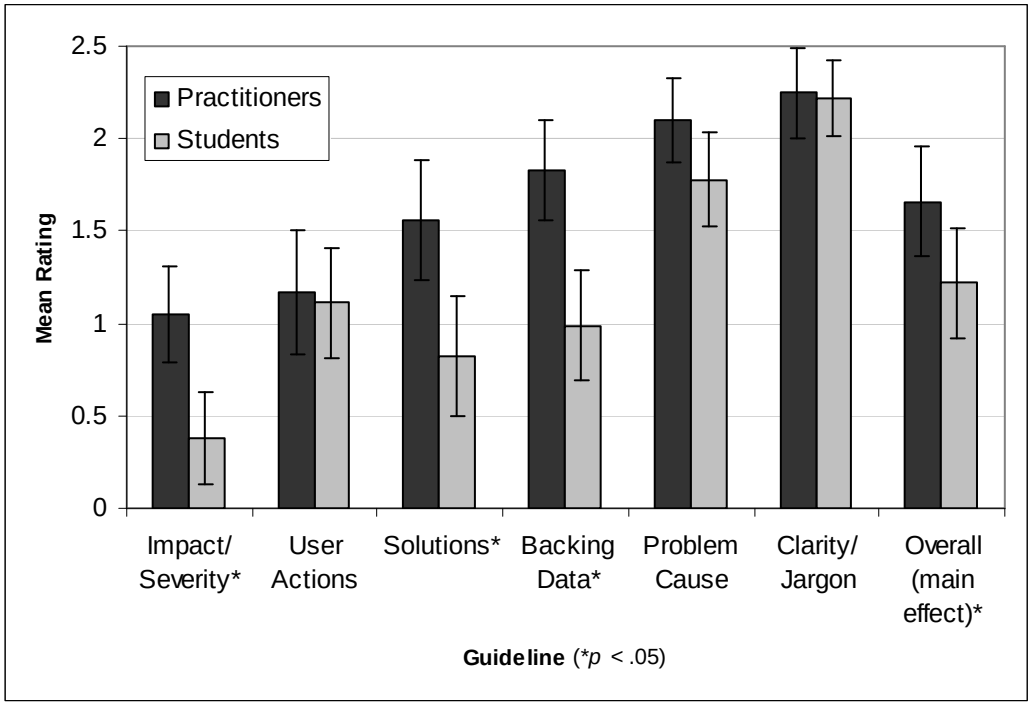


Figure 3.22 Study 4: Mean Report Ratings by Guideline, Evaluator Group

Figure 3.22 shows mean report ratings for each guideline for each evaluator group. Means are based on individual ratings given by each judge, rather than the sum of the three ratings. Judges rated on a 6-point scale, which has been adjusted to a rating from –2.5 to 2.5. Differences in report ratings for practitioners and students were tested as part of a 2x6x3 mixed-factor ANOVA, with group (practitioner and student) as a between-subject factor, guideline (the five *required* guidelines from Study 3 and *Describe a Solution*) and judge as within-subject factors, and evaluator as a repeated measure. Differences due to judge are reported in the next section, Hypothesis 3b: Rating Reliability. Table 3.23 summarizes differences between students and practitioners. Using an alpha level of .05, the main effect of group was significant, $F(1, 42) = 7.27$, $p = .01$, as was the main effect of guideline, $F(5, 210) = 30.34$, $p < .0001$. The interaction between guideline and group was also significant, $F(5, 210) = 3.13$, $p = .01$. Post-hoc tests using slices by guideline to test for simple effects due to evaluator group found that there were significant differences for *Backing Data*, *Describe a Solution,* and *Impact/Severity*, with practitioners receiving higher ratings than students did. There were no significant effects for *Problem Cause*, *Clarity/Jargon*, and *User Actions*. The three-way interaction between judge, guideline and group was not significant, $F(10, 420) = 0.70$, $p = .72$. The SAS code and relevant output are included in Appendix H.

Table 3.23 Study 4: Mean Report Ratings for Practitioners, Students

Guideline	Practitioners		Students		Significant Differences?		
	M	***SD***	***M***	***SD***	***df***	***F***	***p***
Overall (main effect)	**1.66**	1.35	**1.22**	1.44	1, 42	7.27	.01*
Clarity/Jargon	**2.25**	1.12	**2.22**	1.00	1, 210	0.01	.93
Impact/Severity	**1.05**	1.20	**0.38**	1.21	1, 210	7.36	.007**
Backing Data	**1.83**	1.22	**0.99**	1.44	1, 210	11.53	.0008**
Problem Cause	**2.10**	1.06	**1.78**	1.21	1, 210	1.75	.19
User Actions	**1.17**	1.52	**1.11**	1.42	1, 210	0.05	.82
Describe a Solution	**1.56**	1.50	**0.82**	1.55	1, 210	9.05	.003**

Note: guidelines are sorted from most to least *required*. * $p < .05$. ** $p < .01$.

Pearson correlations were used to test for association between the remaining indictors of a good evaluation (experience, thoroughness, validity, and hours spent on the evaluation) and the six guidelines. A conservative alpha level of .01 was used to protect against false positive due to the number of correlations computed (72). None of the correlations were significant. The SAS code and relevant output are included in Appendix H.

Two different measures of validity were calculated, using Equations 3.8 and 3.9 on page 89, with results summarized in Table 3.24. A 2x2 mixed-factor ANOVA was performed to test for effects due to type of evaluator (practitioners vs. students) and validity measure (Equation 3.8 vs. 3.9). Using an alpha level of .05, there was not a significant effect due to type of evaluator, $F(1, 42) = 1.60$, $p = .21$, but there was a significant effect due to validity measure, $F(1, 41) = 38.07$, $p < .0001$. The interaction between the two was not significant, $F(1, 41) = 0.34$, $p = .56$.

Table 3.24 Study 4: Problem Detection Validity

Validity Measure	Equation	Practitioners (n = 21)			Students (n = 23)		
		M	SD	Range	**M**	SD	Range
Severe problems detected vs. all problems detected	3.8	**.72**	.16	.47 – 1	**.67**	.15	.40 – 1
Severe problems marked as severe vs. all severe problems detected	3.9	**.88**	.15	.58 – 1	**.81**	.22	.33 – 1

Equation 3.7 on page 88 concerned a third type of validity which compared problems that could possibly occur to all problems reported. The complete list might include problems that could not occur but are instead due to evaluator misunderstanding about system. The previous section provides several examples of such problems, including believing that the system does not have a shortcut to the “credited with” box. This validity was not calculated for the current study because there were only four instances of evaluator misunderstandings out of 393 problems described.

3.4.4.2 Hypothesis 3b: Rating Reliability

In order to determine suitability of the guidelines for rating usability reports, three aspects of the reliability of the ratings by the three judges were assessed. The first aspect was bias, or the tendencies of individual raters to give higher or lower ratings. This was tested as part of the 2x6x3 ANOVA used for Hypothesis 3a with group (practitioner and student) as a between-subject factor, guideline (the five *required* guidelines from Study 3 and *Describe a Solution*), and judge (A, B, C) as within-subject factors, and evaluator as a repeated measure. Using an alpha level of .05, the main effect due to judge was significant, $F(2, 84) = 14.01$, as was the interaction between judge and guideline, $F(10, 420) = 5.47$, both with $p < .0001$. Post-hoc tests using an alpha level of .05 and Tukey adjusted *p*-values indicated that Judge B gave lower ratings than both judges A and C overall. The interaction between judge and group was not significant, $p = .20$, and the three-way interaction was not significant, $p = .72$. The interaction between judge and guideline was explored using effect slices to test for simple effects due to judge for each guideline. The tests indicated that there was a significant effect due to judge for *Backing Data*, *Clarity/Jargon, Provide a Solution*, and *User Actions.* Post-hoc tests using least-square means indicated that judge B gave lower ratings than A for *Backing Data*, and lower ratings than both A and C for *User Actions*. Ratings are shown in a summary form in Figure 3.23. The SAS code and relevant output are included in Appendix H.

The second aspect of reliability measured was association, or the tendency of pairs of raters to give higher/lower ratings to the same evaluator. This was tested using Pearson's product-moment correlation, with results in Table 3.25. Using an alpha level of .05, all correlations were significant overall and for *Backing Data, User Actions*, and *Describe a Solution*. Judge C was correlated for *Problem Cause* with both judges A and B. Correlations for *Clarity/Jargon* and *Impact/Severity* were either not significant or low ($r = .35$), indicating that the judges used different underlying traits to form their judgments.

Table 3.25 Study 4: Judge Association for Rating Guidelines

Guideline		AB	AC	BC
Overall	*r*	.39	.52	.46
	p	<.0001*	<.0001*	<.0001*
Backing Data	*r*	.46	.50	.50
	p	.00*	.00*	.00*
Clarity/Jargon	*r*	.28	.08	.20
	p	.07	.62	.20
Impact/Severity	*r*	.35	.26	.14
	p	.02*	.09	.36
Problem Cause	*r*	.20	.40	.49
	p	.20	.01*	.00*
Describe a Solution	*r*	.52	.65	.60
	p	.00*	<.0001*	<.0001*
User Actions	*r*	.52	.59	.31
	p	.00*	<.0001*	.04*

Note: The boxes surround correlations that werer not significant.
* $p < 0.05$.

The third aspect of reliability assessed was distribution, or the tendency of each judge to use each point in the rating scale. This was assessed using visual inspection of the distributions of the judges' report ratings, shown in Figure 3.23. Judge A did not use the endpoints on the scale, essentially using a 4-point scale. Judge B used all points on the scale, although distributions appeared to differ among guidelines. Judge C appeared to used the endpoints sparingly. All three appeared to have similar usage of the middle range of ratings.

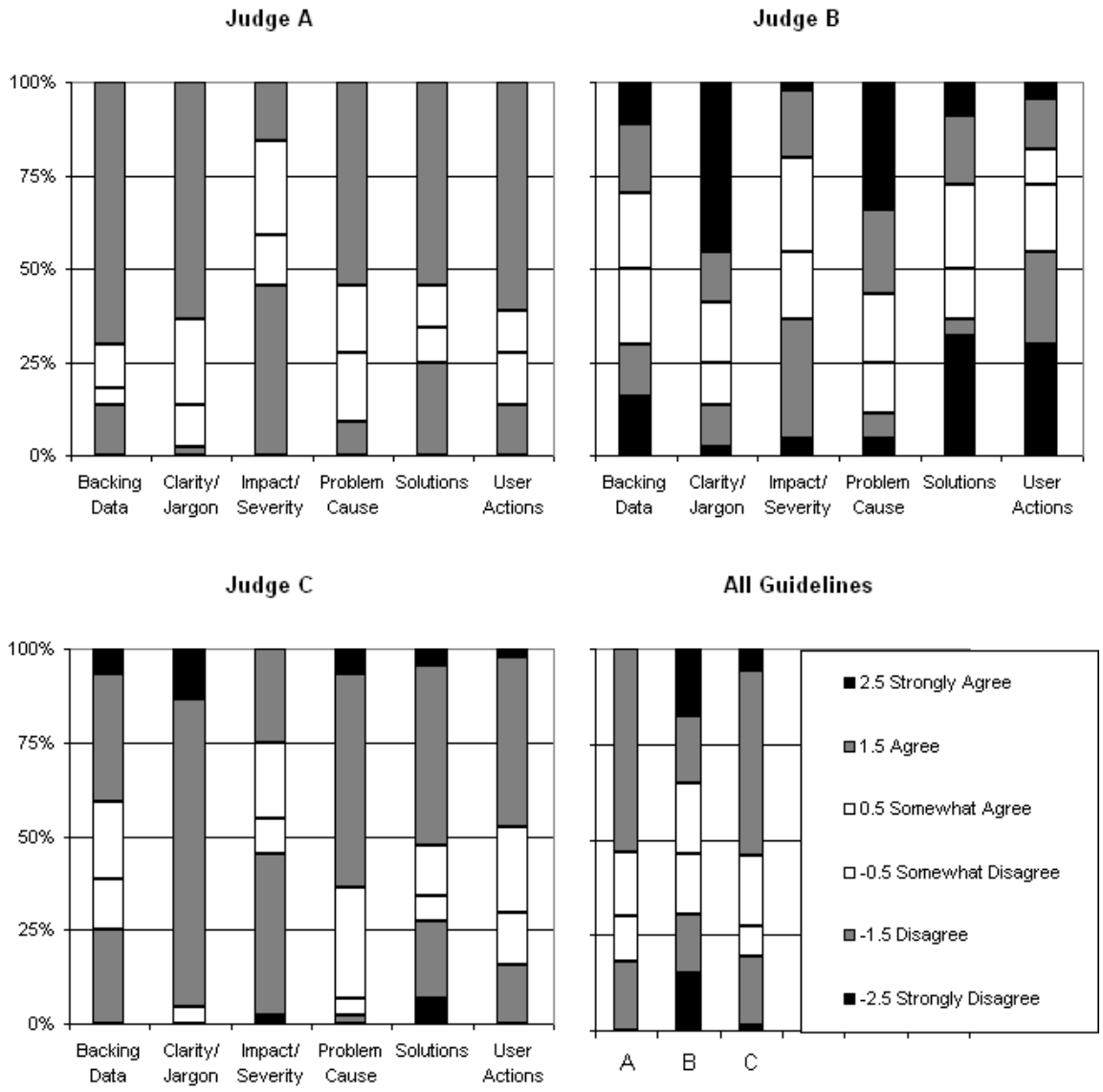


Figure 3.23 Study 4: Judge Distributions for Rating Guidelines

Table 3.26 Study 4: Summary of Judge Reliability for Rating Guidelines

Guideline	Association	Bias	Distribution
Backing Data	r = .46 - .50	B < A	
Clarity/Jargon	None	None	C rated most "agree."
Impact/Severity	r = .35 (AB only)	None	
Problem Cause	r = .40, .49 (AC, BC only)	None	
Describe a Solution	r = .52 - .65	None	
User Actions	r = .31 - .52	B < AC	
Overall	r = .39 - .52	B < AC	All three used neutral ratings similarly. A and C did not use endpoints

3.4.5 Summary of Hypothesis Testing Results

Table 3.27-Table 3.29 summarize the hypothesis testing results for Research Questions 1, 2 and 3.

Table 3.27 RQ1: How Do Practitioners Describe Usability Problems?

Hypothesis	Result
1a: Practitioners in Study 3 vs. Study 4. The practitioners in Study 4 will have the same opinions about the 10 guidelines as the practitioners in Study 3.	**Supported:** There were no significant differences between Study 3 and 4 practitioners for any of the four adjectives for nine of the guidelines: • *Clarity/Jargon* • *Impact/Severity* • *Backing Data* • *Problem Cause* • *User Actions* • *Describe a Solution* • *Professional/Scientific* • *Describe Methodology* • *Evoke Sympathy* • *Politics/Diplomacy** ***Not supported:** Study 4 practitioners gave a different rating than Study 3 practitioners for: • *Politics/Diplomacy* (lower for *difficult*)
1b: Practitioners in Study 3 vs. Students in Study 4. The students in Study 4 will have the same opinions about the 10 guidelines as the practitioners in Study 3.	**Supported:** There were no significant differences in opinion between Study 3 practitioners and Study 4 students for any of the four adjectives for nine of the ten guidelines: • *Clarity/Jargon* • *Impact/Severity* • *Backing Data* • *Problem Cause* • *User Actions* • *Describe a Solution* • *Professional/Scientific** • *Describe Methodology** • *Evoke Sympathy* • *Politics/Diplomacy** ***Not supported:** Study 4 students gave a different ratings than Study 3 practitioners for: • *Professional/Scientific* (lower for *helpful*) • *Describe Methodology* (higher for *difficult*) • *Politics/Diplomacy* (lower for *helpful*)
1c: Following the Guidelines. *T*he Study 4 practitioners will follow the five guidelines rated most *required* by the Study 3 practitioners, and the ratings for *Describe a Solution* will show a split in behavior similar to the split in opinion found in Study 3.	**Supported:** There were three or fewer negative ratings for five guidelines: • *Clarity/Jargon* • *Impact/Severity* • *Problem Cause* • *Backing Data* • *Describe a Solution* *Describe a Solution* received the fewest neutral ratings and the lowest valued negative ratings, supporting the split in behavior. **Not supported:** Six of 21 reports received a negative rating for: • *User Actions*
1d: Opinion vs. Behavior. Practitioner reporting behaviors will be consistent with their opinions, with the highest ratings for following the guidelines they believe are the most *required*.	**Supported:** There were significant correlations with opinions for all four adjectives (*difficult, helpful, relevant* and *required*) and for ratings of: • *Describe a Solution* **Not supported:** There were no significant correlations between opinions for any of the four adjectives (*difficult, helpful, relevant* and *required*) and for ratings of: • *Clarity/Jargon* • *Impact/Severity* • *Problem Cause* • *Backing Data* • *User Actions*

Table 3.28 RQ2: Is There an Evaluator Effect in Usability Testing?

Hypothesis	Result
2a: Problem Discovery Thoroughness and reliability of problem detection will be similar to heuristic evaluation, with multiple evaluators required to find the most severe usability problems.	**Supported:** Mean practitioner measurements were: • Thoroughness, all problems: .22 • Thoroughness, severe problems: .21-.34 • Reliability, all problems: .37 • Reliability, severe problems: .34-.50 Estimated means for practitioners working in pairs: • Thoroughness, all problems: .32 • Thoroughness, severe problems: .29-.45
2b: Problem Severity Practitioners will differ in their severity ratings of usability problems.	**Supported:** Mean practitioner reliability was: • Severe ratings: .57 • Minor ratings: .20

Table 3.29 RQ3: How Can We Assess the Content of UPDs?

Hypothesis	Result
3a: Good evaluators follow the guidelines. Following the guidelines will be associated with the following indicators of a better usability reports: experience of the evaluator (years of experience, number of evaluations), thoroughness of the problem set, validity of the problem set, hours spent on the evaluation, and author (practitioner vs. student).	**Supported:** Practitioners received higher ratings than students for: • Overall • *Backing Data* • *Describe a Solution* • *Impact/Severity* **Not supported:** There were no significant differences between practitioners and students for: • *Clarity/Jargon* • *Problem Cause* • *User Actions* There were no significant correlations between the guideline ratings and the remaining indicators of better evaluations.
3b: Rating reliability. The ratings of the judges will be reliable, in terms of association (considering the same underlying traits in assigning a rating) and bias (tendency to give overall higher or lower ratings).	**Supported:** There were significant but moderate correlations for: • Overall: .39-.52 • *Backing Data* .46-.50 • *Describe a Solution* .52-.65 • *User Actions*: .31-.59 • *Problem Cause*: .40, .49 for 2 pairs of judges **Not Supported:** There were no significant correlations greater than .35 for: • *Clarity/Jargon* • *Impact/Severity* Two judges used only four of the six points on the rating scale, and two judges were positively biased in their responses.

3.4.6 Interesting Observations

This section describes observations about the evaluator reports collected in this phase that fall outside the formal scope of the research questions and hypotheses.

3.4.6.1 Testing Protocol Issues

There were three issues with the testing protocol that would be important to readers of a report summarizing the evaluation results. The first issue was that the facilitator forgot to clear the auto-complete text for form text boxes between user participants. This only affected one user, and the user did not appear to see the auto-complete text, but it should be explained in the usability report. Five of the 44 evaluators noted this issue in their reports, three practitioners and two students. One student mentioned it in post-task comments to the experimenter, but not in the report.

The second testing protocol issue was an intervention by the facilitator. The third participant spent several minutes looking for the right form to make a search. The user was on the correct page, but did not scroll down far enough to find the search box (problem ba, found by 41 of 44 evaluators). The facilitator prompted the user to scroll down to the bottom of the page twice, likely a decision by the facilitator that the user had not succeeded at this portion of the task and that it was time to move on. Facilitator prompting of users is extremely important to mention because it generally indicates a task failure, critical interface flaw, or issue with the testing scenario. This was noted by 16 evaluators: 11 practitioners and 5 students. Three of these evaluators (two practitioners, one student) also noted the previous issue about clearing the auto-complete text.

The third testing protocol issue was another intervention by the facilitator. The fourth user initially searched for the Owen brothers as writers instead of actors (problem fd, found by 22 evaluators). The user went back to the search form and checked the actor boxes, but forgot to uncheck the writer boxes (problem fe, found by 12 evaluators) and was confused when the search results did not generate any movies (problem hb, found by two evaluators). At this point, the facilitator pointed out that the user had forgotten to

uncheck the writer checkboxes. This is probably a violation of testing protocol, since the user was not given time to understand the problem on her own. This incident was noted by seven evaluators, all practitioners. All seven practitioners also noted the other facilitator intervention, but only one practitioner noted all three incidents.

It should be noted that the report collected in this study was not a complete formative evaluation report. The issues discussed in this section are methodological issues, and a case could be made for including these in a report section that is separate from the descriptions of the usability problems. It is possible that some evaluators chose not to mention these issues because they would typically mention them in sections of the report that were not included in the reports collected in the current study.

3.4.6.2 Evaluator Knowledge

The task used in Study 4 was selected because most people are familiar with movies and actors, and so little additional domain knowledge was needed to understand the task. However, most evaluators were unfamiliar with the specific task tested in the study, with 38 evaluators reporting that they had performed the task in the movie two or fewer times. Had the evaluators been part of the development team, they might have had more familiarity with the website and its capabilities. This section describes three examples of ways in which limited evaluator domain knowledge was reflected in the problem descriptions.

The first example of limited evaluator domain knowledge is a small redesign of the website since the session movie was created. The links down the left side of the page with the yellow-on-blue bullets have been redesigned, as shown in Figure 3.24. In the session movie, the pale yellow bullets turn bright green to indicate the current page or location in the current page. In the current website, the green bullets have been replaced by black-on-yellow arrows. None of the evaluators commented on the poor visibility of the small green bullets and none of the evaluators commented on this redesign.

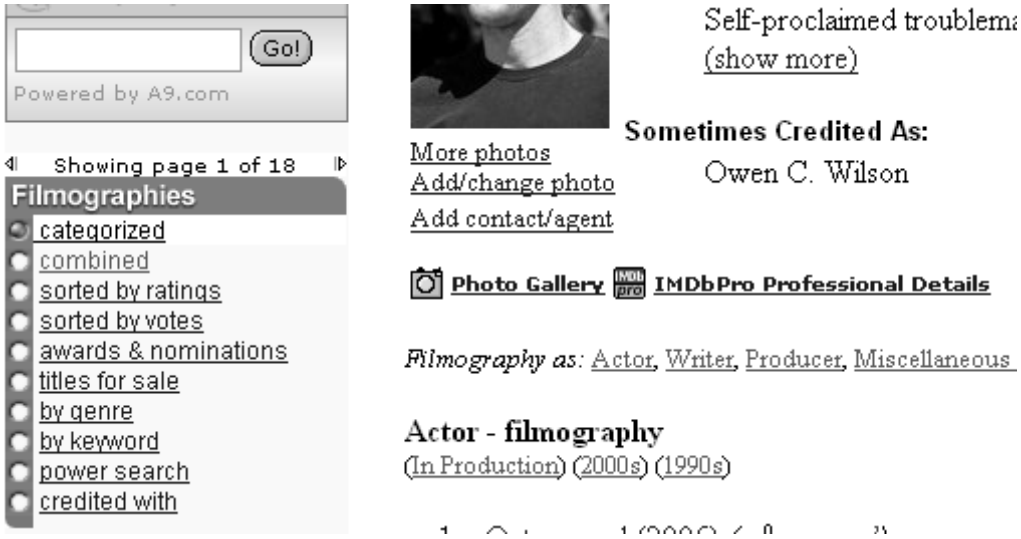
Figure 3.24 Study 4: Bullets Indicating Location, From Movie and Redesigned

The second example of limited evaluator domain knowledge is the design of the actor pages. The third user had a difficult time finding the "credited alongside" search box at the bottom of the Owen Wilson page (problem ba, found by 41 evaluators), and had to be prompted by the experimenter to scroll down to the bottom of the page. Four evaluators suggested the addition of a shortcut to this search box somewhere above the fold. Here is a sample problem description:

> A sidebar option for searching for actors that appear together was missing. Users clicked on a variety of different search options appearing on the left hand side of the screen, but none of them matched this functionality.

It is not true that a sidebar option is missing – there is a "credited with" link/shortcut at the bottom of the "Filmographies" section in the left side of the page, shown in Figure 3.25. The third user clicked on the "keywords" and "combined" links in this section (problems bd and be) but not the "credited with" link (problem bc). Only three evaluators specifically pointed out the existence of this link and its poor design. Several evaluators suggested moving the entire search box above the fold; it is unclear whether they knew the shortcut existed.

Figure 3.25 Study 4: Owen Wilson Page and "Credited With" Shortcut

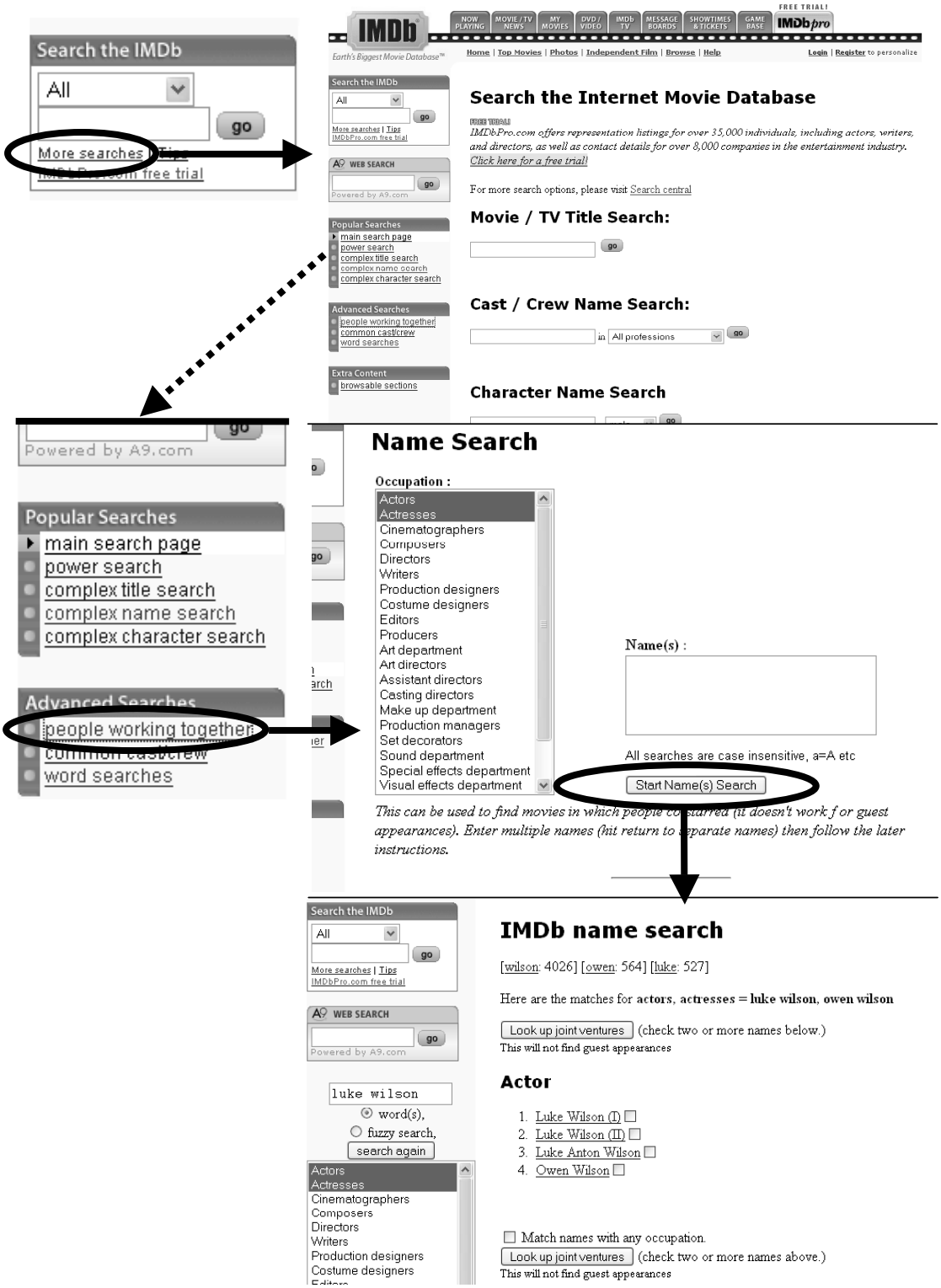


The third example of limited evaluator domain knowledge is that there is a different approach to completing the task than the approach taken by any of the four users, illustrated in Figure 3.26. All four users used the top-left search box to start their task, searching for either Owen or Luke Wilson. The alternate approach, which none of the users used, is:

- Click on "more searches" underneath the top-left search box. This brings up "Search the Internet Movie Database," a page of simple searches.
- Click on "people working together" under the "Advanced Searches" heading on the left side of the page. This brings up a "Name Search" form that is already set up to search for two actors/actresses that appeared in the same movie.
- Type "Luke Wilson<new line>Owen Wilson" into the search box and submit. This brings up the same "IMDb Name Search" page with the checkboxes that users in the session movie encountered (Figure 3.8), but with only actor roles (no writer, director or producer roles).

None of the evaluators mentioned this approach to the task. Knowing that this approach exists might affect redesign suggestions, since the form at the bottom of the actor pages is a shortcut through this alternate process. For example, several evaluators suggested changing terms used in the three pages of the credited alongside/name search process so that they are consistent throughout the entire process and more user-centric. Considering only the pages visited in the movie, this involves coordinating terminology among two web pages and one search results page. However, if the credited alongside search box is a shortcut into a more complex search, then there may be additional related pages that were not visited in the movie. A complete solution should coordinate terminology among all of these affected pages.

Figure 3.26 Study 4: Alternative Approach to Completing Task



Lack of evaluator knowledge can influence both quality of redesign suggestions and acceptance of testing results. Consider a report recipient who reads the suggestion to add a shortcut to the "credited alongside" search box above the fold, but knows that the shortcut already exists. There are several possible outcomes, from best to worst:

1. The recipient realizes that the underlying problem is a poorly designed page – too long, poorly organized, and unclear purpose of the "Filmographies" section. The recipient is skilled in design and so redesigns the page, greatly improving usability.
2. Same as (1) except that the recipient is not skilled in design and the redesign is less successful.
3. The recipient does not realize the underlying problem, and so adds an additional shortcut to the page, increasing clutter.
4. The recipient ignores the suggestion completely, since the shortcut already exists.

The recipient may also conclude that the evaluator does not understand the design of the website. In the very worst case, not only does the recipient ignore the suggestion to add the shortcut, but the evaluator looses credibility with the recipient for making suggestions of features that already exist, and the recipient becomes less receptive to other suggestions by the evaluator.

3.4.6.3 Be Professional, Be Diplomatic

Only six of the ten guidelines were used to rate the reports in Study 4, so the reports were not graded for the degree to which they followed the *Professional/Scientific* and *Politics/Diplomacy* guidelines. However, here are two examples from the same evaluator that illustrate a violation of these guidelines:

> If a search engine – even such a pathetic one as this – doesn't recognize or obey structured queries it should state the fact when the user enters them, rather than uttering garbled responses in the hope that the right answer might be there.

> State the query generated by the search box so that the user can either improve it or gasp at the search engine's deficiencies and flee the site giggling.

The tone of the problem descriptions is informal and very critical of the site. Dumas, Molich and Jeffries (2004) have a suggestion for this type of problem description — express your annoyance tactfully. They point out that emotionally charged statements like these can create tension between evaluators and developers, decreasing developers' receptivity to evaluators' suggestions. Indeed, the practitioner who wrote these descriptions commented in the post-task questionnaire that some of the problem descriptions were " 'first-draft' phrases which I would tone down if the report was going to anyone but a team mate - and I hope you enjoyed them." This practitioner received the highest overall score from the judges for the six guidelines that were scored, and a very high score for finding severe problems. The practitioner also included an excellent executive summary with a synthesis of the problems, explanation of the overall issues with the site, summary of most important problems, and suggestions for the most important changes. While these examples do illustrate a point about the tone of UPDs, it is highly unlikely that these are examples of the practitioner's typical work product.

3.4.6.4 Describing User Actions

User Actions had the lowest mean rating of the six guidelines rated in this study, with the exception of *Impact/Severity* (which had a low mean because evaluators used severity codes instead of describing severity). *User Actions* also had the largest number of negative ratings, larger than *Impact/Severity*. Describing user actions helps explain exactly what happened, and emphasizes the user component of the evaluation, as opposed to expert opinion. Table 3.30 provides five examples of descriptions of problem ab: the user did not realize that the top-left search box did not support "and" searches or exact string searches using double quotes (" "). Three examples were taken from reports with high ratings for *User Actions*, and two from reports with low ratings. Each description has good elements and missing elements, but none of them contains complete information. These are just five examples from the 40 evaluators that reported this problem.

Table 3.30 Study 4: Examples of Describing User Actions for Problem "ab"

Rating	UPD	Features
High	This search entered in top left search box "Owen Wilson" + "Luke Wilson" Gave a search results page that says A search for ""Owen Wilson" + "Luke Wilson"" It's not clear that the search was intended to have external quotes appended for an "Exact Phrase" search. The original search seems to intend and [sic] "And" search with the "+" sign.	✓ Sample text from search ✓ Explains search box used ✓ User/system mismatch • No impact on user
High	User entered compound expressions using a "+" the System the "+" and the fact that the items were in quotes and instead did a "contains" search on each word , effectively an or search. Ignoring the quotes s/b a bug. 	✓ Screen shot of text ✓ Screen shot of search box ✓ User/system mismatch • No impact on user
High	1. *Problem definition*: Poor support for standard search syntax, such as AND, plus sign and double quotation marks. 2. *Problem category*: Implementation/Code 3. *Problem location*: Global. 4. *Why this is a problem*: Experienced Internet users utilise standard search syntax for efficient search. Not supporting conventional syntax may significantly decrease: • Efficiency (users perform task slower than they expect and are able to); • Effectiveness (users are confused and disoriented); • Satisfaction (losses in efficiency and effectiveness may lead to frustration). 5. Examples from evaluation • Clips #3 and #4: notice how confident the user #3 is at the start of his search and how he gets thrown off track after being let down by unsupported search syntax. *Solution recommendations*: enable the full range of standard search syntax.	✓ Why user tried this search ✓ Impact on user ✓ Examples of impact on user ✓ Generic description of search string • Which search box? • No example search string
Low	*2 of 4 Typical Operators Do Not Work* Quotations, +, and "and" did not work to yield results including both names from the search box on the main page. Such functionality is typical of major search engines (e.g., Google). This functionality should be included.	✓ Explanation of mismatch ✓ Why user tried this search • Which search box? • No example search string • No impact on user
Low	Advanced web and computer users expected a search involving "and" to work according to the technical/logical meaning of "and". That is, they expected a search on "x and y" to produce only those instances where both x and y occurred together. However, the search engine gives users what appear to be hits on some random collection of items containing Wilson.	✓ Generic search string ✓ Why user tried this search ✓ Expected and actual results • No specific search string • Which search box? • No impact on user • Vague use of "appear" (see Section 3.4.6.6)

3.4.6.5 Use of Positive Findings

Evaluators sometimes made indirect references to problems in comments that were marked as positive findings. For example, "some participants clicked on the link that is the actor's name (see graphic above) to confirm that Luke Wilson (I) or (II) was the Luke they wanted." This mentions problem gc, that the user does not understand the difference between Luke Wilson I and II. However, the category given to this problem was "positive finding," and problem gc was never mentioned elsewhere by this evaluator. Explicitly mentioning the issue about distinguishing two actors with the same name (problems ga, gb and gc) would increase the likelihood of fixing the problem. Six evaluators mentioned seven problems in comments marked "positive finding." Four of these problems were also mentioned elsewhere in a problem with a *minor/severe/critical* severity rating, three were not.

3.4.6.6 Vague Statements

The judges were instructed to match problem descriptions to master list problems if they could determine which problem was being discussed based on their knowledge. For example, there was only one pull-down menu in the movie, so any description that mentioned a pull-down menu must be talking about the search box in the top-left corner of every page. Similarly, only one page in the movie had checkboxes. Any problem that two of the three judges and the experimenter could not understand was marked "vague." This is a high threshold for vagueness, since the judges and experimenter spent dozens of hours reading hundreds of problem descriptions and became extremely familiar with the movie. In many cases, the judges were able to identify problems that most people would not have understood. There were twenty vague comments from fifteen evaluators. Table 3.31 includes several examples of complete problem descriptions (these are not quotes from longer problem descriptions).

Table 3.31 Study 4: Examples of Vague Problem Descriptions

Problem Description	Issue
This participant also goes directly to the search box only because they remember seeing it.	Which search box? The one in the top-left corner or the one at the bottom of Owen Wilson's page?
Participant clicked on the Luke Wilson (I) and Owen Wilson on the Actor check boxes. Clicked on the search box and got 4 links with only one movie.	This describes the user's actions, but what exactly is the problem?
Provide a "back" button on the error page when no matches are found.	Which error page?
Scrolling was needed to find what was on screen.	Which screen?
Tried clicking on individual movie links from Owen Wilson's page, and did not see any way this would get him the information required.	This describes the user's actions, but what exactly is the problem?

The use of the word "appeared" was problematic in five problem descriptions from three evaluators. For example, "Boolean 'and' search appears to be possible off the home page," and "there appear to be two ways to select an actor - a link and a checkbox." Both of these sentences lack an agent, and so it is unclear to whom the site appears this way. Is it the user or the evaluator that believes the site supports Boolean searches? If it is the user that believes the site supports Boolean searches, it is important to say so explicitly and be clear that the user holds incorrect beliefs about website functionality. If is it the evaluator, then the evaluator is incorrect. The implications of this interpretation are described in Section 3.4.6.2 Evaluator Knowledge. The judges marked these descriptions as examples of evaluators misunderstandings about the website, but the experimenter later decided that the sentence construction was unclear.

3.5 Discussion

3.5.1 RQ1: How Do Practitioners Describe Usability Problems?

The first area of exploration for Phase II was the description of usability problems, in terms of practitioner beliefs about how problems should be described, and observations about the information that practitioners included in the problem descriptions collected in Study 4. Hypothesis 1a was that a re-sampling of usability practitioners

would result in similar opinions about the 10 guidelines as was found in Study 3. This hypothesis was supported, with no significant differences in opinion about any of the guidelines for any of the four adjectives. The one exception was *Politics/Diplomacy*, where Study 4 practitioners gave a lower rating than Study 3 practitioners did. The implication is that opinion about the guidelines is stable within the communities sampled in the study, increasing confidence in the comparison of guidelines rated *required* and *difficult* in Study 3 (Figure 2.8, page 34).

In Study 3, there were no significant effects on opinion about the guidelines due to level of evaluator experience. However, there may have been a floor effect, since there were minimum experience criteria for the respondents. The inclusion of graduate students in Study 4 allowed comparison to evaluators that had less experience than those in Study 3. Hypothesis 1b was that there would be no differences in opinion between students in Study 4 and practitioners in Study 3, as there were no differences due to experience among practitioners in Study 3. This hypothesis was supported, with no significant differences for the six guidelines used to grade reports in Study 4 for any of the four adjectives, and no differences for any of the 10 guidelines for *required* or *relevant*. This suggests that either opinion about the importance of the various elements within a UPD are stable and do not change as evaluators gain experience, or that other factors such as individual differences in opinion are greater than any effects due to experience.

Hypothesis 1c had two parts. The first part of the hypothesis was that Study 4 practitioners would follow the five guidelines rated most *required* by the Study 3 practitioners. This was not supported for *User Actions*, with six of 21 practitioner reports receiving a negative rating. Practitioners received positive ratings for four of the five guidelines (*Clarity/Jargon*, *Impact/Severity*, *Problem Cause*, and *Backing Data*). However, care needs to be taken in concluding that this implies that the practitioners followed the guidelines. Two of the judges were positively biased, giving positive (as opposed to neutral or negative) ratings to over half of the reports. In addition, judge ratings for *Clarity/Jargon* and *Impact/Severity* were, for the most part, uncorrelated, meaning that each judge based their ratings on different underlying traits. This unreliability among judges makes it difficult to draw conclusions about the ratings for

Clarity/Jargon and *Impact/Severity*. The second part of hypothesis 1c was that behavior for *Describe a Solution* would be split. This would be consistent with the split in opinion found in Study 3 about how *required* and *helpful* the guideline is. This hypothesis was supported, with reports receiving the fewest neutral ratings for following this guideline and the lowest valued negative ratings.

The five most *required* guidelines were based on the highest mean ratings from the Study 3 practitioners. However, individual opinion about the guidelines varied. Hypothesis 1d was that individual opinion about the guidelines would affect reporting behaviors, with practitioners that have the strongest belief about how *required* a particular guideline is receiving the higher report ratings for that guideline. This hypothesis was not supported for the five *required* guidelines used to rate reports (*Problem Cause*, *Clarity/Jargon*, *Backing Data*, *Impact/Severity* and *User Actions*), with no significant correlations between opinion and behavior. This suggests that the inclusion or exclusion of certain description elements is due to something other than opinion. In particular, practitioners received low ratings for *User Actions*, despite its high ratings for being *required*, and low ratings for *difficulty*. Further studies are needed to determine if lack of descriptions of user actions is a flaw in usability reports, or if the importance of describing user actions is over-rated by practitioners. Even experts are not always able to articulate what they do, and their explanations of their actions may not match their behavior (Speelman, 1998).

In contrast, Hypothesis 1d was supported for *Describe a Solution*. There were significant though moderate correlations ($|r| = .51$-$.67$) between report ratings for *Describe a Solution* and opinions about *Describe a Solution* for the four adjectives, with a negative correlation for *difficult* and positive for *helpful*, *relevant*, and *required*. This suggests that a solution is something evaluators choose to include or not to include with their UPDs. The post-task questionnaire for Study 4 included an open comment box that asked evaluators whether or not they include solutions with their reports and why. Figure 3.27 summarizes the comments collected from Studies 1, 3 and 4, with more than half of the comments from Study 4 practitioners.

Reasons to include a solution (frequency count):

- The client/team that asked for the evaluation also needs help designing solutions. (7)
- The report is more acceptable when you offer not just criticism but help in the form of suggestions or recommendations. (5)
- You are the one with the design skills, so you should suggest design improvements. (4)
- You can provide some general suggestions or sketches to start the redesign process and foster discussion. (4)
- A solution can help the reader understand the problem. (4)
- Solve trivial problems quickly. (2)
- Development teams want specific, actionable suggestions. (1)
- Knowing the solution will aid fix/leave decisions. (1)
- The reader may not see a solution that seems obvious to you. (1)
- The purpose of a formative evaluation is to suggest solutions to problems. (1)

Reasons not to include a solution (frequency count):

- Wait until you can discuss solutions with the larger team, brainstorm ideas. (5)
- You need to step back and look at the whole picture, not just solve individual problems (3)
- There may be other team members who have better skills for designing a solution, or who know the system better (programmers, graphic design, etc.) (3)
- Only suggest solutions that you are sure of. (2)
- Many solutions are complex and require further research. You do not want to jump to conclusions, and you do not want to delay release of the report until you have time to design quality solutions. (2)
- Recommending a poor solution is worse than recommending no solution. (2)
- Do not suggest a solution when do you not have enough data to support the suggestion. (1)

Figure 3.27 Studies 1, 2 and 4: Reasons to include or not include a solution

The most frequent reason given for including a solution was that the evaluation team is the primary source of usability or design knowledge, such as when the evaluation team is brought in as consultants or the evaluator is the only usability specialist in the organization. Solutions can also make reception of the report more favorable, increasing the likelihood that problems will be fixed. This was noted in a study by Hornbæk & Frøkjær (2005), who found that developers appreciate usability reports that include redesign suggestions more than reports with just problem descriptions. The developers in their study rated problems more severe when described in terms of a redesign than a problem, and commented that they were better able to understand the problem when phrased in terms of a redesign.

The primary reason given for not including solutions with problem descriptions was that you should wait until the entire team can work on solutions together. This allows all stakeholders to provide input and draws on the strength of the group as a whole in designing solutions. Jeffries (1994) criticizes usability reports for suggesting point solutions to single problems without considering trade-offs, optimizing the solution for a particular task or problem without considering its impact on other tasks. This is less likely to happen when developing solutions is postponed until all of the problems can be considered together and the entire team is present. *Describe a Solution* was also rated among the most *difficult* guidelines to follow, which could contribute to a preference for waiting until a solution can be developed properly rather than jumping to conclusions. A moderate approach to providing solutions is to offer general suggestions or alternative approaches, but not specific solutions. These suggestions can start a discussion among the rest of the product team, or provide guidance for designers or programmers in creating solutions that are more detailed.

Outside of the formal research hypotheses, several additional observations were made about information included or not included in the reports. A few evaluators mentioned testing protocol issues such as facilitator interventions, but most did not. Some descriptions were so vague that it was difficult to understand their problem descriptions. The number of images or tables in the report varied from zero to 16, and average words per comment varied from 17 to over 200. Positive (complementary) comments included with the report varied from none to 56. Most of the problem descriptions were free form, but a few evaluators used a standard format for each description, with specific elements such as a title, problem cause, or recommended solution. Some of the reports read like a running commentary of the users' actions with little analysis or insight, others provided detailed analyses of issues such as problem cause and impact on the user.

The diversity of reporting styles and report elements may be a reflection of the diversity of evaluation settings. Recipients of usability reports can vary widely, and may be teams unfamiliar with usability, new consulting clients, teams with an established relationship to the evaluators, executive decision makers, or other usability professionals (Theofanos & Quesenbery, 2005; Theofanos et al., 2005). Companies, departments, and

even individual projects vary in work culture, resources, development timeframe, etc. Many practitioners commented that they tend to tailor their reports based on the particular project, and do not use the same format for every evaluation. Further research is needed to understand how the importance of following the various guidelines changes in different evaluation settings, as well as the importance of other report elements, such as screen shots.

3.5.2 RQ2: Is There an Evaluator Effect in Usability Testing

The second area of exploration for Phase II was the evaluator effect in usability testing. Hertzum and Jacobsen define the evaluator effect as differences in problem detection and problem severity judgments by evaluators using the same UEM (Hertzum & Jacobsen, 2003; Jacobsen et al., 1998). Study 4 was designed to test for the presence of an evaluator effect in usability testing, even when using pre-recorded usability testing sessions to control the test setting. Three measures were taken: thoroughness of problem detection, inter-evaluator reliability of problem detection, and inter-evaluator reliability of problem severity judgments. Three different methods for measuring problem detection were used: counting any problem identified ("all problems"), counting only problems that the judges rated as severe ("severe problems"), and counting only problems that both the judges and the evaluator rated severe ("severe marked severe").

Overall, thoroughness was low, with practitioners having a mean detection rate of .22 (*SD* = .08), meaning that each evaluator found, on average, less than a quarter of the total problems in the movie. Thoroughness for severe problems was higher, with practitioners having a mean detection rate of .34 (*SD* = .12), indicating that practitioners are more likely to detect severe problems than minor problems. A similar pattern was found for reliability of problem detection, or overlap in problems reported by different evaluators. Reliability for practitioners was low (*M* = .37, *SD* = .12) for all problems, and higher for severe problems (*M* = .50, *SD* = .17). The higher reliability for severe problems reflects the smaller number of severe problems (18 vs. 41 overall), and that the minor problems were more likely to be found by fewer practitioners. All of the problems found by at least 10 of the 21 practitioners were severe problems, and 15 of the 18

problems found by one to five practitioners were minor problems. Problems that are not found at all lower thoroughness, but have no effect on reliability scores.

There was not a significant difference between all problems and severe problems marked as severe for thoroughness and reliability of problem detection, although both were significantly lower than thoroughness and reliability for severe problems. The "severe marked severe" measurement relies on both problem detection and severity judgments, which were found to have low reliability. Any-two agreement in severity judgments by practitioners for severe problems was .57 (*SD* = .28). Another indicator of low reliability is that there were 12 severe problems reported by at least 10 evaluators, and for nine of these problems the balance between minor and severe ratings was not significantly different from an equal split.

There were no significant differences between students and practitioners for thoroughness (all problems and severe problems), and no significant differences for reliability across all problems. There were significant differences for reliability of severe problems and severe problems marked severe. The lower reliability for students may be due to the four severe problems that were found by only one student each, and no practitioners. Problems not found at all do not affect reliability, whereas problems found by only one evaluator will lower reliability. Since there were no significant differences between practitioners and students in thoroughness for severe problems, practitioners have higher reliability than student because as a group they found a smaller set of problems, and so are more likely to find the same problems. There were no severe problems found by practitioners that were not found by students.

The results of Study 4 suggest that there is a component of the evaluator effect in usability testing due to the evaluator, as opposed to study design or task selection. This effect was noted for both problem detection and problem severity judgments. While the current study involved evaluators working individually, evaluators frequently work together, particularly in pairs. The discussion of RQ2 continues by extrapolating the results from the individual evaluators from Phase II to groups of evaluators, comparing

the current study to previous studies of the evaluator effect, and discusses possible reasons for an evaluator effect in usability testing.

3.5.2.1 Problem Detection by Evaluators Working in Groups

Problem detection by groups of evaluators working together was simulated using a bootstrap technique. A group of evaluators was selected by randomly picking, with replacement, individual evaluators from the current study. In bootstrap techniques, picking with replacement is used when the sample was a subset of the population; picking without replacement is used when the entire population was measured. The individual evaluators' problem sets were aggregated to simulate an evaluation by a group. The number of problems found by the entire group was calculated using the three interpretations of "found" – all problems, severe problems, and severe problems marked as severe. This process was repeated 20,000 times for each size group (1-20), and each type of evaluator (practitioners, students, and all evaluators). Table 3.32 shows summary statistics for aggregates of up to five practitioner evaluators. Figure 3.28 illustrates the mean problems found as a percentage of all possible problems for the measure (41 problems overall, and 18 severe problems).

Table 3.32 Study 4: Problem Discovery by Aggregates of Practitioners

	All Problems			Severe Problems			Severe Problems Marked Severe		
Simulated Group Size	**M**	**SD**	**%ΔM**	**M**	**SD**	**%ΔM**	**M**	**SD**	**%ΔM**
1	**22%**	8%		**34%**	9%		**22%**	10%	
2	**32%**	10%	43%	**45%**	9%	31%	**29%**	11%	27%
3	**38%**	10%	19%	**50%**	8%	12%	**32%**	11%	10%
4	**42%**	10%	11%	**54%**	8%	6%	**33%**	11%	5%
5	**45%**	10%	8%	**56%**	8%	5%	**35%**	11%	4%

An evaluator designing a usability study similar to the current study might want to be confident in finding a reasonable portion of the usability problems present in the interface. Table 3.33 lists the percent of problems detected by groups of one, two and

three practitioners for the 95% confidence level. Figure 3.29 illustrates the 90% confidence interval around the mean number of problems found by groups of one to four practitioner evaluators, with the bottom of each vertical line representing a 95% confidence in finding at least that many problems. Having a 95% confidence in finding at least 50% of severe problems would require seven practitioners. Having a 95% confidence in finding at least 49% of all problems would require 17 practitioners. The maximum percentage of problems found by a pair of evaluators (not necessarily the same pair of evaluators) was 61% of all problems, 78% of severe problems, and 56% of severe problems marked severe.

Table 3.33 Study 4: Practitioner Problem Detection, 95% Confidence Level

Number of Practitioners	Type of Problem		
	All Problems	Severe	Severe Marked Severe
1	10%	17%	6%
2	17%	28%	11%
3	24%	33%	11%

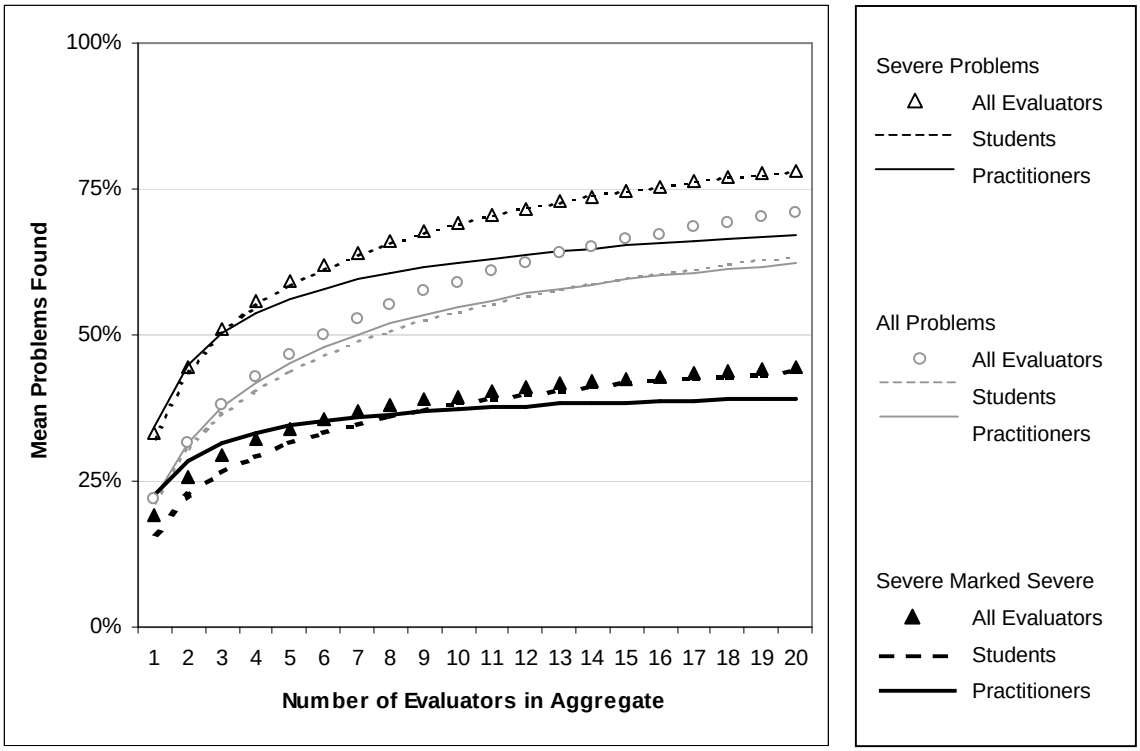


Figure 3.28 Study 4: Problem Discovery by Simulated Groups of Practitioners

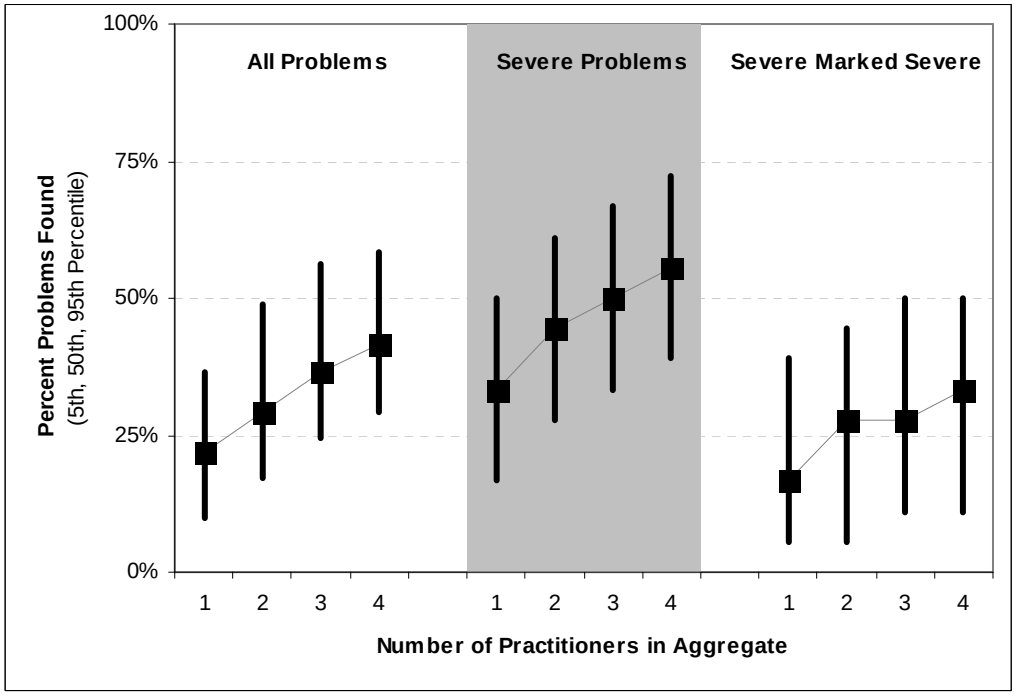


Figure 3.29 Study 4: 90% Confidence Intervals for Simulated Group Discovery

Table 3.34 Study 4: Comparison of Thoroughness, Reliability Compared to Previous Studies of UT, HE

Study	UEM	#, type of evaluators		# Probs.		Thoroughn.		Rel.
				All	Sev.	All	Sev.	All
Current Study	UT(Y)	21 practitioners		41	18	.22	.34	.37
		23 graduate students		41	18	.22	.33	.37
Long, Styles, Andre & Malcolm (2005)	UT(Y)	12 usability students		11	—	.34[a]	—	—
same movie as the current study		12 usability students		11[a]	—	.33[a]	—	—
Skov & Stage (2005)	UT(Y)	14 undergraduate CS students		32	17	.27[a]	.34[a]	—
Jacobsen, Hertzum & John (1998)	UT(Y)	2 experienced evaluators 2 beginner evaluators		93	37	.52	.72[b]	.42[b]
Vermeeren, Kesteren & Bekker (2003)	UT(Y)	2 unspecified individuals, Study 1		30	—	.78[a]	—	.59[a]
		2 unspecified individuals, Study 2		33	—	.82[a]	—	.64
		2 unspecified individuals, Study 3		93	—	.87[a]	—	.75[a]
Molich et al. (1998): CUE-1	UT(N)	3 professional teams[c]		146	—	.36[a]	—	.06[a]
Molich et al. (2004): CUE-2	UT(N)	6 industry/university teams[c]		186	—	.22[b]	.43[b]	.07[b]
Kessner, Wood, Dillon & West (2001)	UT(N)	6 professional teams, same system as CUE-2		36	18	.36[a]	.43[a]	—
Hornbæk & Frøkjær (2004b)	HE	43 1st-year CS undergraduates		341	—	.03[a]	—	.07
	MOT	44 1st-year CS undergraduates		341	—	.03[a]	—	.09
Molich & Nielsen (1990),	HE	77 readers of Computerworld	(Mantel)	30	—	.38	.44	.45[b]
Nielsen & Molich (1990),		37 CS students in UI course	(Teledata)	52	—	.51	.49	—
Nielsen (1992),		34 CS students in UI course	(Savings)	48	—	.26	.32	.26[b]
Nielsen (1994b)		34 CS students in UI course	(Transport)	34	—	.20	.32	—
		31 CS students, no training	(Banking)	16	8	.22	.29	—
		19 usability specialists	(Banking)	16	8	.41	.46	.33[b]
		14 usability & IVR "double experts"	(Banking)	16	8	.60	.61	—

Note: This table only includes UT and HE studies for which thoroughness (mean problems per evaluator or mean evaluators per problem) or reliability (any-two agreement only) is available, and for HE only studies with at least 30 evaluators. A dash (—) indicates that the measure could not be calculated from published data. UT(Y/N) = usability testing (same task? Yes/No). HE = Heuristic Evaluation. MOT = Metaphors of human thinking. Insp. = Inspection. [a]Calculated based on data published in the article. [b]As reported by Hertzum & Jacobsen (2003). [c]Other teams participated, but only these teams used UT exclusively.

3.5.2.2 Comparisons of Problem Detection to Previous Studies

The movie for the current study is the same movie used by Long, Styles, Andre & Malcolm (2005), except that the version for this study had approximately one minute cut from a very lost participant. Thoroughness in the current study (.22) was lower than thoroughness in Long et al. (.33). However, the current study based its thoroughness calculations on a set of 41 problems, as opposed to the 11 used by Long et al. The larger set in the current study could be due to finer-grained interpretations of individual problems, or it could be that the students in Long et al. found only a small subset of the problems reported by the graduate students and practitioners in the current study. Either explanation would account for the lower thoroughness in the current study. The higher thoroughness in Long et al. could also be due to the homogeneity of the student population in that study, drawn from a single advanced human factors course. This seems an unlikely explanation in light of the lack of significant differences in thoroughness and reliability between students and practitioners in the current study (across all problems only – the practitioners were more reliable than students in finding severe problems). Most of the student evaluators in the current study had taken Usability Engineering with the same professor while the practitioners came from many different companies, resulting in more heterogeneity in the practitioner group than the student group.

The three other studies that used pre-recorded sessions and reported thoroughness and reliability (or that reported enough data to calculate these measures) are Skov & Stage (2005), Jacobsen, Hertzum & John (1998) and Vermeeren, Kesteren & Bekker (2003). Skov & Stage had similar results with 14 evaluators for severe problems (.34) and slightly higher thoroughness across all problems (.27) than the current study. The other studies had much higher thoroughness, from .52 to .87, and higher reliability, from .42 to .75, even for two studies with over 90 problems in the complete problem set.

Given the low thoroughness generally measured in comparison studies, a problem set created as the union of a small number of evaluators is unlikely to contain all of the problems present in the interface. For example, Skov & Stage compared their 14 undergraduate evaluators to two experienced usability experts, and found that the

usability experts detected 10 problems that the undergraduates did not and that the undergraduates detected four problems that the experts did not, with an overlap of 28 problems. The problem set used in the current study should be a large subset of a theoretically complete set of problems in the usability session movie, since it was created based on reports by 58 evaluators. Studies with many fewer evaluators cannot make this same claim.

A problem set created by a small number of evaluators is likely to be smaller than a set created by a larger number of evaluators, and so an estimate of individual thoroughness in such a study is likely to be higher than an estimate from a study that used a larger group to create the problem set. Thoroughness is the ratio of individual problems found to the entire set of problems. In larger studies, the numerator (individual problem detection) stays the same, but the denominator (number of problems in the master list) grows. In the bootstrap simulation of thoroughness by groups of evaluators in Section 3.4.3.1, an aggregate of two evaluators found, on average, 32% of the problems found across all the evaluators and pilot evaluators, and four evaluators found 42% of the total problems. Had the current study used fewer evaluators and based the complete problem set on that smaller group, estimates of individual thoroughness would have been higher than recorded in the current study.

A second bootstrap simulation was performed to illustrate the degree of over-estimation of thoroughness that would have occurred by using smaller sample sizes and ad-hoc problem sets based on the union of problems found by the evaluators in the simulated study. Since the goal of this simulation was to predict what would have happened with the same evaluators that participated in the current study, groups of evaluators were created using sampling without replacement. Figure 3.30 illustrates the aggregate thoroughness of groups of individual evaluators for severe problems and across all problems, for group sizes from 2 to 44.

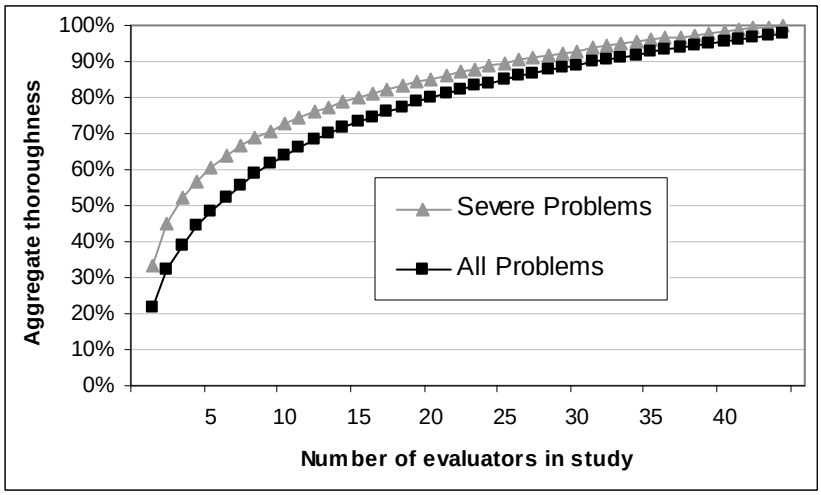


Figure 3.30 Study 4: Simulation of aggregate thoroughness vs group size

Figure 3.31 illustrates the thoroughness calculations based on the simulated groups, with drop lines indicating the difference between the thoroughness based on the aggregate problems from the sample group and thoroughness calculated against the complete master problem list used in the current study. These estimates of measured thoroughness at a smaller sample size bring the results of the current study closer to the thoroughness values for previous studies of usability testing reported in Jacobsen, Hertzum & John (1998) and Vermeeren, Kesteren & Bekker (2003), as shown in Figure 3.32 and in Table 3.35.

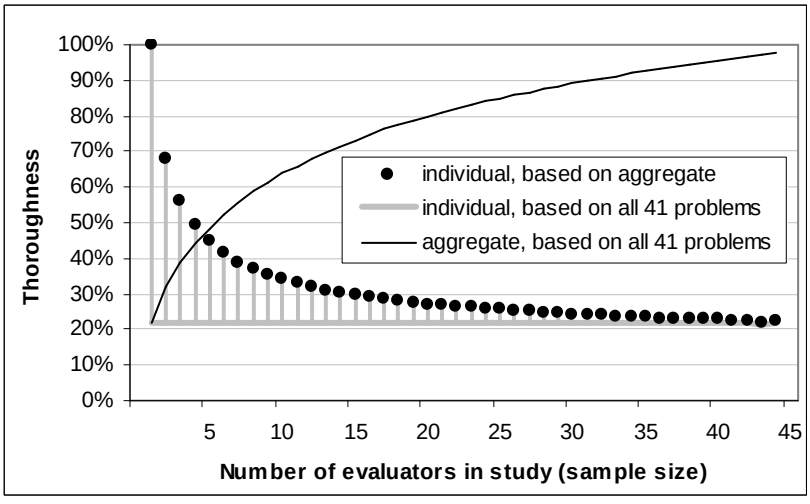


Figure 3.31 Study 4: Simulation of thoroughness over-estimation due to sample size

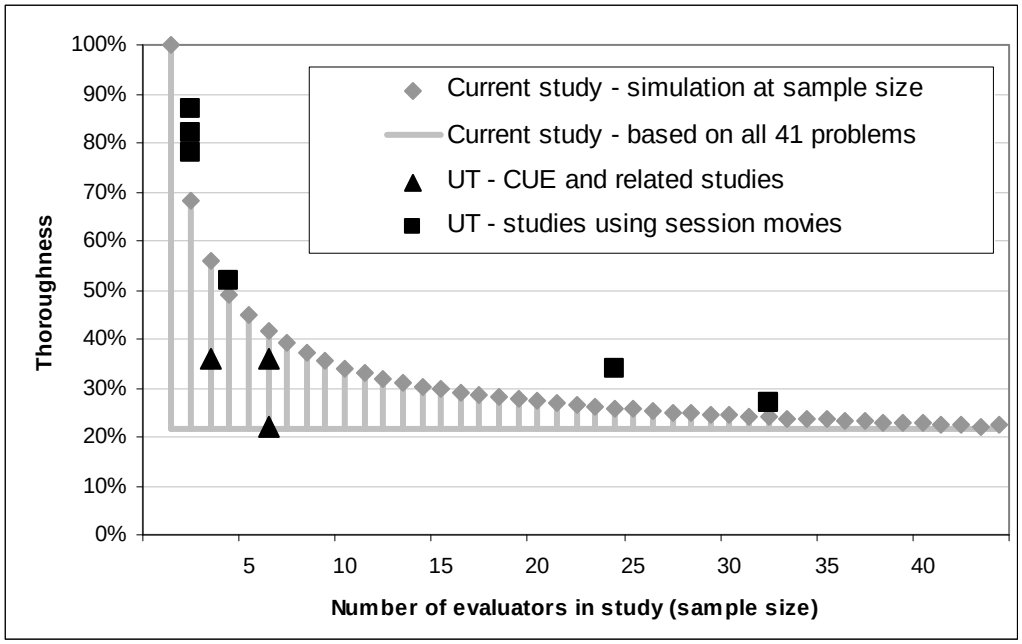


Figure 3.32 Study 4: Sample Size-Corrected Thoroughness Compared to UT Studies

Table 3.35 Study 4: Sample Size-Corrected Thoroughness Compared to UT Studies

Study	UEM	Number of Evaluators (sample size)	Reported mean thoroughness (all problems)	Bootstrap estimate of current study thor. at sample size
Current Study	UT(Y)	44	.22	.22
Long, Styles, Andre & Malcolm (2005)	UT(Y)	24	.34	.26
Skov & Stage (2005)	UT(Y)	32	.27	.24
Jacobsen, Hertzum & John (1998)	UT(Y)	4	.52	.49
Vermeeren, Kesteren & Bekker (2003): 1	UT(Y)	2	.78	.68
Vermeeren, Kesteren & Bekker (2003): 2	UT(Y)	2	.82	.68
Vermeeren, Kesteren & Bekker (2003): 3	UT(Y)	2	.87	.68
Molich et al. (1998): CUE-1	UT(N)	3	.36	.56
Molich et al. (2004): CUE-2	UT(N)	6	.22	.41
Kessner, Wood, Dillon & West (2001)	UT(N)	6	.36	.41

Note: UT(Y/N) = usability testing (same task? Yes/No).

The CUE-like studies of usability testing (Kessner et al., 2001; Molich et al., 1998; Molich et al., 2004) used a different approach to designing their study than was used in the current study. The systems were large, tasks and users varied, and evaluators' hours on the study ranged from 10-200, rather than 0.5 – 6.5. The goal of these studies was to compare evaluations conducted in real-world settings (high external validity), as opposed to the current study, which favored tighter experimental control (high internal validity). Despite the variability in tasks, users, testing approach, etc., thoroughness in CUE-1 and Kessner et al. was higher (.36) than the current study (.22), and CUE-2 was similar (.22). However, when using the same process for correcting the thoroughness for number of evaluators in the study, the thoroughness numbers in the CUE-like studies are lower than the current study (corrected value: .41-.56).

Evaluator reliability in CUE-1 and CUE-2 (.06, .07) was lower than the current study (.37). Reliability in the studies that also used recorded sessions was similar to or higher than reliability in the current study. Reliability in Jacobsen, Hertzum & John (1998) was slightly higher (.42) than in the current study, and reliability in Vermeeren, Kesteren & Bekker (2003) was much higher (.59 – .75). Reliability calculations are not subject to the same estimation issues with small sample size as thoroughness is, since any-two agreement is an average across pairs of evaluators and not truly a measure of the entire group (this was verified using bootstrap simulations of smaller size groups).

Reliability of problem severity judgments was .57 for severe problems. This is higher than for the three studies summarized by Hertzum and Jacobsen (2003), which had agreements ranging from .20-.28. These studies asked all to evaluators rate the severity of a set of problems. In the current study, evaluators only rated the severity of problems that they themselves found, which was, on average, 22% of all the problems. Lesaigle and Biers (2000) collated the problems found by their fifteen usability professionals and then had them rate the severity of all the problems. They found the coefficient of concordance to be .18, and that four evaluators gave significantly higher ratings to the problems they had detected than they gave to problems they had not detected. It is possible that having all evaluators rate the severity of all problems would result in lower reliability than was measured in the current study.

In summary, the setup in the current study (short pre-recorded movie, small target system) increased thoroughness as compared to the CUE and related studies, and greatly increased evaluator reliability compared to the CUE studies. However, neither thoroughness nor reliability could be characterized as "high," and levels were comparable to heuristic evaluation. Therefore, there is an evaluator effect, even when using a session recording to control the test setting. Why is there an evaluator effect? The next section uses the similarity of usability diagnosis to medical diagnosis to provide one possible explanation for the evaluator effect observed in Study 4.

3.5.2.3 The Evaluator Effect: Unreliability in Expert Judgment

The usability community has a tendency to focus on the user in usability testing. There is ongoing debate about the number of users to include in usability testing (Faulkner, 2003; Law & Hvannberg, 2004a; Smilowitz, Darnell, & Benson, 1994; Spool & Schroeder, 2001; Virzi, 1990, 1992; Woolrych & Cockton, 2001). Several UEM comparison studies equate number of users in usability testing to number of experts in expert evaluation (Beer et al., 1997; Fu et al., 1998; Karat et al., 1992). Some articles describing studies comparing usability testing to expert evaluation go so far as to discuss their results in terms of problems found by users as opposed to problems found by evaluators (Catani & Biers, 1998; Fu et al., 1998; Law & Hvannberg, 2004b; Mankoff et al., 2005).

If it is the users who find the problems in usability testing, then it should follow that evaluators watching the same users would report the same usability problems. This was not observed in Study 4. Reliability of problem detection was low, with different evaluators finding different problems. This suggests that the evaluator has a significant role in the evaluation, and choice of evaluator or number of evaluators can affect the outcome of an evaluation, an additional threat to internal validity. Hertzum and Jacobsen (2003) reported similar findings in a study where they counted problems found by each evaluator from each user in a session recording. They suggest that the number of problems found will approximate C x SQRT(number of users x number of evaluators).

In this regard, usability testing is yet another form of expert review of an interface, a review that includes observing users in addition to direct examination of the interface. An evaluator may find a problem because she notices a violation of design principles, or because she observes a user having a critical incident. However, it is the evaluator that detects the problem, diagnoses the problem cause, and describes the problem so that it can be fixed. For every task and every user, the evaluator re-examines the interface, and uses her judgment to identify problems and rate their severity. Every problem identified is likely to be due to a combination of user actions and the evaluator's review of the interface. The first few users and the first few tasks serve to generate many usability problems, but problem discovery will diminish over time as the number of new problems due to expert review diminishes, as observed by Dumas and Sorce (1995) and the number of new problems due to user variability also diminishes. Involving multiple evaluators results in more problems found than involving a single evaluator, whether or not users are involved in the evaluation, because different evaluators generate different hypotheses about usability problems in the interface. Involving end-users in the evaluation process is extremely helpful, and can identify problems that an evaluator might not have considered in a purely expert review, but the problems are still filtered through the evaluator's personal judgment.

Why is the evaluator so important? Why does evaluator opinion play such a large role, even when all evaluators are watching the same user sessions? One answer to this question can be found in analogous research about the role of the medical practitioner in medical diagnosis. If analyzing usability testing sessions is a form of diagnostic decision-making, then biases in decision-making can help explain the existence of the evaluator effect in usability testing. A common model of medical diagnosis is that diagnosticians form hypotheses based on review of early data, use these hypotheses to predict further findings that would be present if the hypotheses were true, and use these predictions to guide their acquisition of additional data, deducing the accuracy of each hypothesis (Elstein, 1994). Practitioners need to know what questions to ask, what signs and symptoms distinguish between hypothesized diagnoses, what tests to order, and so on.

They refine their conception of the problem as they gather more information, develop hypotheses, and identify limiting factors on their solution (Kurfiss, 1988).

Usability diagnosis is similar to medical diagnosis. The usability practitioner forms hypotheses about problems in the interface, and uses these hypotheses to guide further data collection, either through personal exploration (expert review) or observation of users (usability testing). A closer analogy than the traditional one-doctor/one-patient relationship is an epidemiologist studying an outbreak of a contagious disease. The epidemiologist studies the environment and the many people in it to understand how the disease is transmitted, why some people catch the disease but others do not, and why different people show different symptoms for the same disease. The usability practitioner studies the interface and the users interacting with it to understand why some people experience usability problems but others do not, and why different users have different reactions to the same interaction flaw.

Examination of sources of errors in medical decision-making suggests that medical practitioners are subject to many cognitive biases that can affect the diagnostic outcome. Croskerry lists 32 *cognitive dispositions to respond*, which are biases in interpreting information and making decisions (Croskerry, 2002, 2003). Table 3.36 provides four examples of cognitive dispositions to respond in both the medical and usability realm. Many of these are based on cognitive heuristics. While heuristics can help the diagnostician make decisions quickly, they can also impair judgment, leading to diagnostic errors. Usability practitioners can also experience cognitive dispositions to respond. In terms of hypothesis-driven data collection and analysis, they can affect the problems the practitioner notices, the hypotheses that the evaluator forms, the causes the practitioner ascribes to problems, and the solutions the practitioner suggests. Based on the similarities between medical diagnosis and usability diagnosis, it is not surprising that there is an evaluator effect in usability testing.

Table 3.36 Examples of Cognitive Dispositions to Respond in Diagnosis

Cognitive Disposition to Respond		Medical Example (Croskerry, 2003)	Usability Example
Availability (Tversky & Kahneman, 1974)	The tendency of events that happen frequently or recently to come to mind more readily.	Can lead to over-diagnosis of common ailments or ailments within your specialty, and under-diagnosis of uncommon ailments or ailments outside your specialty.	You are more likely to notice problems/causes that you encounter frequently or select solutions that you have used before.
Diagnosis momentum (Croskerry, 2003)	Once a diagnosis has been selected, it tends to remain selected, even if the initial diagnosis was tentative.	A tentative diagnosis may become a final diagnosis as it is passed among patient, paramedics, nurses, doctors, etc., without being properly verified.	An initial diagnosis of problem cause or a tentative solution may be adopted as the final answer. Similar to the tendency of prototype design elements to become final design elements.
Over-confidence bias (Howell, 1971)	The belief that our decisions are more accurate than they really are, placing too much faith in opinion instead of evidence.	Can lead to selecting a final diagnosis too early (anchoring), selecting a common diagnosis based on insufficient information (availability), and failure to seek more information to rule out alternate diagnoses (commission).	Can lead to selecting a solution too quickly and over-reliance on expert evaluations as opposed to testing with users.
Loss aversion (Tversky & Kahneman, 1991) or sunk costs	The more that you invest in a strategy, the less likely you are to consider alternatives.	Once a physician has invested time and treatments in a particular diagnosis, she is less likely to consider alternative diagnoses.	Once a design or solution has been discussed, created, tested and redesigned, usability practitioners (or developers, managers, etc.) may be less willing to consider radically different approaches.

These cognitive dispositions to respond may have a great effect on an evaluator performing a heuristic evaluation, since these evaluations are strongly based on expert opinion. It might be expected that they would have a smaller effect on interpretation of usability testing sessions, as it is the user rather than the evaluator that leads the exploration of the interface. Indeed, some articles describing studies comparing usability testing to expert evaluation discuss their results in terms of problems found by users as opposed to problems found by evaluators (Catani & Biers, 1998; Fu et al., 1998; Law & Hvannberg, 2004b; Mankoff et al., 2005). If this were strictly true, then any evaluator watching a usability session would find the same problems as any other evaluator, which

has not been found in previous studies of the evaluator effect in usability testing with identical user sessions (Jacobsen et al., 1998; Lesaigle & Biers, 2000; Long et al., 2005; Palacio et al., 1993; Skov & Stage, 2005; Vermeeren et al., 2003). Usability testing is more than a simple signal-detection task where the evaluator either sees or misses problems experienced by users. User interactions with the interface require problem identification and diagnosis, which may differ from one evaluator to another. Many usability studies involve small numbers of users, and it is the responsibility of the usability expert to examine each usability problem and make judgments such as whether it will be experienced by other users, and is it specific to the testing situation and unlikely to occur in actual usage. The evaluator may also observe interaction flaws that did not cause problems during usability testing, but would cause problems for other users, based on the evaluator's knowledge and experience. These decisions are subject to the same cognitive dispositions to respond that affect decision-making in medical diagnosis.

Even with user-reported data, there is a need for expert decision-making, leading to a possibility of expert bias. Capra (2002) examined critical incidents reported by 24 users and found that users commonly reported positive interactions, such reporting having figured out how to complete a task without reporting the underlying problem that caused the delay in task completion. Problem descriptions were often vague or incomplete, and sometimes used inexact terminology, such as "pull-down menu"(a single-selection device) to refer to a multiple-selection list box. The user can be seen as assisting with data collection, but it is still the evaluator's role to distill problems, diagnose them, explain them, and (in some projects or organizations) solve them. The evaluator has to infer usability problems from the user's descriptions, in the same way that a medical practitioner has to infer a disease or condition from a patient's description of his symptoms.

There is a large body of research on medical decision-making and errors (see Croskerry, 2002; Kuhn, 2002; Patel, Arocha, & Kaufman, 2001). Considering the similarities between medial diagnosis and usability diagnosis, the field of usability might benefit from an understanding of the decision-making processes used by evaluators, including differences between experts and novices, and information that is necessary for

diagnosing usability problems. The usability community could also benefit from tools similar to those developed by the medical community. For example, structured reporting is used in the medical community as a way to reduce diagnostic errors and loss of clinical observations. The Digital Imaging and Communications in Medicine (DICOM) standard for storing and sharing medical images (National Electrical Manufacturers Association, 2004) is used to share radiology data, and standardized templates for evidence documents and reports are being developed to supplement stored images with clinical observations (Loef & Loef, 2005). The National Institute of Health's National Library of Medicine's Unified Medical Language System (UMLS) contains standardized vocabulary, semantic networks, and lexicons for computer-aided processing of medical information (National Library of Medicine, 2006). Structured reporting and taxonomies in usability (Andre et al., 2001; Cockton et al., 2004; Hvannberg & Law, 2003; Lavery et al., 1997; Sutcliffe, 2000) are in the early stages of development, but may provide similar benefits to the field of usability. Once we understand the process of usability diagnosis, we may be able to modify usability training, tools and techniques to support the diagnostic process and improve expert decision-making in usability diagnosis.

3.5.3 RQ3: How can we assess the content of UPDs?

The third area of exploration for Phase II was the usefulness of the guidelines developed in Phase I for grading usability reports in terms of the content of usability problem descriptions. Hypothesis 3a was that "good" evaluators follow the guidelines, and the operational test of this was that following the guidelines would be associated with several indicators of a better evaluation. The first test of this hypothesis was whether report ratings differed for students and practitioners, with the assumption that practitioners write better reports than students do. The hypothesis was supported for three of the guidelines (*Backing Data, Describe a Solution* and *Impact/Severity*) and across all six guidelines, with practitioners receiving higher scores than students. Although *Describe a Solution* is correlated with opinions for the four adjectives, this is unlikely to be a factor in student/practitioner differences for this guideline since there were no significant differences in opinion for this guideline between students and practitioners in Study 4 (see Appendix H). The hypothesis was not supported for *Clarity/Jargon*,

Problem Cause, or *User Actions*. This hypothesis was also tested by measuring correlations between the report ratings and other indicators of better evaluations: evaluator experience (years of experience, number of evaluations), thoroughness, validity, and hours spent on the evaluation. The hypothesis was not supported for these measures.

Hypothesis 3b was that the ratings of the judges would be reliable in terms of association (considering the same underlying traits in assigning a rating) and bias (tendency to give overall higher or lower ratings). This hypothesis was supported for four of the guidelines in terms of association, with significant correlations for providing *Backing Data*, *Describe a Solution*, *User Actions*, and *Problem Cause* (r = .31-.60) and overall (r = .39-.52). There were no significant correlations greater than .3 for *Clarity/Jargon* or *Impact/Severity*, which suggests that the judges based their ratings on different underlying traits. This unreliability could mask differences between students and practitioners, and makes the conclusion that hypothesis 3a was not supported for these two guidelines suspect. One of the judges gave lower overall scores than the other two judges. Further studies are needed to develop reliable means of rating reports for these guidelines and testing whether reports written by practitioners and students differ for these guidelines.

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 or rely on different skills. Figure 3.33 shows a scatter plot comparing thoroughness in finding severe problems and mean rating across the five *required* guidelines. The lines represent median values, and a slight random jitter (0.015) was added to offset overlapping points. The best reports are likely those in quadrant I, reports that received high marks for both finding severe problems and describing the problems. The two evaluators at the right edge of the graph received very high marks for finding severe problems, but an analysis of their report grades indicates that they did an average job of describing the problems that they found. Similarly, the evaluator at the left edge of the graph had the lowest thoroughness score but received a higher than average score for

following the guidelines. The weakest reports are likely those in quadrant III, reports that received low marks for both thoroughness and following the guidelines. Figure 3.34 is a contour graph of the thoroughness scores and report ratings, sorted by thoroughness for severe problems. This graph highlights not only evaluators whose guideline scores do not match their thoroughness scores, but also evaluators who have consistently high or low scores across all guidelines. The combination of measures provides a more complete picture of each evaluation.

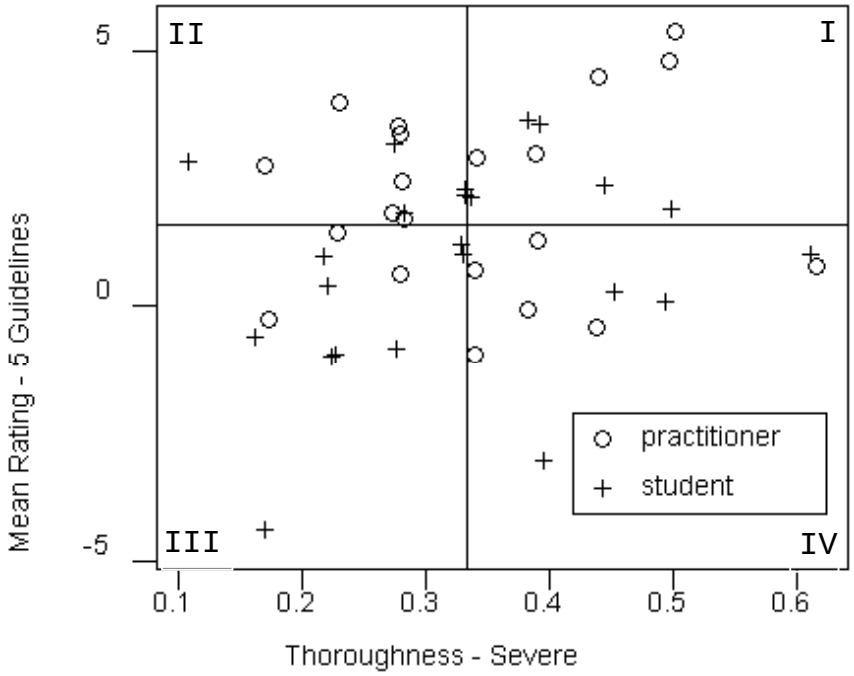
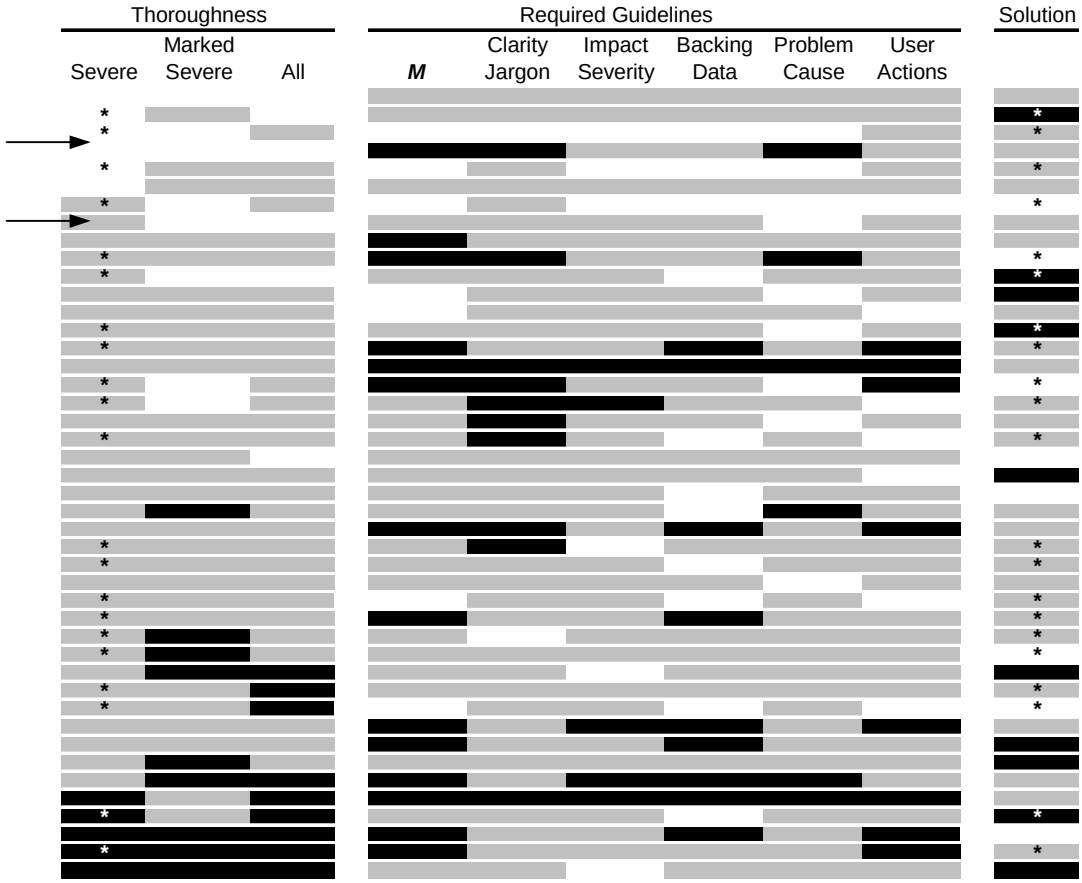


Figure 3.33 Study 4: Comparison of Thoroughness and Mean Guideline Ratings



Note: White: high score; Gray: $M \pm SD$; Black: low score. * Practitioner

Figure 3.34 Study 4: Contour Graph of Thoroughness, Guideline Scores

Table 3.37 provides several sample problem descriptions from a pair of evaluators, marked by arrows on Figure 3.34. The two evaluators received identical thoroughness scores for all problems and for severe marked severe problems, and similar scores for severe problems. Both were in the top five evaluators for overall thoroughness, both were students, yet they received very different scores for *Problem Cause* and *Clarity/Jargon*. One evaluator describes what happened, how many users encountered the problem, and provided some suggestions for fixing the problems. The other evaluator gave very brief descriptions of the problem, with little context or explanation of what happened. Assessment using only thoroughness would conclude that their evaluations were equally effective. However, further assessment with the report ratings reveals that the content of the reports are not equivalent, with one report containing descriptions that are more complete.

Table 3.37 Study 4: UPDs from Evaluators with Similar, High Thoroughness

	Descriptions from evaluator with…		
Matching Problems →	**…high rating for *Problem Cause***	**…low ratings for *Problem Cause, Clarity/Jargon***	**Commentary**
fa, fc	The results from "Find where X is credited alongside another name (Y)" present a list of links with checkboxes to matches for the search. When searching from X's page, the user assumes that clicking on exactly none or one of the checkboxes (Y) will provide the movies where X and Y are credited together, but in fact doing so displays a message "Need 2 or more names." Only then does the user realize that he has to check two boxes. The following participants encountered this problem: 1, 2, 3, 4.	On Name search screen, he then found the check box, but did not check two names.	There are several pages that have the title "Name Search." Although only one has checkboxes, the shorter description provides little context for the problem.
fd, fe	The availability to search for people who served different roles in a movie tripped up participant 4. She mistakenly searched for films in which Owen Wilson and Luke Wilson were both writers, not actors, and incorrectly concluded that there was only one movie where the brothers appeared as actors (there are actually 5). She also mistaken searched for films where both acted and wrote, again reaching an incorrect conclusion. Potential Solution: search for all pairs and present the results in more descriptive terms ("Films in which both Owen and Luke appeared as actors" or "Films in which Owen acted and Luke wrote").	Did not notice Actor, Writer, Director categories. Selected from the wrong category. Did not uncheck the boxes in the wrong category before checking the actor boxes.	The shorter description does not describe the user's actions and consequences, and does not explain why the user did not notice the actor, writer and director categories.

Word count is another distinguishing trait of these two evaluators. Both evaluators submitted approximately the same number of words (~825), but the first evaluator submitted nine problem descriptions of approximately 90 words each, while the second evaluator submitted 50 descriptions with an average of 17 words each. The shorter descriptions may be enough to jog the evaluator's memory about the problem, but do not provide enough detail for someone that has not seen the usability session movie to understand the problem without examining the website. For someone with only slight familiarity with the task, or a setting where many different tasks were tested, the shorter descriptions may not provide sufficient context about what web page or feature is being discussed. The evaluator herself may have increased difficult understanding her terse UPDs as time passes and memory fades.

Several additional report aspects that differentiate reports were described in Section 3.4.6: Interesting Observations. Evaluator misunderstandings can decrease evaluation effectiveness and harm the reputation of the evaluator in the eyes of the report recipients. Evaluators should mention testing protocol issues that could affect the interpretation of the results. Reports that use unprofessional language or are overly critical may be poorly received, decreasing the likelihood of implementation of suggested changes. Good reports do not have to have solutions or screen shots, but are indicative of different styles of usability reports.

3.6 Limitations

The Phase II study used pre-recorded usability sessions to provide all evaluators with identical usability sessions for their evaluations. Thus, Study 4 tested evaluators' performance in finding problems in this particular session movie, as opposed to designing and running studies to find problems in the target interface. Evaluators also had little context for the evaluation, such as user profiles, usability goals, task frequency, and other information that would generally be available to an evaluation team. The evaluators were given a general description of the report recipients, but no specifics. Many evaluators commented that they write different reports for different audiences, tailoring their reporting style based on whether the report is for other usability professionals, managers, developers, or clients.

The evaluators did not know that this task was one in a series of tasks performed during the test session; some evaluators guessed this, based on the disparity in performance among users, but some did not. The movie was also very short. Usability sessions are usually at least half an hour, more often 1-2 hours, and most usability evaluations involve more than four users. Although there were 41 usability problems in the movie, the 10-minute movie combined with the 2-hour estimate of study length and lack of reimbursement for study participation may have led some evaluators to write a shorter, less thorough report than they might have in a genuine work setting.

The goal of the study was to collect finished usability problem descriptions, descriptions that included some analysis of the problems, according to the scope illustrated in Figure 3.35. The evaluators that submitted executive summaries clearly performed some degree of analysis of the evaluation, diagnosing problem causes and interpreting actions based on the interaction design. However, some evaluators did not seem to have performed analysis, just data collection. For instance, the report that described the most problems also had extremely short descriptions, and was almost a running commentary of the users' actions, with no analysis of problem causes or discussion of the events. For evaluators that did not diagnose the usability problems in their reports, it is impossible to know whether they submitted unfinished reports or they do not typically conduct thorough analyses of their usability evaluations.

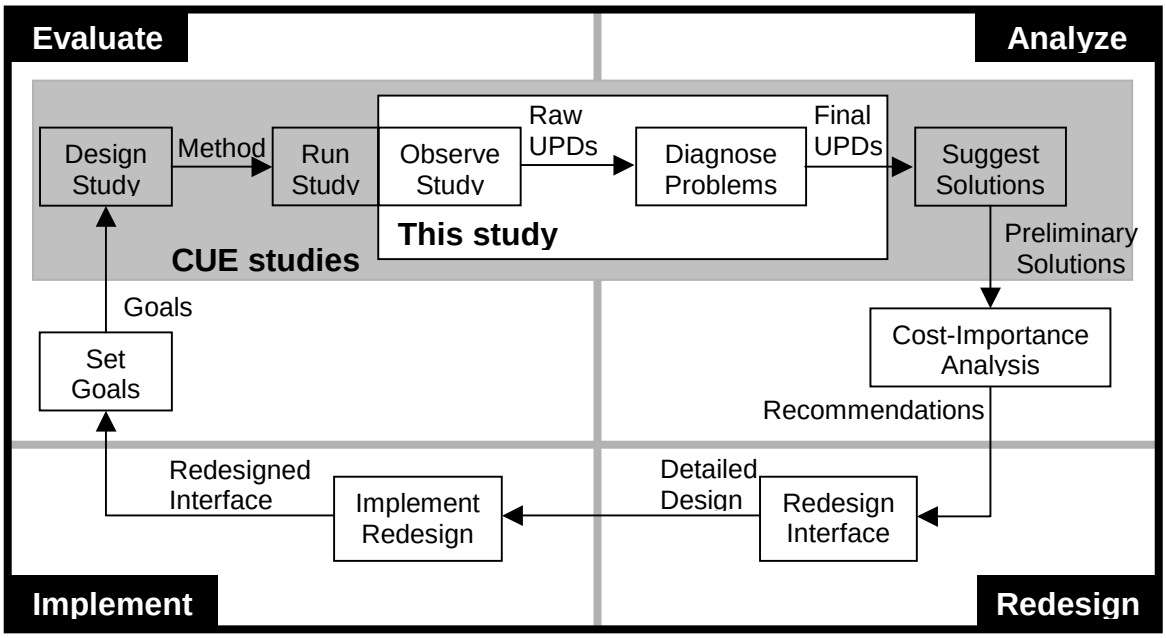


Figure 3.35 Study 4: Intended Scope of Study

Most of the student evaluators had the same instructor for their Usability Engineering class, which may have increased reliability of problem detection and problem severity judgments for the students. All three of the judges had the same instructor as most of the student evaluators, and all three judges completed Usability Engineering at Virginia Tech. This may have resulted in more similarity of severity

judgments between the students and judges, increasing thoroughness of severe problems and severe marked severe problem for the student evaluators.

This study examined inter-evaluator differences in problem detection, problem description, and severity judgments. A component of the observed variability may be due to intra-evaluator variability. In other words, an individual usability practitioner evaluating the same interface multiple times might generate different results. Intra-evaluator reliability would be difficult to study because earlier evaluations would affect later evaluations.

CHAPTER 4. Conclusion

The first goal of this research was to develop guidelines for describing usability problems. Ten guidelines were developed during Phase I through consultation with usability practitioners and a few senior PhD students in HCI via a questionnaire and a card sort. The five most *required* guidelines were identified using a second questionnaire. The 10 guidelines, sorted from most to least *required*, are (briefly):

- Be clear and precise while avoiding wordiness and jargon.
- Describe the impact and severity of the problem.
- Support your findings with data.
- Describe the cause of the problem.
- Describe observed user actions.
- Consider politics and diplomacy when writing your description.
- Be professional and scientific in your description.
- Describe a solution to the problem, providing alternatives and tradeoffs.
- Describe your methodology and background.
- Help the reader sympathize with the user's problem.

The complete text of the guidelines is listed in Table 2.12 on page 35. Phase II explored the evaluator effect in usability testing, comparing usability reports collected from 21 usability practitioners and 23 graduate students watching a pre-recorded usability session. Phase II also explored the guidelines further by asking evaluators to give their opinions about the guidelines, and by rating their reports for the five most *required* guidelines and *Describe a Solution*. This section summarizes three findings of this research and suggests future research in this area.

4.1 Usability Testing Benefits from Multiple Evaluators

Donaghue (2002) argues that usability testing is generally constrained by time and resources, and so the goal of testing is not to find all usability problems, but rather to find a reasonable number with a reasonable amount of effort. When using usability testing for

evaluation, it is generally accepted that multiple participants are needed to find a reasonable proportion of usability problems. Involving multiple users accounts for the "wild card effect" of variability in individual users (Gray & Salzman, 1998). Several studies have found that 5-16 users are enough to find a reasonable number of problems in small-to-medium systems (Faulkner, 2003; Nielsen, 1994a; Virzi, 1990, 1992; Woolrych & Cockton, 2001), with more needed for complex systems and diverse user populations (Spool & Schroeder, 2001), or to ensure finding a high percentage of problems in the system (Faulkner, 2003; Woolrych & Cockton, 2001). Hundreds of users may be necessary to find severe but rare problems in systems with highly diverse user populations, such as elections (Bailey, 2000, November).

An effect similar to the user "wild card effect" has been found for evaluators in previous studies of the evaluator effect in analytical methods and usability testing (Hertzum & Jacobsen, 2003). Hartson et al. (2003) suggest that unreliability in expert evaluations can be leveraged by combining multiple evaluators for higher overall thoroughness than any individual evaluator. Dumas and Sorce (1995) draw a similar conclusion, finding that having more evaluators spend fewer hours is more effective than having fewer evaluators spend more hours on an expert evaluation.

The low reliability between evaluators in Study 4 suggests that usability testing may benefit from multiple evaluators, as expert evaluation does, with a group of evaluators finding more problems than any individual evaluator does. Based on the results of the bootstrap simulation for Study 4 which used aggregates of individuals to simulate groups of evaluators, adding a second evaluator results in a 30-43% increase in problem detection, or four additional minor problems and two additional severe problems. Gains decreased with each additional evaluator, with a 12-20% increase from adding a third evaluator (three minor problems and one severe), a 7-12% increase for adding a fourth evaluator (two minor problems and one severe), and a 5-9% increase for adding a fifth evaluator (two minor problems and 70% chance of a severe problem).

The session recording used in Study 4 was 10 minutes long and involved a constant number of users (four) performing a single task. Hertzum and Jacobsen (2003)

used an hour-long recording and simulated varying the number of users by asking evaluators to timestamp the point in the recording where they identified each problem. They suggest that the number of problems found will approximate C x SQRT(number of users x number of evaluators). However, they point out that this equation was derived from a study of four evaluators. Further research is needed to explore several aspects the impact of the evaluator effect on usability testing:

- How do the number of users and evaluators affect the number of problems found through usability testing? Does the benefit from additional evaluators increase or decrease with larger numbers of users? Future studies could vary both of these factors to verify Hertzum and Jacobsen's equation or develop a new one.
- What factors affect the shape of the number-of-evaluators/thoroughness curve? Future studies should study the evaluator effect in different settings – longer sessions, different types of interfaces, different stages in the life cycle (from early prototypes to released systems), websites versus desktop software versus physical products, etc.
- How many evaluators are needed to find a reasonable number of problems in a usability session? From a cost-benefit standpoint, at what point does the cost of adding another evaluator become greater than the benefit of finding additional usability problems? This is contingent on knowing the return-on-investment of finding a usability problem, part of cost-justifying usability, an aspect of usability that itself needs further research.
- It is necessary to involve multiple evaluators to compensate for the "wild card effect" of individual evaluator variability? Can involving a large number of users compensate for a small number of evaluators? Can involving a large number of evaluators compensate for a small number of users?

When increasing the number of users in a usability test is undesirable due to limitations in time, resources, or available users, adding additional evaluators may be an appropriate strategy for increasing the number of problems found. Further research should verify the

results of this research using longer and more complex usability sessions and varied evaluation settings.

4.2 Is Usability Testing the "Gold Standard" of Evaluation?

Usability testing has been called the "gold standard" (Nickerson & Landauer, 1997, p. 17) of usability evaluations for its thoroughness in finding usability problems and focus on the user's experience with the interface. Usability testing with end users has been used as a benchmark or baseline to assess other formative evaluation methods, particularly analytical methods (Cuomo & Bowen, 1994; Desurvire et al., 1991; Desurvire et al., 1992; Law & Hvannberg, 2004b; Mankoff et al., 2005; Nielsen, 1994b). Several previous authors have suggested that usability testing may not be appropriate to use as a yardstick for other techniques because it is not perfect, being subject to limitations of the laboratory environment, study design flaws, and evaluator biases (Hartson et al., 2003; Jacobsen et al., 1998; Newman, 1998). The results of this research suggest that usability testing is subject to an evaluator effect similar to expert review techniques. Usability testing may find false alarms and incorrect problems, and it may miss real usability problems. Different evaluators find different problems, even when watching the same users. Simulations of the current study with smaller sample sizes found that individual thoroughness would have been over-estimated, which is a threat to internal validity. Studies that compare evaluation techniques to usability testing with a single evaluator may under-estimate the effectiveness of the other technique as compared to usability testing, which is also a threat to internal validity.

The implication for usability researchers is that usability testing may not be a reliable benchmark for other UEMs. In particular, a small usability evaluation (fewer than 10 users) with a single evaluator is not appropriate for creating a benchmark set of problems to assess the effectiveness of other UEMs. Further research is needed to determine whether increasing the number of users or evaluators is more effective and to determine appropriate numbers for various size interfaces and study purposes. Creating a comprehensive benchmark set of usability problems may require more users and evaluators than a simple formative evaluation as part of the product design cycle. We

need to carefully examine previous UEM comparison research and limit the implications we draw from studies that used a single evaluator for usability testing to generate the master problem list. It is also important to know the number of evaluators/evaluations that were involved in creating a master problem list. For example, while the current study involved 44 evaluators, the master problem list was based on 58 evaluators (44 participants, 7 pilot participants, and 7 pilot participants from an external researcher). Hornbæk and Frøkjær (2004b) tested two different techniques with 43-44 evaluators each, but created a problem list using all reports, and so their list was based on 87 evaluators.

4.3 Measure Both Quantity and Quality

The Phase II study found no differences between students and practitioners using measures of problem detection alone – thoroughness and reliability for all problems and severe problems. There were, however, differences between the groups across the six guidelines used in Phase II, and also for *Backing Data*, *Describe a Solution*, and *Impact/Severity*. Measuring adherence to the guidelines can provide a more complete picture of evaluation effectiveness than measuring problem detection alone. Several additional features of problem description content were identified, including use of images and screenshots, correctness, vagueness, and professionalism.

The guidelines developed for this study can be used in future studies to compare evaluation effectiveness. They can also be useful in training usability evaluators, providing instructors with a means to grade usability reports and provide feedback to students, and providing students with guidelines for writing problems descriptions in class projects. The guidelines could be used to select examples of good problem descriptions, bad problem descriptions, and different styles of writing problem descriptions for creating training materials for usability evaluators. For example, student evaluators could watch the movie used in Study 4, write their own problem descriptions, and then compare their descriptions to sample descriptions to note weaknesses in their own descriptions. While following each and every guideline might be too time-consuming for a practicing evaluator, it would be a useful exercise for usability students.

Usability practitioners could use the guidelines to modify problem reporting forms or checklists, and to evaluate their work products to ensure that they are writing effective problem descriptions in their usability reports.

Further work is needed to develop the guidelines as a tool for grading usability reports, particularly the guidelines where the judges' ratings were not reliable, *Clarity/Jargon* and *Impact/Severity.* Additional training or group discussion to achieve consensus before judging might increase reliability. If not, the statements for the Likert-type scales could be rephrased, additional explanations of the guidelines could be provided, or the guidelines could be rewritten from the original questionnaire responses. While opinions about the guidelines were generally positive, this study consulted primarily usability practitioners. Other stakeholders in the usability process, such as developers, managers, designers, marketing and product support, may have different opinions about the information that is important to them. The importance of the various guidelines may also differ depending on the type of product or type of company, and it is important to understand which guidelines are most important in various evaluation settings.

CHAPTER 5. References

Andre, T. S., Hartson, H. R., Belz, S. M., & McCreary, F. A. (2001). The user action framework: a reliable foundation for usability engineering support tools. *International Journal of Human-Computer Studies, 54*(1), 107-136.

ANSI. (2001). *Common Industry Format for Usability Test Reports* (ANSI NCITS 354-2001). New York: author.

ATLAS.ti. (2000). Cologne, Germany: ATLAS.ti Scientific Software Development.

Bailey, B. (2000, November). *UI Design Newsletter (2000 Presidential Election).* Fairfield, IA: Human Factors International. Retrieved October 31, 2004 from http://www.humanfactors.com/downloads/nov00.asp.

Bailey, G. (1993). Iterative methodology and designer training in human-computer interface design. In S. Ashlund & K. Mullet & A. Henderson & E. Hollnagel & T. White (Eds.), *Proceedings of the INTERCHI 93: Conference on Human Factors in Computing Systems (INTERACT 93 and CHI 93)* (pp. 198-205). New York: ACM.

Bailey, R. W., Allan, R. W., & Raiello, P. (1992). Usability Testing vs. Heuristic Evaluation: A Head-To-Head Comparison. In *Proceedings of the 36th Annual Meeting of the Human Factors and Ergonomics Society* (pp. 409-413). Santa Monica, CA: HFES.

Baker, K., Greenberg, S., & Gutwin, C. (2002). Empirical Development of a Heuristic Evaluation Methodology for Shared Workspace Groupware. In *Proceedings of the 2002 ACM Conference on Computer Supported Cooperative Work (CSCW 2002)* (pp. 96-105). New York: ACM.

Bastien, J. M. C., & Scapin, D. L. (1995). Evaluating a User Interface With Ergonomic Criteria. *International Journal of Human-Computer Interaction, 7*(2), 105-121.

Bastien, J. M. C., Scapin, D. L., & Leulier, C. (1996). Looking for usability problems with the ergonomic criteria and with the ISO 9241-10 dialogue principles. In M. J. Tauber & V. Bellotti & R. Jeffries & J. D. Mackinlay & J. Nielsen (Eds.), *Proceedings of the Conference on Human Factors in Computing Systems (CHI '96)* (pp. 77-78). New York: ACM.

Beer, T., Anodenko, T., & Sears, A. (1997). A Pair of Techniques for Effective Interface Evaluation: Cognitive Walkthroughs and Think-Aloud Evaluations. In *Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting* (pp. 380-384). Santa Monica, CA: Human Factors and Ergonomics Society.

Bobby 5.0. (2005). Waltham, MA: Watchfire. Retrieved on October 23 from http://www.watchfire.com/products/desktop/accessibilitytesting/default.aspx.

Capra, M. G. (2002). Contemporaneous versus retrospective user-reported critical incidents in usability evaluation. In *Proceedings of the Human Factors and Ergonomics Society 46th Annual Meeting* (pp. 1973-1977). Santa Monica, CA: Human Factors and Ergonomics Society.

Capra, M. G. (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.

Catani, M. B., & Biers, D. W. (1998). Usability Evaluation and Prototype Fidelity: Users and Usability Professionals. In *Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting* (pp. 1331-1335). Santa Monica, CA: Human Factors and Ergonomics Society.

Cattell, R. B. (1966). The Scree Test for the Number of Factors. *Multivariate Behavioral Research, 1*, 245-276.

Chattratichart, J., & Brodie, J. (2002). Extending the heuristic evaluation method through contextualisation. In *Proceedings of the 46th Annual Meeting of the Human Factors and Ergonomics Society (HFES 2002)* (pp. 641-645 CHECK!!!!!!!). Santa Monica, CA: HFES.

Chattratichart, J., & Brodie, J. (2004). Applying User Testing Data to UEM Performance Metrics. In E. Dykstra-Erickson & M. Tscheligy (Eds.), *Proceedings of the Conference on Human factors in computing systems (CHI 2004)* (pp. 1119-1122). New York: ACM.

Chin, J. P., Diehl, V. A., & Norman, L. K. (1988). Development of an instrument measuring user satisfaction of the human-computer interface. In E. Soloway & D. Frye & S. B. Sheppart (Eds.), *Proceedings of the Conference on Human Factors in Computing Systems (CHI '88)* (pp. 213-218). New York: ACM.

Cockton, G., Woolrych, A., & Hindmarch, M. (2004). Reconditioned Merchandise: Extended Structured Report Formats in Usability Inspection. In E. Dykstra-Erickson & M. Tscheligy (Eds.), *Proceedings of the Conference on Human factors in computing systems (CHI 2004)* (pp. 1433-1436). New York: ACM.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational & Psychological Measurement, 20*, 37-46.

Connell, I. W., & Hammond, N. V. (1999). Comparing Usability Evaluation Principles with Heuristics: Problem Instances vs. Problem Types. In M. A. Sasse & C. Johnson (Eds.), *Proceedings of the Human Computer Interaction - INTERACT '99* (pp. 621-629): IOS.

Cook, R. J. (1998). Kappa and Its Dependence on Marginal Rates. In P. Armitage & T. Colton (Eds.), *Encyclopedia of biostatistics* (pp. 2166-2168). Chichester, West Sussex, England: John Wiley.

Cordes, R. E. (2001). Task-selection bias: a case for user-defined tasks. *International Journal of Human-Computer Interaction, 13*(4), 411-419.

Croskerry, P. (2002). Achieving Quality in Clinical Decision Making: Cognitive Strategies and Detection of Bias. *Academic emergency medicine, 9*(11), 1184-1204.

Croskerry, P. (2003). The Importance of Cognitive Errors in Diagnosis and Strategies to Minimize Them. *Academic Medicine, 78*(8), 775-780.

Cuomo, D. L., & Bowen, C. D. (1992). Stages of User Activity Model as a Basis for User-System Interface Evaluation. In *Proceedings of the 36th Annual Meeting of the Human Factors and Ergonomics Society* (pp. 1254-1258). Santa Monica, CA: Human Factors and Ergonomics Society.

Cuomo, D. L., & Bowen, C. D. (1994). Understanding usability issues addressed by three user-system interface evaluation techniques. *Interacting With Computers, 6*(1), 86-108.

De Angeli, A., Matera, M., Costabile, M. F., Garzotto, F., & Paolini, P. (2000). Validating the SUE Inspection Technique. In *Proceedings of the Working Conference on Advanced Visual Interfaces (AVI 2000)* (pp. 143-150). New York: ACM.

De Angeli, A., Matera, M., Costabile, M. F., Garzotto, F., & Paolini, P. (2003). On the advantages of a systematic inspection for evaluating hypermedia usability. *International Journal of Human-Computer Interaction, 15*(3), 315-335.

Desurvire, H. (1994). Faster! Cheaper! Are Usability Inspection Methods as Effective as Empirical Testing? In J. Nielsen & R. L. Mack (Eds.), *Usability Inspection Methods* (pp. 173-202). New York: John Wiley & Sons.

Desurvire, H., Lawrence, D., & Atwood, M. (1991). Empiricism Versus Judgment: Comparing User Interface Evaluation Methods on a New Telephone-Based Interface. In S. P. Robertson & G. M. Olson & J. S. Olson (Eds.), *Proceedings of the Conference on Human Factors in Computing Systems (CHI'91)* (pp. 58-59). New York: ACM.

Desurvire, H., & Thomas, J. C. (1993). Enhancing the performance of interface evaluators using non-empirical usability methods. In *Proceedings of the 37th Annual Meeting of the Human Factors and Ergonomics Society* (pp. 1132-1136). Santa Monica, CA: Human Factors and Ergonomics Society.

Desurvire, H. W., Kondziela, J. M., & Atwood, M. E. (1992). What is Gained and Lost when Using Evaluation Methods Other than Empirical Testing. In A. M. Monk & D. Diaper & M. D. Harrison (Eds.), *Proceedings of the HCI '92 Conference on People and Computers VII*. UK: Cambridge University.

Donaghue, K. (2002). *Built for use : driving profitability through the user experience*. New York: McGraw-Hill.

Doubleday, A., Ryan, M., Springett, M., & Sutcliffe, A. (1997). A Comparison of Usability Techniques for Evaluating Design. In G. C. van der Veer & A. Henderson & S. Coles (Eds.), *Proceedings of the Symposium on Designing Interactive Systems: Processes, Practices, Methods, & Techniques (DIS 1997)* (pp. 101-110). New York: ACM.

Dumas, J., & Sorce, J. (1995). Expert reviews: how many experts is enough? In *Proceedings of the Human Factors and Ergonomics Society 39th Annual Meeting* (pp. 228-232). Santa Monica, CA: Human Factors and Ergonomics Society.

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

Dumas, J. S., & Redish, J. C. (1993). *A Practical Guide to Usability Testing*. Norwood, NJ: Ablex.

Durso, F. T., & Gronlund, S. D. (1999). Situation Awareness. In F. T. Durso & R. S. Nickerson & R. W. Schvaneveldt & S. T. Dumais & D. S. Lindsay & M. T. H. Chi (Eds.), *The Handbook of Applied Cognition* (pp. 283-314). Chichester, Sussex, UK: John Wiley & Sons.

Dutt, A., Johnson, H., & Johnson, P. (1994). Evaluating Evaluation Methods. In G. Cockton & S. W. Draper & G. R. S. Weir (Eds.), *Proceedings of the HCI '94 Conference on People and Computers VII* (pp. 109-121). UK: Cambridge University.

Elstein, A. S. (1994). What goes around comes around: return of the hypothetico-deductive strategy. *Teaching and Learning in Medicine, 6*(2), 121-123.

Elstein, A. S., & Bordage, G. (1979). Psychology of clinical reasoning. In G. C. Stone & F. Cohen & N. E. Adler (Eds.), *Health psychology, a handbook : theories, applications, and challenges of a psychological approach to the health care system* (pp. 333-367). San Francisco: Jossey-Bass.

Elstein, A. S., Shulman, L. S., & Sprafka, S. A. (1978). *Medical Problem Solving: An Analysis of Clinical Reasoning*. Cambridge, MA: Harvard University.

Everitt, B. (1974). *Cluster Analysis*. London: Heinemann Educational Books.

Faulkner, L. (2003). Beyond the five-user assumption: Benefits of increased sample sizes in usability testing. *Behavior Research Methods, Instruments, & Computers, 35*(3), 379-383.

Frøkjær, E., & Hornbæk, K. (2005). Cooperative usability testing: complementing usability tests with user-supported interpretation sessions. In *Proceedings of the Conference on Human factors in computing systems: Extended Abstracts (CHI 2005)* (pp. 1383-1386). New York: ACM.

Fu, L., Salvendy, G., & Turley, L. (1998). Who Finds What in Usability Evaluation. In *Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting* (pp. 1341-1345). Santa Monica, CA: Human Factors and Ergonomics Society.

Gambrill, E. (1990). *Critical thinking in clinical practice: improving the accuracy of judgments and decisions about clients*. San Francisco: Jossey-Bass.

Garb, H. N. (1998). *Studying the clinician: judgment research and psychological assessment*. Washington, DC: American Psychological Association.

Gerhardt-Powals, J. (1996). Cognitive engineering principles for enhancing human-computer performance. *International Journal of Human-Computer Interaction, 8*(2), 189-221.

Gilhooly, K. J. (1990). Cognitive psychology and medical diagnosis. *Applied Cognitive Psychology, 4*(4), 261-272.

Gray, W. D., & Salzman, M. C. (1998). Damaged merchandise? A review of experiments that compare usability evaluation methods. *Human-Computer Interaction, 13*(3), 203-261.

Griffin, T., Schwartz, S., & Sofronoff, K. (1998). Implicit Processes in Medical Diagnosis. In K. Kirsner & C. Speelman & M. Maybery & A. O'Brien-Malone & M. Anderson & C. MacLeod (Eds.), *Implicit and Explicit Mental Processes* (pp. 329-341). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Hammond, N., Hinton, G., Barnard, P., MacLean, A., Long, J., & Whitefield, W. (1984). Evaluating the interface of a document processor: A comparison of expert judgment and user observation. In B. Shackel (Ed.) *Proceedings of the Human Computer Interaction - INTERACT '84* (pp. 725-729). North-Holland: Elsevier Science.

Hartson, H. R., Andre, T. S., & Williges, R. C. (2003). Criteria for evaluating usability evaluation methods. *International Journal of Human-Computer Interaction, 13*(4), 145-181.

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

Henderson, R. D., Smith, M. C., Podd, J., & Varela-Alvarez, H. (1995). A Comparison of the Four Prominent User-Based Methods for Evaluating the Usability of Computer Software. *Ergonomics, 38*(10), 2030-2044.

Hertzum, M., & Jacobsen, N. E. (1999). The Evaluator Effect during First-Time Use of the Cognitive Walkthrough Technique. In H.-J. Bullinger & J. Ziegler (Eds.), *Proceedings of the HCI International '99 (HCII99): Human-Computer Interaction: Ergonomics and User Interfaces* (pp. 1063-1067). London: Lawrence Erlbaum Associates.

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.

Hertzum, M., Jacobsen, N. E., & Molich, R. (2002). Usability Inspections by Groups of Specialists: Perceived Agreement in Spite of Disparate Observations. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI 2002) [extended abstracts]* (pp. 662-663). New York: ACM Press.

Hix, D., & Hartson, H. R. (1993). *Developing user interfaces: ensuring usability through product & process*. New York: John Wiley & Sons.

Hornbæk, K., & Frøkjær, E. (2004a). Two psychology-based usability inspection techniques studied in a diary experiment. In *Proceedings of the NordiCHI '04* (pp. 3-12). New York: ACM.

Hornbæk, K., & Frøkjær, E. (2004b). Usability inspection by metaphors of human thinking compared to heuristic evaluation. *International Journal of Human-Computer Interaction, 17*(3), 357-374.

Hornbæk, K., & Frøkjær, E. (2005). Comparing Usability Problems and Redesign Proposals as Input to Practical Systems Development. In *Proceedings of the Conference on Human factors in computing systems (CHI 2005)* (pp. 391-400). New York: ACM.

Howell, W. C. (1971). Uncertainty from Internal and External Sources: A Clear Case of Overconfidence. *Journal of Experimental Psychology, 89*(2), 240-243.

Hutchinson, T. P. (1993). Kappa Muddles Together Two Sources of Disagreement: Tetrachoric Correlation is Preferable. *Research in Nursing & Health, 16*, 313-315.

Hvannberg, E. T., & Law, L.-C. (2003). Classification of Usability Problems (CUP) Scheme. In M. Rauterberg (Ed.) *Proceedings of the Human-Computer Interaction - INTERACT'03* (pp. 655-662): IOS.

ISO. (1998). *Ergonomic requirements for office work with visual display terminals (VDTs) -- Part 11: Guidance on usability* (ISO 9241-11:1998(E)). Geneva: author.

ISO. (1999a). *Ergonomic requirements for office work with visual display terminals (VDTs) -- Part 10: Dialogue principles* (ISO 9241-10:1998(E)). Geneva: author.

ISO. (1999b). *Human -centred design processes for interactive systems* (ISO 134071999(E)). Geneva: author.

Jaccard, P. (1912). The distribution of flora in the alpine zone. *The New Phytologist, 11*(2), 37-50.

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.

Jeffries, R. (1994). Usability Problem Reports: Helping Evaluators Communicate Effectively with Developers. In J. Nielsen & R. L. Mack (Eds.), *Usability Inspection Methods* (pp. 273-294). New York: John Wiley.

Jeffries, R., Miller, J. R., Wharton, C., & Uyeda, K. M. (1991). User Interface Evaluation in the Real World: A Comparison of Four Techniques. In S. P. Robertson & G. M. Olson & J. S. Olson (Eds.), *Proceedings of the Conference on Human Factors in Computing Systems (CHI'91)* (pp. 119-124). New York: ACM.

John, B. E., & Kieras, D. E. (1996). The GOMS family of user interface analysis techniques: comparison and contrast. *ACM Transactions on Computer-Human Interaction (TOCHI), 3*(4), 320-351.

John, B. E., & Marks, S. J. (1997). Tracking the effectiveness of usability evaluation methods. *Behaviour & Information Technology, 16*(4/5), 188-202.

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika, 23*, 187-200.

Karat, C.-M., Campbell, R., & Fiegel, T. (1992). Comparison of empirical testing and walkthrough methods in user interface evaluation. In P. Bauersfeld & J. Bennett & G. Lynch (Eds.), *Proceedings of the Conference on Human Factors in Computing Systems (CHI 92)* (pp. 397-404). New York: ACM.

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: John Wiley & Sons.

Keenan, S. L., Hartson, H. R., Kafura, D. G., & Schulman, R. S. (1999). The usability problem taxonomy: A framework for classification and analysis. *Empirical Software Engineering, 4*(1), 71-104.

Kessner, M., Wood, J., Dillon, R. F., & West, R. L. (2001). On the Reliability of Usability Testing. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI 2001) [extended abstracts]* (pp. 97-98). New York: ACM.

Kuhn, G. (2002). Diagnostic Errors. *Academic emergency medicine, 9*(7), 740-750.

Kurfiss, J. G. (1988). *Critical Thinking: Theory, Research, Practice, and Possibilities* (ASHE-ERIC Higher Education Report No. 2). Washington, DC: The George Washington University, Graduate School of Education and Human Development.

Lavery, D., Cockton, G., & Atkinson, M. P. (1997). Comparison of evaluation methods using structured usability problem reports. *Behaviour & Information Technology, 16*(4/5), 246-266.

Law, E. L.-C. (2004). A Multi-Perspective Approach to Tracking the Effectiveness of User Tests: A Case Study. In K. Hornbæk & J. Stage (Eds.), *Proceedings of the Workshop on Improving the Interplay between Usability Evaluation and User Interface Design, NordiCHI 2004 (HCI-Lab Report no. 2004/2)* (pp. 36-40). Denmark: Department of Computer Science, Aalborg University.

Law, E. L.-C., & Hvannberg, E. T. (2004a). Analysis of Combinatorial User Effect in International Usability Tests. In E. Dykstra-Erickson & M. Tscheligy (Eds.), *Proceedings of the Conference on Human factors in computing systems (CHI 2004)* (pp. 9-16). New York: ACM.

Law, E. L.-C., & Hvannberg, E. T. (2004b). Analysis of Strategies for Improving and Estimating the Effectiveness of Heuristic Evaluation. In *Proceedings of the NordiCHI '04* (pp. 241-250). New York: ACM.

Law, L.-C., & Hvannberg, E. T. (2002). Complementarity and Convergence of Heuristic Evaluation and Usability Test: A Case Study of UNIVERSAL Brokerage Platform. In *Proceedings of the second Nordic conference on Human-computer interaction* (pp. 71-80). New York: ACM.

Lesaigle, E. M., & Biers, D. W. (2000). Effect of Type of Information on Real-Time Usability Evaluation: Implications for Remote Usability Testing. In O. Brown, Jr. (Ed.) *Proceedings of the International Ergonomics Association XIVth Triennial Congress and Human Factors and Ergonomics Society 44th Annual Meeting* (p. 585). Santa Monica, CA: Human Factors and Ergonomics Society.

Lewis, C. H., Polson, P., Wharton, C., & Rieman, J. (1990). Testing a walkthrough methodology for theory-based design of walk-up-and-use interfaces. In J. C. Chew & J. Whiteside (Eds.), *Proceedings of the Conference on Human factors in computing systems (CHI '90)* (pp. 235-242). New York: ACM.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 140*, 55.

Loef, C., & Loef, R. (2005). Evidence and Diagnostic Reporting in the IHE Context [Technical Report]. *Academic Radiology, 12*(5), 620-625.

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.

Lorr, M. (1983). *Cluster Analysis for Social Scientists*. San Francisco: Jossey-Bass.

Mack, R., & Montaniz, F. (1994). Observing, Predicting, and Analyzing Usability Problems. In R. L. Mack & J. Nielsen (Eds.), *Usability Inspection Methods* (pp. 295-340). New York: John Wiley & Sons.

Mankoff, J., Fait, H., & Tran, T. (2005). Is Your Web Page Accessible? A Comparative Study of Methods for Assessing Web Page Accessibility for the Blind. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI 2005)*. New York: ACM.

Mayhew, D. J. (1999). *The Usability Engineering Lifecycle*. San Francisco: Morgan Kaufman.

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

Molich, R., Bevan, N., Curson, I., Butler, S., Kindlund, E., Miller, D., et al. (1998). Comparative Evaluation of Usability Tests. In *Proceedings of the Usability Professionals' Association 1998 (UPA 1998) Conference* (pp. 189-200): Retrieved on May 9, 2005 from Rolf Molich's Dialogue Design website, http://www.dialogdesign.dk/tekster/cue1/cue1paper.doc.

Molich, R., & Dumas, J. (2006). *Comparative usability evaluation (CUE-4)* Manuscript submitted for publication.

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

Molich, R., & Nielsen, J. (1990). Improving Human-Computer Dialogue. *Communications of the ACM, 33*(3), 338-348.

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.

Monk, A. M. (1998). Experiments are for small questions, not large ones like "what usability evaluation method should I use?" In G. M. Olson & T. P. Moran (Eds.), *Commentary on "Damaged Merchandise?" Human-Computer Interaction, 13*, 263-323.

National Electrical Manufacturers Association. (2004). *Digital Imaging and Communications in Medicine (DICOM)* (Industry Standard). Rosslyn, Virginia: author.

National Library of Medicine. (2006). *Unified Medical Language System*, [website]. Retrieved February 19, 2006 from http://www.nlm.nih.gov/research/umls/.

Newman, W. M. (1998). On simulation, measurement and piecewise usability evaluation. In G. M. Olson & T. P. Moran (Eds.), *Commentary on "Damaged Merchandise?" Human-Computer Interaction, 13*, 263-323.

Nickerson, R. S., & Landauer, T. K. (1997). Human-Computer Interaction: Background and Issues. In M. Helander & T. K. Landauer & P. Prabhu (Eds.), *Handbook of Human-Computer Interaction* (2nd ed., pp. 3-31). Amsterdam: Elsevier Science.

Nielsen, J. (1992). Finding Usability Problems Through Heuristic Evaluation. In P. Bauersfeld & J. Bennett & G. Lynch (Eds.), *Proceedings of the Conference on Human Factors in Computing Systems (CHI 92)* (pp. 373-380). New York: ACM.

Nielsen, J. (1993). *Usability Engineering*. Boston: Academic.

Nielsen, J. (1994a). Estimating the number of subjects needed for a thinking aloud test. *International Journal of Human-Computer Studies, 41*(3), 385-397.

Nielsen, J. (1994b). Heuristic Evaluation. In J. Nielsen & R. L. Mack (Eds.), *Usability Inspection Methods* (pp. 25-62). New York: John Wiley & Sons.

Nielsen, J., & Landauer, T. K. (1993). A Mathematical model of the finding of usability problems. In S. Ashlund & K. Mullet & A. Henderson & E. Hollnagel & T. White (Eds.), *Proceedings of the INTERCHI 93: Conference on Human Factors in Computing Systems (INTERACT 93 and CHI 93)* (pp. 206-213). New York: ACM.

Nielsen, J., & Molich, R. (1990). Heuristic Evaluation of User Interfaces. In J. C. Chew & J. Whiteside (Eds.), *Proceedings of the Empowering people: Human factors in computing system: special issue of the SIGCHI Bulletin* (pp. 249-256). New York: ACM.

Nielsen, J., & Phillips, V. L. (1993). Estimating the relative usability of two interfaces: heuristic, formal, and empirical methods compared. In S. Ashlund & K. Mullet & A. Henderson & E. Hollnagel & T. White (Eds.), *Proceedings of the INTERCHI 93: Conference on Human Factors in Computing Systems (INTERACT 93 and CHI 93)* (pp. 214-221). New York: ACM.

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The Measurement of Meaning*. Urbana, IL: University of Illinois.

Palacio, F. L., Bloch, D. R., & Righi, C. (1993). A Comparison of Interobserver Agreement and Quantity of Usability Data Obtained Using Graphics-Based and Text-Based Data Collection Tools. In A. Gawman & E. Kidd & P.-Å. Larson (Eds.), *Proceedings of the 1993 Conference of the Centre for Advanced Studies on Collaborative Research: Distributed Computing* (pp. 1053-1058): IBM.

Patel, V. L., Arocha, J. F., & Kaufman, D. R. (2001). A primer on aspects of cognition for medical informatics. *Journal of the American Medical Informatics Association : JAMIA, 8*(4), 324-343.

Patel, V. L., Kaufman, D. R., & Magder, S. A. (1996). The acquisition of medical expertise in complex dynamic environments. In K. A. Ericsson (Ed.), *The road to excellence: the acquisition of expert performance in the arts and sciences, sports and games* (pp. 127-165). Mahwah, NJ: Lawrence Erlbaum Associates.

Preece, J. (1994). *Human-Computer Interaction*. Wokingham, England: Addison-Wesley.

Quesenbery, W. (2005). Usability Standards: Connecting Practice Around the World. In *Proceedings of the IEEE International Professional Communication Conference* (pp. 451-457). Piscataway, NJ: IEEE.

Rieman, J., Davies, S., Hair, D. C., Esemplare, M., Polson, P., & Lewis, C. (1991). An automated cognitive walkthrough. In S. P. Robertson & G. M. Olson & J. S. Olson (Eds.), *Proceedings of the Conference on Human Factors in Computing Systems (CHI'91)* (pp. 427-428). New York: ACM.

Roget's New Millennium™ Thesaurus. (2005). (First (v1.1.1) ed.). Los Angeles, CA: Lexico. Available via http://thesaurus.reference.com/.

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.

Rubin, J. (1994). *Handbook of usability testing: how to plan, design, and conduct effective tests*. New York: John Wiley & Sons.

Scapin, D. L., & Bastien, J. M. C. (1997). Ergonomic criteria for evaluating the ergonomic quality of interactive systems. *Behaviour & Information Technology, 16*(4/5), 220-231.

Sears, A. (1995). Heuristic walkthroughs: Combining the advantages of existing evaluation techniques. In *Proceedings of the 3rd Annual Mid-Atlantic Human Factors Conference* (pp. 29-35). Blacksburg: Virginia Tech.

Sears, A. (1997). Heuristic Walkthroughs: Finding the Problems Without the Noise. *International Journal of Human-Computer Interaction, 9*(3), 213-234.

Sears, A., & Hess, D. J. (1998). The Effect of Task Description Detail on Evaluator Performance with Cognitive Walkthroughs. In C.-M. Karat & A. Lund & J. Coutaz & J. Karat (Eds.), *Proceedings of the CHI 98: Conference on Human Factors in Computing Systems* (pp. 259-260). New York: ACM.

Skov, M. B., & Stage, J. (2003). Enhancing Usability Testing Skills of Novice Testers: A Longitudinal Study. In *Proceedings of the 2nd International Conference on Universal Access in Human-Computer Interaction (UAHCI 2003)* (pp. 1035-1039). Mahwah, NJ: Lawrence Erlbaum.

Skov, M. B., & Stage, J. (2004). Integrating Usability Design and Evaluation: Training Novice Evaluators in Usability Testing. In K. Hornbæk & J. Stage (Eds.), *Proceedings of the Workshop on Improving the Interplay between Usability Evaluation and User Interface Design, NordiCHI 2004 (HCI-Lab Report no. 2004/2)* (pp. 36-40). Denmark: Department of Computer Science, Aalborg University.

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.

Smilowitz, E. D., Darnell, M. J., & Benson, A. E. (1994). Are we overlooking some usability testing methods? A comparison of lab, beta, and forum tests. *Behaviour & Information Technology, 13*(1-2), 183-190.

Smith, S. L., & Mosier, J. N. (1986). *Guidelines for Designing User Interface Software* (Tech. Rep. ESD-TR-86-278). Bedford, MA: The MITRE Corporation.

Somervell, J., & McCrickard, D. S. (2004). Comparing Generic vs. Specific Heuristics: Illustrating a New UEM Comparison Technique. In *Proceedings of the Human Factors and Ergonomics Society 48th Annual Meeting (HFES '04)* (pp. 2480-2484). Santa Monica, CA: HFES.

Speelman, C. (1998). Implicit expertise: do we expect too much from our experts? In K. Kirsner & C. Speelman & M. Maybery & A. O'Brien-Malone & M. Anderson & C. MacLeod (Eds.), *Implicit and Explicit Mental Processes* (pp. 135-147). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Spool, J., & Schroeder, W. (2001). Testing web sites: five users is nowhere near enough. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI 2001) [extended abstracts]* (pp. 285-286). New York: ACM.

Spool, J. M., & Schaffer, E. M. (2005). The Great Debate: Can Usability Scale Up? In *Proceedings of the SIGCHI conference on Human factors in computing systems (CHI '05) [extended abstracts]* (pp. 1174-1175). New York: ACM.

Stone, D., Jarrett, C., Woodroffe, M., & Minocha, S. (2005). *User Interface Design and Evaluation*. San Francisco: Morgan Kaufmann.

Sutcliffe, A. (2000). On the Effective Use and Reuse of HCI Knowledge. *ACM Transactions on Computer-Human Interaction (TOCHI), 7*(2), 197-221.

Sutcliffe, A., Ryan, M., Doubleday, A., & Springett, M. (2000). Model mismatch analysis: toward a deeper explanation of users' usability problems. *Behaviour & Information Technology, 19*(1), 43-55.

Theofanos, M., & Quesenbery, W. (2005). Towards the Design of Effective Formative Test Reports. *Journal of Usability Studies, 1*(1), 27-45.

Theofanos, M., Quesenbery, W., Snyder, C., Dayton, D., & Lewis, J. (2005). *Reporting on Formative Testing. A UPA 2005 Workshop Report*. Bloomingdale, IL: UPA. Retrieved on October 3, 2005 from the UPA website http://www.upassoc.org/usability_resources/conference/2005/formative%20reporting-upa2005.pdf.

Tullis, T., & Wood, L. (2004). How Many Users Are Enough for a Card-Sorting Study? In *Proceedings of the 13th Annual Conference of the Usability Professionals' Association (UPA 2004) [conference CD]*. Bloomingdale, IL: UPA.

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science, 185*, 1124-1131.

Tversky, A., & Kahneman, D. (1991). Loss Aversion in Riskless Choice: A Reference-Dependent Model. *The Quarterly Journal of Economics, 106*(4), 1039-1061.

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.

Virzi, R. A. (1990). Streamlining the design process. Running fewer subjects. In M. E. Wiklund (Ed.) *Proceedings of the Human Factors Society 34th Annual Meeting* (pp. 291-294). Santa Monica, CA: Human Factors Society.

Virzi, R. A. (1992). Refining the Test Phase of Usability Evaluation: How Many Subjects Is Enough? *Human Factors, 34*(4), 457-468.

Virzi, R. A., Sorce, J. F., & Herbert, L. B. (1993). A Comparison of Three Usability Evaluation Methods: Heuristic, Think-Aloud, and Performance Testing. In *Proceedings of the 37th Annual Meeting of the Human Factors and Ergonomics Society* (pp. 309-313). Santa Monica, CA: Human Factors and Ergonomics Society.

Waloszek, G. (2003, 11/06/2003). *User Interface Design - Is It a Science, an Art, or a Craft?* SAP AG, Product Design Center: Walldorf, Germany. Retrieved October 10, 2005 from SAP website, http://www.sapdesignguild.org/community/editorials/editorial_03_2003.asp.

WebXACT. (2004). Waltham, MA: Watchfire. Retrieved on October 23, 2005 from http://webxact.watchfire.com/.

Wharton, C., Rieman, J., Lewis, C., & Polson, P. (1994). The cognitive walkthrough method: a practitioner's guide. In J. Nielsen & R. L. Mack (Eds.), *Usability Inspection Methods* (pp. 105-140). New York: John Wiley.

Wixon, D. R. (2003). Evaluating Usability Methods: Why the Current Literature Fails the Practitioner. *interactions, 10*(4), 28-34.

Woolrych, A., & Cockton, G. (2001). Why and When Five Test Users aren't Enough. In J. Vanderdonckt & A. Blandford & A. Derycke (Eds.), *Proceedings of the People and Computers XV - Interaction without Frontiers : Joint Proceedings of HCI 2001 and IHM 2001* (pp. 105-108). Toulouse, France: Cépadèus Éditions.

Wright, P. C., & Monk, A. M. (1991). A cost-effective evaluation method for use by designers. *International Journal of Man-Machine Studies, 35*(6), 891-912.

Zhang, Z., Basili, V., & Shneiderman, B. (1998). Empirical Study of Perspective-Based Usability Inspection. In *Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting* (pp. 1346-1350). Santa Monica, CA: Human Factors and Ergonomics Society.

Zhang, Z., Basili, V., & Shneiderman, B. (1999). Perspective-based Usability Inspection: An Empirical Validation of Efficacy. *Empirical Software Engineering, 4*(1), 43-69.

This page intentionally left blank.

APPENDIX A. UEM Comparison and Evaluator Effect Studies

A.1 Summary of Studies

This table lists previous studies comparing UEMs or measuring the evaluator effect within a UEM. This table excludes studies that focused on the user effect or determining how many users to include in a usability test (Faulkner, 2003; Law & Hvannberg, 2004a; Smilowitz et al., 1994; Spool & Schroeder, 2001; Virzi, 1990, 1992; Woolrych & Cockton, 2001), focus soley on the effects of design and redesign (Bailey, 1993; John & Marks, 1997) and studies that used already-found problems for severity judgments (Hassenzahl, 2000), evaluation of competing heuristics (Somervell & McCrickard, 2004) or classification reliability (Andre et al., 2001). This table also excludes UEM comparison studies that only used one evaluator per UEM, which are too numerous to list individually. Abbreviations for specific UEMs are explained in a separate table in section A.2.

Study	UEM	#, Type of Evaluators*
Bailey, Allan & Raiello (1992)	UT	1 unspecified, 4 rounds of testing/redesign
	HE	22 computer professionals (programmers, SA, HF) compares to Molich & Nielsen (1990)
Baker, Greenberg & Gutwin (2002)	HEG	16 CS HCI undergraduates 2 HCI professors 9 HCI graduate students
Bastien & Scapin (1995)	EE	9 usability specialists
	EC	
	EE	10 usability specialists
	EE	
Bastien, Scapin & Leulier (1996)	EC	6 ergonomics students
	DP	5 ergonomics students
	EE	6 ergonomics students
Beer, Anodenko & Sears (1997)	CW	6 software developers/testers
	UT	1 author
Catani & Biers (1998)	EE	5 usability professionals
	UT	1 evaluator
Chattratichart & Brodie (2004)	HE	5 graduate students
	HE-Plus	4 graduate students
Connell & Hammond (1999)	HE	8 novice undergrads, 5 experienced researchers
	CHP	8 novice undergrads, 5 experienced researchers
	EE	8 novice undergrads, 5 experienced researchers
	HE	8 novice undergrads
	CHP	8 novice undergrads
	EE	7 novice undergrads
Cuomo & Bowen (1992; 1994)	SMG	1 HF professional
	EE	2 HF professionals
	ACW	2 HF professionals
	UT	1 unspecified
De Angeli, Matera, Costabile, Garzotto & Paolini (2003)	HE	14 HCI seniors
	SUE	14 HCI seniors

Study	UEM	#, Type of Evaluators*
Desurvire & Thomas (1993), Desurvire(1994)	UT	1 unspecified (same as next two studies)
	PAVE	3 HFE experts 3 system developers 3 non experts
Desurvire, Kondziela & Atwood (1992), Desurvire(1994)	UT	1 unspecified (same as previous and study)
	ACW	3 HFE experts 3 system developers** 3 non-experts
	HE	3 HFE experts 3 system developers** 3 non-experts
Desurvire, Lawrence & Atwood (1991), Desurvire(1994)	UT	1 unspecified (same as previous two studies)
	SMG	>1 UI expert
	SMG	>1 UI non-expert
Doubleday, Ryan, Springett & Sutcliffe (1997)	UT	1 unspecified
	HE	5 HCI experts
Dumas, Molich & Jeffries (2004), Rourke (2003): CUE-4	UT	9 professional teams
	EE	8 professional teams
Dumas & Sorce (1995)	EE	5 usability experts
Dutt, Johnson & Johnson (1994)	HE	3 graduate students, 2 in HCI
	CW	3 graduate students, 2 in HCI
Frøkjær & Hornbæk (2005)	UT	4 mixed backgrounds
	CUT	4 mixed backgrounds
Fu, Salvendy & Turley (1998)	UT	1 team unspecified
	HE	6 usability experts
Hammond, Hinton, Barnard, MacLean, Long & Whitefield (1984)	UT	1 unspecified
	EE	6 HF experts
Henderson, Smith, Podd & Varela-Alvarez (1995)	LD	40 psychology students across 3 interfaces
	HSQ	39 psychology students across 3 interfaces
	INT	37 psychology students across 3 interfaces
	VP	32 psychology students across 3 interfaces
Hertzum & Jacobsen (1999)	CW	11 graduate students, half with industry design experience
Hertzum, Jacobsen & Molich (2002): CUE-3	HE	11 professional usability specialists
Hornbæk & Frøkjær (2004a)	CW	10 CS graduate students
	MOT	10 CS graduate students
Hornbæk & Frøkjær (2004b)	HE	43 1st-year CS undergraduates
	MOT	44 1st-year CS undergraduates
Jacobsen, Hertzum & John (1998)	UT	2 experienced evaluators 2 beginner evaluators

Study	UEM	#, Type of Evaluators*
Jeffries, Miller, Wharton & Uyeda (1991)	HE	4 HCI researchers
	HPG	1 team of 3 SWEs
	GCW	1 team of 3 SWEs
	UT	1 HF UI specialist
Karat, Campbell & Fiegel (1992)	UT	1 team 2 usability engineers
	UW	12 GUI users, developers & UI specialists
	UW	12 pairs of ″
		(x2 different interfaces**)
Kessner, Wood, Dillon & West (2001)	UT	6 professional teams
		6 professional teams reanalyzed from CUE-2
Law & Hvannberg (2002)	UT	2 experienced evaluators**
	HE	2 experienced evaluators**
Law & Hvannberg (2004b)	UT	1 unspecified
	HE	8 undergraduate students**
		8 graduate students**
		2 post-graduate students**
	GPP	8 undergraduate students**
		8 graduate students**
		2 post-graduate students**
Lesaigle & Biers (2000)	UT-1	5 usability professionals
	UT-2	5 usability professionals
	UT-3	5 usability professionals
Lewis, Polson, Wharton & Rieman (1990)	UT	1 unspecified
	CW	3 creators of CW
		1 other evaluator
Long, Styles, Andre & Malcolm (2005)	UT-1	15 usability students
	UT-2	15 usability students
Mack & Montaniz (1994)	IW	5 usability specialists
		5 non-usability specialists
	CW	5 usability specialists
		5 non-usability specialists
Mankoff, Fait & Tran (2005)	UT	1 team of authors
	EE-1	9 web developers
	EE-2	9 web developers
	EE-3	9 blind users
	Bobby	(automated evaluation)
Molich et al. (1998): CUE-1	UT	3 professional teams
Molich et al. (1999), Molich, Ede, Kaasgaard & Karyukin (2004): CUE-2	UT	6 industry/university teams
		2 student teams
	QUIS	1 industry/university team
Molich & Nielsen (1990), Nielsen & Molich (1990)	HE	77 readers of Computerworld
Nielsen (1992)	HE	31 CS students, no training
		19 usability specialists
		14 usability & IVR specialists

Study	UEM	#, Type of Evaluators*
Nielsen (1994b)	HE	11 usability specialists
	UT	1 team of authors
Nielsen & Molich (1990)	HE	37 CS students in UI course
	HE	34 CS students in UI course
	HE	34 CS students in UI course
Nielsen & Phillips (1993)	UT	1 unspecified
	HE-1	12 experienced usability specialists
	HE-2	10 experienced usability specialists
	HE-3	15 experienced usability specialists
	GOMS	19 upper-division undergraduates
Palacio, Bloch, & Righi (1993)	UT-1	4 temps with 3 months experience**
	UT-2	
Sears (1995)	HW	3 graduate CS students in UI evaluation class
	HE	3 graduate CS students in UI evaluation class
	CW	3 graduate CS students in UI evaluation class
Sears (1997)	HW	6-7 CS graduate students
	HE	6-7 CS graduate students
	CW	6-7 CS graduate students
Sears & Hess (1998)	CW-1	8 students (CS, IS, HCI)
	CW-2	9 students (CS, IS, HCI)
Skov & Stage (2004)	UT	36 teams 1st year students
		8 professional teams reanalyzed from CUE-2
Skov & Stage (2005)	UT-1	8 2nd-year undergraduate CS students
	UT-2	6 2nd-year undergraduate CS students
	UT-1	2 usability experts
Vermeeren, Kesteren & Bekker (2003)	UT	2 unspecified individuals **
	UT	2 unspecified individuals **
	UT	2 unspecified individuals **
Virzi, Sorce & Herbert (1993)	EE	6 usability professionals
	UT-1	1 research assistant**
	UT-2	1 research assistant**
Wright & Monk (1991)	UT	1 team of HF specialists (authors)
		7 teams of just-trained undergrads
	CE	6 teams of just-trained undergrads
	UT	8 teams of undergrads
Zhang, Basili & Shneiderman (1999)	PBI	12 professionals – mixed backgrounds
	HE	12 professionals – mixed backgrounds

*All UEMs are between-subjects except where ** indicates within-subject study. Numbers for UT are numbers of evaluators or teams of evaluators (not users). Abbreviations for types of evaluators: Computer Science (CS), Human-Computer Interaction (HCI), Human Factors/Engineer(HF/E), Information Systems (IS), Software Engineer (SE), User Interace (UI). The table in A.2 explains the UEM abbreviations.

A.2 Description of UEMs

Abbre-viation	Description	Source for Technique	Family
ACW	Automated cognitive walkthrough	Rieman, Davies, Hair, Esemplare, Polson, & Lewis (1991)	W
Bobby	Bobby online tool for testing website accessibility, now available as both WebXACT and Bobby	Bobby 5.0 ("Bobby 5.0," 2005) , WebXACT ("WebXACT," 2004)	—
CE	Cooperative evaluation, think-aloud testing where the evaluator helps the user walk through a prototype	Wright & Monk (1991)	UT
CHP	Principles from Connell & Hammond (1999)	Connell & Hammond (1999)	EE
CUT	Cooperative usability testing	Frøkjær & Hornbæk (2005)	UT
CW	Cognitive walkthrough	Lewis, Polson, Wharton & Rieman (1990), Wharton, Rieman, Lewis & Polson (1994)	W
DP	ISO 9241Dialogue Principles	ISO 9241:10 (1999a)	EE
EC	Ergonomic criteria	Scapin & Bastien (1997)	EE
EE	Expert evaluation with no formal method or method name. This includes studies that use a technique called "heuristic evaluation" but do not specify that they are using Nielsen's heuristic evaluation method (Nielsen, 1994b; Nielsen & Molich, 1990)		EE
GOMS	Goals, operators, methods and selection	John & Kieras (1996)	—
GPP	Gerhardt-Powal's Principles	Gerhardt-Powals (1996)	EE
GCW	Cognitive walkthrough performed by a group		W
HE	Heuristic evaluation	Nielsen (1994b), Nielsen & Molich (1990)	EE
HE-Plus	Heuristic evaluation plus a usability problems profile	Chattratichart & Brodie (2002)	EE
HEG	Heuristic evaluation for groupware	Baker, Greenberg & Gutwin (2002)	EE
HPG	Hewlett Packard's software guidelines	Hewlett-Packard Company's Softguide: Guidelines for usable software (as cited in Jeffries et al., 1991)	EE
HSQ	Questionnaire from the Henderson et al. (1995)	Henderson, Smith, Podd & Varela-Alvarez (1995)	—
HW	Heuristic walkthrough	Sears (1995; 1997)	EE/W
INT	Review of interview transcripts		—

Abbre-viation	Description	Source for Technique	Family
LD	Review of logged data		—
MOT	Metaphors of human thinking	Hornbæk & Frøkjær (2004a; 2004b)	W
PAVE	Programmed Amplification of Valuable Experts	Desurvire & Thomas (1993)	EE
PBI	Perspective-Based Inspection	Zhang, Basili, & Shneiderman (1999), Zhang et al. (1998)	EE
QUIS	Questionnaire for User Interaction Satisfaction	Chin, Diehl & Norman (1988)	—
SUE	Systematic Usability Evaluation inspection	De Angeli, Matera, Costabile & Garzotto (2000)	EE
SMG	Smith & Mosier Guidelines	Smith & Mosier (1986)	EE
UT	Usability testing, think aloud or performance testing (usability testing without think aloud)		UT
UW	Usability walkthrough	Karat, Campbell & Fiegel (1992)	W
VP	Review of verbal protocol transcripts	Henderson, Smith, Podd & Varela-Alvarez (1995)	—
W	Walkthrough – an expert evaluation using specific tasks to focus the evaluation process		

APPENDIX B. Study 1

B.1 IRB Approval for Studies 1, 3

Virginia Tech
VIRGINIA POLYTECHNIC INSTITUTE AND STATE UNIVERSITY

Institutional Review Board
Dr. David M. Moore
IRB (Human Subjects) Chair
Assistant Vice Provost for Research Compliance
CVM Phase II - Duckpond Dr., Blacksburg, VA 24061-0442
Office: 540/231-4991; FAX: 540/231-6033
e-mail: moored@vt.edu

February 19, 2003

MEMORANDUM

TO: Tonya Smith-Jackson ISE 0118
Miranda Capra ISE 0118

FROM: David M. Moore

SUBJECT: IRB EXEMPTION APPROVAL – "Practitioner Survey of Usability Qualities" – IRB # 03-082

I have reviewed your request to the IRB for exemption for the above referenced project. I concur that the research falls within the exempt status. Approval is granted effective as of February 18, 2003.

cc: file
Department Reviewer: RJ Beaton ISE 0118

B.2 Recruiting Letter

Dear Usability Colleague:

I am a Ph.D. candidate in Human Factors at Virginia Tech, studying under Dr. Tonya L. Smith-Jackson in the Department of Industrial and Systems Engineering and Dr. H. Rex Hartson in the Department of Computer Science. My dissertation research focuses on developing a training system for new usability practitioners to improve the effectiveness of their formative usability analyses. As part of this research, I would like to identify the important qualities of usability problem descriptions generated during formative usability analyses.

I ask that you fill out this survey and share your opinions about what constitutes a good or bad usability problem description; it should take 20-30 minutes. Once I have collated the responses, I will send out a second, shorter survey to ask you to rate the qualities I have collected from all of my respondents. I will post the final results of the survey to this mailing list, when they are available.

The survey is available at this URL:
http://hci.ise.vt.edu/~mcapra/UsabilitySurvey/

Please fill out the survey by 10pm EST on <DATE>

Feel free to forward this email to other usability professionals that may be interested in filling out this survey.

Thank you for your consideration,
Miranda Capra

Miranda Capra <mcapra@vt.edu>
Ph.D. candidate and Alexander E. Walter Fellow
Grado Department of Industrial and Systems Engineering
Virginia Tech -- http://www.ise.vt.edu/

B.3 Questionnaire

Usability Survey [redacted]

Section 1 (of 3): Please Tell Us About Yourself

1. Approximate years of usability experience. Please only count years that you have spent conducting usability evaluations of user interfaces in any context (industry, university, etc.).
2. Approximate number of usability evaluations you have conducted:
3. What usability technique(s) do you commonly use when conducting usability evaluations?

   Cognitive Walkthrough
   Expert Review
   Focus Groups
   Heuristic Evaluation
   Heuristic Inspection
   Laboratory Testing With Users
   Naturalistic Observation
   Remote Usability Testing
   Surveys
   User Interviews
   Other:

4. Where do you currently work?

   Industry
   Government
   Military
   University
   Other:

5. What is your job position/title?
6. Do you have a degree in a usability-related field? If so, please describe your highest usability-related degree. If not, please explain how you have gained your knowledge and experience in usability evaluations.

   Please select one
   - Please select one
   - Yes, I have a Bachelors degree, and my area/field/discipline/specialization is:
   - Yes, I have a Masters degree, and my area/field/discipline/specialization is:
   - Yes, I have a Doctoral degree, and my area/field/discipline/specialization is:
   - No, I do not have a related degree, but my usability experience comes from the following:

Submit Section 1

Usability Survey [redacted]

Section 2 (of 3): Usability Problem Description Qualities

Imagine the following scenario:

> You, a usability practitioner, are part of a team assessing the usability of a user interface (website, software, consumer product, etc.) that you may or may not have designed. The ultimate goal of your assessment is to develop a set of improvements to the interface.
>
> You have just completed a formative usability evaluation using usability testing with end-users. Your current goal is to generate a list of usability problem descriptions that summarizes the data collected during the study both from user interactions and from expert observations during the test.
>
> The people reading the descriptions (e.g. usability practitioners, developers, marketing, product development, management, clients) may or may not have usability training. They may use this list for many purposes, such as deciding which problems to fix in the next product release, designing interface changes, or adding features to the software.

Given this scenario, consider the individual usability problem descriptions that you will write. What should you include in these descriptions? What should you avoid? What makes one usability problem description better or worse than another problem description?

If you can think of more qualities than there are spaces below, you can
Add another blank entry

Name of the quality: [ ] is a [Select one / Select one / good / bad] quality.
Describe this quality. What is it? What makes it important? How does it contribute to or detract from your usability evaluation process?

[ ]

Name of the quality: [ ] is a [Select one / Select one / good / bad] quality.
Describe this quality. What is it? What makes it important? How does it contribute to or detract from your usability evaluation process?

[ ]

Select one
Select one
good
bad

Name of the quality: [ ] is a [Select one] quality.
Describe this quality. What is it? What makes it important? How does it contribute to or detract from your usability evaluation process?

[ ]

Select one
Select one
good
bad

Name of the quality: [ ] is a [Select one] quality.
Describe this quality. What is it? What makes it important? How does it contribute to or detract from your usability evaluation process?

[ ]

Select one
Select one
good
bad

Name of the quality: [ ] is a [Select one] quality.
Describe this quality. What is it? What makes it important? How does it contribute to or detract from your usability evaluation process?

[ ]

If you can think of more qualities than there are spaces above, you can
Add another blank entry

Submit Section 2

Usability Survey [redacted]

Section 3 (of 3): Usability Skills

What are important skills for usability practitioners to have when analyzing qualitative data collected from usability testing with end users, when identifying, diagnosing, analyzing and describing usability problems?

What are common mistakes made by inexperienced usability practitioners when identifying, describing and analyzing usability problems found through usability tests with end-users?

What are common mistakes made by all usability practitioners, even experienced practitioners, when identifying, describing and analyzing usability problems found through usability tests with end-users?

Do you have any other comments about the identification, analysis and description of usability problems collected through usability testing with end-users that you would like to share? Or any other comments about this survey?

Submit Section 3

B.4 Questionnaire Responses

Respondent #1

Title: summary information
Type: good
most of the audience you described is not fond of reading. If you do not catch their attention in a summary, they will never get to the individual descriptions

Title: readability
Type: good
Clear concise description needed for actionability

Type: bad
Title: technical or system centric
Those who decide whether or not action will be taken often are not familar with any technical (e.g., software, hfe) or system jargon so write without it.

Type: good
Title: concise
can not expect that audience will read carefully

Title: bounded
Type: good
To what extent can results be extrapolated to others ... and why. Common refrain: this does not apply to me because you did not test ...

Title: descriptive
Type: good
No description of issue, no action. Need to what's issue, what's impact on user, what's business impact, how is the issue classified, how often does it happen, which users did it happen to, etc.

Title: prioritized
Type: good
Action plans require knowing what things are most important to change and why.

Title: participatory
Type: good
Evaluations and evaluation reports done off in a dark corner without the inclusion of key stakeholders do not result in real change. Take a participatory approach to maximize agreement before the report goes out ... reduces prob of free for alls afterwards

Title: user centric
Type: good
Make the report about the user and organized around user work. Helps people understand the work done with the tool and is a good framework for discussion

Title: prescriptive
Type: good
Suggest next steps at both the summary and detailed level. Limit your guessing in recommendations to those that you are sure of and suggest follow-up work if needed to determine between different options.

Title: interconnected

Type: good
Do not treat each issue as an island onto itself. Its part of a larger system and your recommendations need to take that into account.

Important Skills

- human factors
- system thinking
- good writing skills
- good at seeing patterns
- user-centric thinking
- good knowledge of the system being tested
- ability to think at both the macro and micro levels
- understanding of the user work context

Mistakes that everyone makes (even experienced practitioners)

1. Focus on the user but ignore how larger work context relates.
2. Do not get a second opinion.
3. Do not use the evaluation as a tool to get customer thinking about users in a more proactive fashion.

Mistakes that novices make

1. they believe that the solution suggested by the user must be the right solution as opposed to a symptom
2. They get lost at the micro level and fail to see the macro implications.
3. fail to take into account how user and organizational qualities interact with the qualities of the system
4. Write up every little detail but fail to pull out the trends. They do not realize that analysis is more about reducing the data into something understandable rather than documenting every detail.
5. Do not consider their user (report audience) when writing up
6. Recommendations are trite rewordings of Nielsen's heuristics.
7. Make recommendations for specific actions when they do not have enough data to suppport ... should be recommending follow-up studies not bandaids.

Respondent #2

Title: technical HFE language
Type: bad
I've seen problem descriptions that a non-HFE could not understand. Since the purpose of writing it down is to communicate, the report becomes less than usable product.

Title: Prioritized
Type: good
The dev team needs data to decide whether to fix the defect.

Title: Description of impact
Type: good
Again, the dev team needs data to decide whether to fix the defect.

Title: Quantified
Type: good
The audience of the evaluation (if a u-test) will always want to know how many people it happened to (e.g., 3 of 7 users).

Title: Misleadingly quantified

Type: bad
If you're testing 8 users and 2 users ran into an issue, saying that 25% of the users had the issue is misleading, since it implies an extrapolation to the whole population of users that you cannot make based on so small a sample (unless the pop. is exceedingly small).

Title: Solution
Type: good
Dev teams don't want you to beat around the bush -- they want you to tell them how to solve the problem. So make a very specific suggestion for the UI.

Important Skills

- Observation skills, noticing everything without the "oh, that's irrelevant" filter turned on
- Knowledge of the domain, allowing them to make better hypotheses about mental models and causes for errors
- Ability to aggregate and generalize without losing the specificity that keeps the findings accurate
- Ability to draw distinctions between things to seem similar in one way but are dissimilar in an important way

Mistakes that everyone makes (even experienced practitioners)

- Attacking the surface problem rather than getting to the underlying problem
- Concentrating too much on tasks rather than higher-level user goals

Mistakes that novices make

- Giving frequency of occurrency in percentages
- Failing to give the big-picture summary of how usable the product is
- Missing subtle problems that could have been found by watching the users' eyes, reading between the lines, noticing the seemingly irrelevant small details
- Jumping to conclusions too quickly

Respondent 3

Title: Severity
Type: good
How serious is the problem - will it prevent users from accomplishing their goal or can they bypass it. Helps prioritize issues

Title: Scope
Type: good
How many users will be affected by this problem. Helps prioritize issues

Title: Hypothesized cause of problem
Type: good
If users are having trouble finding a web page, what is causing the problem? This has important implications for the proposed solution.

Title: Proposed solution
Type: good
This feature is much less important that the others, but it helps to start discussions with developers. The usability engineer should be careful not to invest too much of their self-esteem in their solutions because there are often other people on the team who have better experience in programming, graphic design, etc. and can come up with better solutions.

Title: Problem description

Type: good
What exactly is the problem?

Respondent 4

Title: impacted audience / scope
Type: good
define what segment of user population is affected by the problem. Helps development team assess important and market impact of fix / nofix decision.

Title: severity
Type: good
provide a graded scale score(high, medium, low, etc) with definitions (e.g. high = unable to complete step of core task) of impact of problem on users ability to complete a task or activity.

Title: heuristic
Type: bad
not necessarily a "bad" quality of a problem statement, but not always a useful one. In context of heuristic reviews, identifying the heuristic that has been violated by the reported error may have some value for usability professionals, but by itself is not useful information for the wider development team.

Title: error type
Type: good
creating a general typology or errors or findings from the data gathering process is essential whenever you have open ended response data. You need a way to characterize, organize, and quantify the findings to understand overall trends and patterns

Important Skills
need to have some kind of defined, repeatable process, so that the same

Mistakes that everyone makes (even experienced practitioners)
losing sight of the objectives of the product and objectives of the analysis, failing to define and ask questions via the analysis that will provide useful pertinent actionalbe information

Mistakes that novices make
losing sight of the objectives of the product and objectives of the analysis, failing to define and ask questions via the analysis that will provide useful pertinent actionalbe information

Respondent 5

Title: valid
Type: good
The problem description (pd) must give a valid view of the real problems experienced by the users/observed by the experts.

Title: visual
Type: good
I prefer to present problems on screen printouts, so the designers can easily see where in the program they appear.

Title: usable
Type: good

The problem description should be usable: designers must know what to do with it. So: diagnostic information and revision proposals make pb's better usable than just problem detections.

Important Skills

- a scientific training and/or attitude (a usability tester does evaluation research). Must be able to distinguish own impressions from user observations. Must be able to discriminate between main and minor problems. To work systematically.
- must be good in dealing with people (users, designers, etc.)
- must have a very good overview of potential usability problems and underlying factors (eg. human factors)

Mistakes that everyone makes (even experienced practitioners)

- forget to take situational context of REAL use into account. Thinks that successful use of program in lab conditions predict a good reception in the market.

Mistakes that novices make

- mix own preferences/dislikes with users' problems
- unsystematical, biased analysis of data (not listening to tapes, but just working from notes or impressions)

Respondent 6

Title: Task
Type: good
The task the problem relates to

Title: Participant
Type: good
Which participant experienced the problem

Title: Principle
Type: good
What principle the usability problem relates to

Title: Recommendation
Type: good
A recommendation for the fix.

Important Skills
Important quotes, screens, number of times the issue repeats itself and recommendations.

Mistakes that novices make
Only stating the problems.

Other Comments
Ways to improve the ways we can move the notes from the logging sheets or video and into a summary table of issues and recommendations.

Respondent 7

Title: brevity
Type: good

The problem description should be brief. Most people will not begin reading a report that looks difficult to read.

Title: level of severity
Type: good
The problem description should indicate the severity of the problem so that other members of the development team can decide whether it should be fixed now or later.

Title: recommended fixes
Type: good
If possible, the problem description should be followed by one or more problem solutions. The potential solutions will help other members of the development team can decide whether it should be fixed now or later.

Title: hypothesized causes
Type: good
The problem description should suggest a cause for the problem being described. Describing the cause of the problem helps the reader evaluate the proposed solutions.

Title: reference to non-test factors
Type: bad
By "test", I mean the evaluation methodology. The problem description should not refer to anything that is not part of the evaluation methodology or enviroment. In other words, the reader should be able to read the evaluation and understand its implications without having to know the history of the product development. This also prevents the evaluation report from containing "I told you so" editorials which have no place in a usability evaluation.

Title: non-pragmatic information
Type: bad
The problem description, cause, or solution should not contain information that would be of interest to HFES members only. If the development team believes that the report writer is a usability expert, then the expert does not have to demonstrate that expertise by using psychology/engineering jargon that must be explained to the reader.

Title: failure to separate well-informed opinions from guesses
Type: bad
The problem description, proposed solution, or hypothesized cause should be qualified by the usability expert's level of certainty. It is better to say "I don't know" than to lead people to believe that you are certain about something.

Important Skills

- mastery of the work domain that the system being designed is intended to support
- an full understanding of the roles played by users of the systems, their goals, their skills and abilities
- parametric and non-parametric data analysis
- observational skills, i.e., the ability to understand what is being said compared to how it is said
- understanding of how organizations work - this may be useful only to people developing software for large organizations

Mistakes that novices make

- emphasizing the trivial - not all usability problems are important
- misunderstanding the role of usability - usability is one of several factors affecting system design
- inability to take criticisim without taking it personally – som

Mistakes that novices make

The problems I described for inexperienced practitioners are problems for all practioners. I don't know that I that experience is the distinction that is important. The differences I've observered, between practitioners, have been based in skills rather than experience. Most of the differences arise from a lack of basic training or education, on the part of the less skilled practitioner.

Respondent 8

Title: Users Goal
Type: good
Provide a clear and unambiguous description of the goal that the user was trying to achieve

Title: User's outcome
Type: good
Describe the eventual outcome that the user achieved

Title: The primary usability problem(s)
Type: good
Details the precise problem that the user encountered. Describe this in terms of the navigation flow and/or interaction architecture.

Title: Severity Ratings
Type: good
Include ratings of problem severity (pref 1-2 scale) so that design/Engineering knows where to focus resources

Title: Provide Recommendations
Type: good
List recommended fixes/design improvements. Provide specific details.

Title: Create a check-list
Type: good
Create a "problem-fixing" check list so that you can track development of changes. Arrange with engineering/design for you to stay in the loop until problems are either fixed or you hae explanation as to why fix not possible.

Important Skills
A thorough background and understanding ogf human psychology, and experimental design/data analysis

Mistakes that everyone makes (even experienced practitioners)
Failing to assign severity ratings to the problems

Mistakes that novices make
Failure to control variables in the test design; inability to know when they are looking at a problem; cursory descriptions and over use of "cut and paste"

Respondent 9

Title: State What Happened
Type: good
Explain what happened. Bring everyone to a common place.

Title: State Why This Is Not Correct

Type: good
Level the playing field between experts and non-experts.

Title: State What Should be Done to Correct
Type: good
Without this, you are no good as a UE. Lots of folks can identify problems, the key is to be able to propose solutions.

Title: Be Concise
Type: good
Keep it concise - otherwise people 1) ignore it or 2) read too much into it.

Title: Be Precise
Type: good
Language is very tricky - especially if you're working with non-english speaking developers - define your terms and be very certain you abide by your definitions.

Important Skills
An open mind. It is far too easy to become drawn into the tunnel of self-righteousness. Do not presume. Try to establish a pattern but be ready to discard that theory if another develops. Always try to develop multiple theories.

Mistakes that everyone makes (even experienced practitioners)
1. Believing that users are designers and taking their input literally.
2. Inability to abstract a solution from fuzzy data
3. Statistical malfeasance
4. Imprecise use of language
5. Verbosity
6. Good experimental design.

Mistakes that novices make
1. Believing that users are designers and taking their input literally.
2. Falling in love with their designs.
3. Getting defensive when users critisize a design.
4. Failure to recognize the big issues from the small ones.
5. Inability to abstract a solution from fuzzy data.
6. Statistical malfeasance
7. Imprecise use of language
8. Verbosity.

Other Comments
Overall, usability (as practiced in the wild) is really more of an art than a science. People are sloppy as a general rule about their behaviors and interpreting the behaviors of others. This needs to change...

Respondent 10

Title: Supports finding way through task
Type: good
Helps user recognize first step and subsequent steps with no trial and error. This is important because it reduces the cognitive strain on figuring out what to do and where to go in the interface.

Title: Helps user recover from errors
Type: good

User is able to easily recover from the error. This is important because it prevents the user from making a catastrophic error.

Title: System interaction is similar to other models of interaction
Type: good
When user's can interact with a system that is similar to other conventions and models, it provides a more efficient interaction experience.

Important Skills
Familiarity with guidelines and standards.
Descriptive stats.

Mistakes that everyone makes (even experienced practitioners)
Discounting the severity/impact of the problem because they think the user is stupid.

Mistakes that novices make
Focusing on "emotional" problems or problems of preference.

Respondent 11

Title: Task
Type: good
Describes the task the user/reviewer performed that uncovered the usability problem. This provides a context for the description of the problem and helps others recreate the problem.

Title: Objective data
Type: good
When available, it is important to present objective data about the task up front. This helps to set the tone that the problem can be quantified -- it isn't "fuzzy" and/or only one person's view of the user interface.

Title: Critical incidents or Theory/Heuristics-based issues
Type: good
Many people have difficulty "seeing" a problem only from objective data. Presenting critical incidents in an anecdotal manner as well as theory- and heuristics-based usability issues is important to fill in the details of the problem and illustrate its effects. This part of the description can also be used to impart some principles of good user interface design for those who are actually reading the description (i.e., the people who are most interested in usability).

Title: Impact
Type: good
Estimating the impact of the problem is key. If the usability expert ranks usability problems this helps them classify the severity of problems. For other cross-functional team members, this is mostly what they care about -- how's it going to affect the overall project.

Title: Fix
Type: good
Once the usability problem has been explained and justified, it's important to offer a solution. It's also good to duplicate all of the fixes in a separate list for easy reference, preferably grouped by severity.

Title: Vague suggestions
Type: bad
Being too general about how to fix the problem makes it difficult to adopt the solution. Well defined solutions are easier to implement.

Title: Presenting opinions
Type: bad
Without specific justification, it leaves one open to the charge of "that's just your opinion -- I'm just as much a user as you. " Make it clear that the problem isn't just one person's opinion.

Important Skills

- Understanding the principle that underlies the observation
- seeing how the same problem may manifest itself in multiple ways
- based on the problem, inventing a solution that fixes that specific problem
- culling which observations are problems for some users, but its fix would create problems for a much larger number of users
- understanding the different problems encountered by novice and experienced users

Mistakes that everyone makes (even experienced practitioners)

- estimating the impact of a problem on systems that doesn't exist yet
- failing to recognize that increased user control usually means more complexity
- getting default settings right -- most users won't change them
- providing designs to developers which can be implemented with a minimum of consultation

Mistakes that novices make

- thinking that describing the problem makes it self-evident that it is a problem and how to fix it -- that usability somehow sells itself
- thinking that if users had a problem it automatically means there is a usability problem. While that's the case most of the time, one shouldn't become an extremist -- to err is human. Usability seeks to minimize design-induced mistakes, not correct human nature.
- estimating how long a participant will take to perform a set of tasks

Other Comments
Although identification, analysis, and description are key to writing a good report, the usability professional needs to advocate for the user and seeing changes made to fix the usability problems. It's one thing to describe the problem, another to offer a solution for it, and still another to get it implemented when one is competing for resources.

Respondent 12

Title: Component
Type: good
You need to know which component of the product the problem is associated with (page, dialog box, field, etc.). This is important so the developers know what to fix.

Title: Granularity
Type: good
The granularity of a problem is the "size" of the problem or level of detail. A problem might be fine-grained (problem with a field) or large grained (bad navigation architecture). Problem reporting needs to address different levels of granularity for different audiences. Those fixing the problem need to see fine-grained problem reports. Presenting the wrong level of granularity can impeded fixes.

Title: Severity
Type: bad
Severity is the amount of loss in time, data, reputation, that a problem will cause. Knowing how severe a problem is sometimes requires domain knowledge which the usability engineer does not have.

Title: Frequency

Type: bad
The number of times that a problem occurs or is predicted to occur based on a sample of participants.

Title: Breadth of impact
Type: bad
Breadth is the extent to which a problem affects many things. For example, a bad control might affect the product in many places.

Important Skills
- Domain knowledge
- Sense of granularity
- Good language skills
- Description that is factual

Mistakes that everyone makes (even experienced practitioners)
- Not clear about the impact
- State multiple problems as a single problem
- Interpret rather than describe a problem

Mistakes that novices make
- Not clear about the impact
- State multiple problems as a single problem
- Interpret rather than describe a problem
- No clear problem definitions

Respondent 13

Title: exemplars
Type: good
The problem description has examples of the problem based on user performance (or violation of guideline)

Title: taxonomic
Type: good
the problems are grouped together so that possible solution covers multiple problems where appropropriate. this is difficult to do well, but important in formative evaluation.

Title: vague
Type: bad
description is too high level or not obvious why bad or how it might be fixed.

Title: suggestive of a solution
Type: good
purpose of formative is to suggest how a problem might be solved.

Important Skills
- ability to synthesize results
- ability to explain outliers and whether these are important or not

Mistakes that novices make
- wrong level: too specific or too general
- no solution in mind.
- contradictory

Respondent 14

Title: Functional
Type: good
How completely a user can achieve their goals/objectives with the interface.

Type: good
How fast and accurately a user can complete their tasks with the interface, both initially and when it's mastered.
Title: Efficient

Title: Learnable
How the interface helps the user to become competent using the product, both initially and when it's mastered.

Title: Forgiving
How the interface helps the user prevent errors from occurring and how it helps users recover from errors should they occur.

Title: Appealing
How pleasant and satisfying the interface is to use and how confident users are when interacting with the interface.

Important Skills

- To be able to group problems into categories that address the different aspects of the interface and how the problems identified affect the user's experience with the interface - e.g., navigation, terminology, interaction, architecture, other, etc
- To be able to assign a severity rating to each problem and how the severity will affect the interface.
- To provide statistics on success or failure rates, as well as ROI when it comes time to address the issues.
- To be able to translate all of this into something the project manager and client can understand.

Mistakes that everyone makes (even experienced practitioners)
Coming up with solutions too fast before they've taken time to look at the whole picture.

Mistakes that novices make
Not correctly interpreting what the user's expressed or experienced during the study.

Respondent 15

Title: traceability
Type: good
Can be traced to one or more observations in a test. Includes references so others can go to source material (e.g., recordings) and get more detail or see for themselves.

Title: non-judgemental
Type: good
State what happened and your interpretation. Don't judge the design.

Title: evocative

Type: good
Communicates the essence and impact of the issue clearly & succinctly.

Title: Describing only an alternate design and not clearly identifying the breakdown.
Type: bad
A problem description should not simply give an alternative design. It should focus on what is wrong with the current design. An alternate design idea is optional.
design (only)

Respondent 16

Title: Severity
Type: bad
The degree to which the interface inhibits teh ability of the average user to succeed at the desired task.

Title: Issue
Type: bad
The usability issue identified.

Title: Recommendation
Type: good
What suggested chagnes should be considered as a means to correct the idetified issue

Title: Issue
Type: good
Sometimes its just as improtant to identify what was done correctly and point it out in the issues column so taht the desingers can leverage that success.

Important Skills
Being able to see beyond the data and rethink the problem. Too often designers solve page level problems when in fact the individual screens may have little to do with the real problme that users are having. For instance, you can redesign a report generator, but maybe the users need a graph and not a report. SO no matter how well you desing teh reprot generator, it'll be the wring solution.

Mistakes that everyone makes (even experienced practitioners)
Even experienced practitioners misidentify what the real task is for the users and focus ont he screen level issues. Most usability professionals don't do user research and task analysis as part of every evaluation. This is a mistake. You cannot give a representative evaluation or prepare a representative test without user observations and tasks analysis.

Mistakes that novices make
Focusing on the button clicks and not ont he psychology of the user's actions. Also, designing at the screen level without having a clear vision of the big picture, as mentioned in the previos question.

Respondent 17

Type: good
how important is this problem, how important is it to fix it
Title: priority/rank/criticality

Title: descriptiveness
Type: good

what is the problem, use image if possible to explain where and how problem arises

Title: impact/cost
Type: good
what will happen if problem persists, in terms that are relevant to client - e.g. fewer completed orders, more errors in inventory, more staff for data entry

Title: suggested fix
Type: good
obvious/standard ways to fix problem if possible - not all problems are easy to fix!

Important Skills

- pragmatism!
- succinctness
- verbal/written fluency/persuasivness

Mistakes that everyone makes (even experienced practitioners)
not understanding technical constraints & possibilities

Mistakes that novices make
ignoring/not understanding client drivers, priorities and constraints, usability for its own sake

---

Respondent 18

Title: Conceptual Model
Type: good
Major organization of the software. Is it clear and understandable. Does it help the user understand and organize their work?

Title: Efficient Navigation
Type: good
Mouse and Keyboard navigation should be possible. Should be efficient - few keystrokes, mouse movements appropriate to fitt's law.

Title: efficient data entry
Type: good
Use standard interaction elements. Provide edit masks to indicate expected field format. Provide data entry assistance tools such as a calendar control where appropriate.

Title: System efficiency
Type: good
Short response time from the system. Appropriate feedback for longer wait times.

Title: Screen layout
Type: good
Follows a layout grid to align screen elements. Appropriate use of whitespace - not too much or to little. Appropriate grouping and labelling.

Title: language
Type: good
text is clear, concise, and in user's language. Not too many abbreviations or acronyms. Labels are brief but complete. Potentially new terms have definitions available.

Title: Consistency

Type: good
To the level appropriate, the system is consistent both within itself and with other applications the user will likely use in conjunction with the app. The same element should look and act the same wherever it occurs.

Title: user assistance
Type: good
online help, tips, printed references, etc. are provided as needed in a context appropriate way.

Title: accessibility
Type: good
Section 508 and/or W3C WAI guidelines are supported to make the product useful to people with a range of physical abilities.

Important Skills
Knowledge of user's context of use - preferably obtained through ethnographic field studies. Knowledge of a wide range of devices, their usability attributes, and their success/failure. Knowledge of industry standards, guidelines, and best practices. Knowledge of the development environment capabilities - e.g. html, visual basic, java, etc.

Mistakes that everyone makes (even experienced practitioners)
Some describe issues without providing potential solutions. Perception can be "its easy to criticize but difficult to solve". Be able to do design, not just evaluate. Mock up potential solutions and include screen shots in the document. Note that there can be a range of successful solutions, not just the one a famous company uses or your favorite stock solution.

Mistakes that novices make
Know the audience you are communicating to - they are the user's of your information. A long report is OK, but it must have a short executive summary or powerpoint that hits the key points. Be able to describe the key usability issues in 2 minutes or less. If you are working in industry, this isn't a science project or a school paper. It should't be in APA format. Try to use a standard company template and writing style that is familiar to your audience.

Respondent 19

Title: State the task and context
Type: good
For tests in lab, provide readers with information about what users were asked to do, and in what context they were doing it. (For evaluations, what would users do and why.)

Title: State the success rate with reasons for lack of success
Type: good
List how many users didn't succeed, and why users couldn't (or wouldn't) perform the task successfully.

Title: Provide general recommendations for changes
Type: good
Provide recommendations for improving the design, without necessarily providing specific designs. Refer to good usability practices in the recommendations.

Type: good
Understanding who the real users are is sometimes a problem. If you do not test with a sample of truly representative users, results may be tainted. Would real users have more domain knowledge, and therefore make less mistakes, or vice versa?

Important Skills

- Knowledge of good usability practices
- Knowledge of common usability problems
- Ability to get to the real issue when users struggle; the problems aren't always as obvious as they might seem.

Mistakes that novices make

- Not realizing the real issue when users struggle. This skill often only comes from experience, or a usability person's own experience in related situations.
- Also, forming conclusions without collecting enough data. Don't necessarily make judgements based on one or two users.

APPENDIX C. Study 2

C.1 IRB Approval for Study 2

Virginia Tech
VIRGINIA POLYTECHNIC INSTITUTE AND STATE UNIVERSITY

Institutional Review Board

Dr. David M. Moore
IRB (Human Subjects)Chair
Assistant Vice Provost for Research Compliance
CVM Phase II- Duckpond Dr., Blacksburg, VA 24061-0442
Office: 540/231-4991; FAX: 540/231-6033
email: moored@vt.edu

DATE: February 10, 2004

MEMORANDUM

TO: Tonya L. Smith-Jackson Industrial and Systems Engineering 0118
Miranda Capra ISE 118

FROM: David Moore

SUBJECT: **IRB Exempt Approval:** "Utilizing Usability Problem Descriptions to Measure Formative Usability Evaluation Effectiveness" IRB # 04-042

I have reviewed your request to the IRB for exemption for the above referenced project. I concur that the research falls within the exempt status. Approval is granted effective as of February 9, 2004.

cc: File
Department Reviewer R.J. Beaton ISE 0118

C.2 Recruiting Letter

Dear Usability Colleague:

I previously sent to this list a request to fill out a survey to identify the important qualities of a good usability problem description generated as part of a formative usability evaluation. However, the list that I collected is much too detailed.

Please visit the following URL and do a card sort on this list of UPD qualities. This will help me develop a shorter list of more general UPD qualities.

http://hci.ise.vt.edu/~mcapra/UsabilitySurvey/

Some people have taken 20 minutes to complete the survey, some have taken an hour. However, you can save your work and complete the survey in multiple sessions, if you need to.

Feel free to share this URL with other members of the usability community that may be interested in filling out this survey.

I will post the results of this survey to this list, when available.

BACKGROUND: I am a Ph.D. candidate in Human Factors at Virginia Tech, studying under Dr. Tonya L. Smith-Jackson in the Department of Industrial and Systems Engineering and Dr. H. Rex Hartson in the Department of Computer Science. My dissertation research focuses on methods for comparing usability evaluation methods (UEMs) and for measuring the output of UEMs, with the eventual goal of trying to understand what contributes to a good evaluation and what makes a good usability practitioner. This set of surveys is the first part of a series of studies.

Thank you for your consideration,
Miranda

Miranda Capra <mcapra@vt.edu>
Ph.D. candidate and Alexander E. Walter Fellow
Grado Department of Industrial and Systems Engineering
Virginia Tech -- http://www.ise.vt.edu/

C.3 Questionnaire

Usability Survey [redacted]

Section 1 of 3: Please Tell Us About Yourself

1. Approximate years of usability experience. Please only count years that you have spent conducting usability evaluations of user interfaces in any context (industry, university, etc.).
   [ ]
2. Approximate number of usability evaluations you have conducted:
   [ ]
3. What usability technique(s) do you commonly use when conducting usability evaluations?

   Cognitive Walkthrough
   Expert Review
   Focus Groups
   Heuristic Evaluation
   Heuristic Inspection
   Laboratory Testing With Users
   Naturalistic Observation
   Remote Usability Testing
   Surveys
   User Interviews
   Other: [ ]

4. Where do you currently work?

   Industry
   Government
   Military
   University
   Other: [ ]

5. What is your job position/title?
   [ ]
6. Do you have a degree in a usability-related field? If so, please describe your highest usability-related degree. If not, please explain how you have gained your knowledge and experience in usability evaluations.

   Please select one
   - Please select one
   - Yes, I have a Bachelors degree, and my area/field/discipline/specialization is:
   - Yes, I have a Masters degree, and my area/field/discipline/specialization is:
   - Yes, I have a Doctoral degree, and my area/field/discipline/specialization is:
   - No, I do not have a related degree, but my usability experience comes from the following:

   [ ]

Submit Section 1

Usability Survey [redacted]

Section 2 of 3: Usability Problem Description Qualities[2]

Imagine the following scenario:

> You, a usability practitioner, are part of a team assessing the usability of a user interface (website, software, consumer product, etc.) that you may or may not have designed. The ultimate goal of your assessment is to develop a set of improvements to the interface.
>
> You have just completed a formative usability evaluation using usability testing with end-users. Your current goal is to generate a list of usability problem descriptions that summarizes the data collected during the study both from user interactions and from expert observations during the test.
>
> The people reading the descriptions (e.g. usability practitioners, developers, marketing, product development, management, clients) may or may not have usability training. They may use this list for many purposes, such as deciding which problems to fix in the next product release, designing interface changes, or adding features to the software.

Given this scenario, consider the individual usability problem descriptions that you will write. What should you include in these descriptions? What should you avoid? What makes one usability problem description better or worse than another problem description?

The following list of UPD qualities was collected through a previous survey, but the list below is much too detailed. Please perform a card sort on these items to help me develop a shorter list of UPD qualities.

Instructions

- Please organize the following items into 5-20 categories
- Please give each category a descriptive name
- It's OK, if necessary, to:
  - put one item into multiple categories (mark with DUP or similar)
  - make ONE level of sub-categories
  - add a few of your own items (mark with ADDED or similar)
- Note that these are qualities of INDIVIDUAL problem descriptions, and not of an entire usability report (which includes many other things than just UPDs)
- You may cut-and-paste the list into your favorite text editor and then paste the results in the box below when you are done.

---

[2] This is an approximation of the web page text and layout for Section 2. Actual text may have differed somewhat. The 70 items were ordered randomly. The textbox for the 70 items was smaller and used vertical scrolling.

EXAMPLE OF SORTED ITEMS - Ice Cream

CATEGORY: Fruit flavors
Strawberry
Raspberry
Blueberry

CATEGORY: Chocolate-based
Chocolate
Double chocolate chip (DUP)
Chocolate fudge brownie
Chocolate fudge swirl

CATEGORY: Ice cream with chips
SUBCATEGORY: Chocolate
Chocolate chip
Double chocolate chip (DUP)
Mint chocolate chip
Chocolate chunk
SUBCATEGORY: Butterscotch
Butterscotch chip (ADDED)
Maple butterscotch chip (ADDED)

Avoid dictating a specific solution – don't alienate other team members
Avoid guessing about the problem cause or user thoughts
Avoid jargon or technical terms
Avoid judging the system or decisions made by other team members
Avoid just listing the heuristic violated; explain why it's is a problem
Avoid misleading statistics and presentation of results
Avoid pointing fingers or assigning blame
Avoid so much detail that no one will want to read to description
Avoid your own opinions or subjective statements
Be clear and precise
Be concise, avoid wordiness
Be evocative, help the reader understand/sympathize with what happened
Be pragmatic/practical; avoid theories/jargon that non-HCI people wouldn't appreciate
Be specific enough about a solution to be helpful, without jumping to a design conclusion
Define any terms that you use
Describe a potential solution to the problem
Describe advantages and disadvantages of alternative solutions
Describe critical incidents
Describe exactly which system components are affected/involved
Describe how many users that experienced the problem
Describe how the interaction architecture contributed to the problem
Describe the breadth of components of the system involved in the problem
Describe the business effects – support costs, time loss, etc.
Describe the eventual outcome of the user's actions
Describe the impact of the problem
Describe the importance of the task the user was performing
Describe the main usability issue involved in this problem
Describe the problem's cause

Describe the user groups that could be affected during system usage
Describe the user groups that were affected during testing
Describe the user's navigation flow through the system
Describe the users' subjective reactions
Described observed behaviors
Don't dictate a specific solution – don't jump to conclusions
Don't guess about a good design solution
Don't use vague terms and descriptions; be concrete
Include a category from an outside taxonomy/classification scheme
Include definitions of severity/importance/impact to avoid confusion
Include objective data from the study to support your arguments
Include several alternate solutions, if possible
Include sufficient detail to understand exactly what happened
Include the cause of the problem
Include time spent on the task
Describe limitations of your domain knowledge
Make sure that you are complete in your descriptions
Make sure that you impart good design principles
Make sure the description is readable/understandable
Mention follow-up work that should be done to understand the problem
Mention good design elements and successful user interactions
Mention usability practices/previous research to back your explanations
Mention usability practices/previous research to back your suggestions
Mention how often the problem occurred during testing
Mention how often the problem will occur during usage
Mention the context of the problem, such as the user's task
Mention the number of task attempts
Mention the testing context (laboratory, field study, inspection, etc.)
Mention whether or not this problem should be fixed
Mention whether the problem was user reported or experimenter observed
Mention whether the user succeeded or failed at the task
Provide traceability of problems to observed data
Use a user-centric perspective
Use anecdotes to make the problem seem real, promote sympathy
Use only facts from the study, rather than your opinions or guesses
Use pictures/captures of interface to describe a suggested solution
Use pictures/screen shots of the user interface to describe the problem
Use precise terminology
Use quantitative data to support your arguments
Use specific examples from the study
Use supporting data
Use user-centered descriptions rather than system-centric

Section 3 of 3: Conclusion

Do you have any other comments you would like to share with me about the survey you have just completed, usability problem descriptions, or about the formative usability process?

Submit Section 3

C.4 Study 2: Factor Analysis Eigenvalues

	Eigenvalue	Difference	Proportion	Cumulative
1	9.31	3.24	0.133	0.13
2	6.08	0.71	0.087	0.22
3	5.37	2.32	0.077	0.30
4	3.05	0.27	0.044	0.34
5	2.78	0.22	0.040	0.38
6	2.56	0.28	0.037	0.42
7	2.28	0.09	0.033	0.45
8	2.19	0.25	0.031	0.48
9	1.93	0.19	0.028	0.51
10	1.74	0.04	0.025	0.53
11	1.70	0.17	0.024	0.56
12	1.53	0.12	0.022	0.58
13	1.41	0.10	0.020	0.60
14	1.31	0.11	0.019	0.62
15	1.20	0.02	0.017	0.64
16	1.19	0.09	0.017	0.65
17	1.10	0.08	0.016	0.67
18	1.02	0.04	0.015	0.68
19	0.98	0.01	0.014	0.70
20	0.97	0.05	0.014	0.71
21	0.92	0.04	0.013	0.72
22	0.88	0.03	0.013	0.74
23	0.85	0.03	0.012	0.75
24	0.82	0.04	0.012	0.76
25	0.78	0.07	0.011	0.77
26	0.71	0.02	0.010	0.78
27	0.69	0.02	0.010	0.79
28	0.68	0.04	0.010	0.80
29	0.63	0.01	0.009	0.81
30	0.62	0.03	0.009	0.82
31	0.59	0.02	0.008	0.83
32	0.57	0.02	0.008	0.84
33	0.55	0.01	0.008	0.84
34	0.54	0.02	0.008	0.85
35	0.51		0.007	0.86

APPENDIX D. Study 3

D.1 Recruiting Letter

Dear Usability Colleague:

Through previous research I have created a list of 10 general issues that should be addressed when describing a usability problem description. Please take 20-30 minutes to visit the following URL and let me know which of these are the most important.

http://hci.ise.vt.edu/~mcapra/UsabilitySurvey/

Feel free to share this URL with other members of the usability community that may be interested in filling out this survey.

I will post the results of this survey to this list, when available.

BACKGROUND: I am a Ph.D. candidate in Human Factors at Virginia Tech, studying under Dr. Tonya L. Smith-Jackson in the Department of Industrial and Systems Engineering. My dissertation research focuses on methods for comparing usability evaluation methods (UEMs) and for measuring the output of UEMs.

The first part of this research was a survey to collect qualities of a good usability problem description. The second part was a card sort to group the items in this detailed list into general, higher-level categories. The current 10-item list was developed from the card sort responses.

Thank you for your consideration,
Miranda Capra

Miranda Capra <mcapra@vt.edu> http://filebox.vt.edu/users/mcapra/
Ph.D. candidate, Human Factors and Ergonomics http://hfec.vt.edu/
Assessment and Cognitive Ergonomics Laboratory http://ace.ise.vt.edu/
Grado Dept. of Industrial & Systems Engineering http://www.ise.vt.edu/

D.2 Questionnaire

Usability Survey [redacted]

Section 1 of 3: Please Tell Us About Yourself

1. Approximate years of usability experience. Please only count years that you have spent conducting usability evaluations of user interfaces in any context (industry, university, etc.).
2. Approximate number of usability evaluations you have conducted:
3. What usability technique(s) do you commonly use when conducting usability evaluations?

   Cognitive Walkthrough
   Expert Review
   Focus Groups
   Heuristic Evaluation
   Heuristic Inspection
   Laboratory Testing With Users
   Naturalistic Observation
   Remote Usability Testing
   Surveys
   User Interviews
   Other:

4. Where do you currently work?

   Industry
   Government
   Military
   University
   Other:

5. What is your job position/title?
6. Do you have a degree in a usability-related field? If so, please describe your highest usability-related degree. If not, please explain how you have gained your knowledge and experience in usability evaluations.

   Please select one
   - Please select one
   - Yes, I have a Bachelors degree, and my area/field/discipline/specialization is:
   - Yes, I have a Masters degree, and my area/field/discipline/specialization is:
   - Yes, I have a Doctoral degree, and my area/field/discipline/specialization is:
   - No, I do not have a related degree, but my usability experience comes from the following:

Submit Section 1

Usability Survey [redacted]

Section 2 of 3: Usability Problem Description Qualities

Imagine the following scenario:

> You, a usability practitioner, are part of a team assessing the usability of a user interface (website, software, consumer product, etc.) that you may or may not have designed. The ultimate goal of your assessment is to develop a set of improvements to the interface.
>
> You have just completed a formative usability evaluation using usability testing with end-users. Your current goal is to generate a list of usability problem descriptions that summarizes the data collected during the study both from user interactions and from expert observations during the test.
>
> The people reading the descriptions (e.g. usability practitioners, developers, marketing, product development, management, clients) may or may not have usability training. They may use this list for many purposes, such as deciding which problems to fix in the next product release, designing interface changes, or adding features to the software.

Given this scenario, consider the individual usability problem descriptions that you will write. What should you include in these descriptions? What should you avoid? What makes one usability problem description better or worse than another problem description?

Through previous research I have collected the following 10 qualities of a description of a usability problem (in random order):

1. Describe the cause of the problem
2. Describe a solution to the problem
3. Justify the problem with data from the study
4. Be clear and precise while avoiding wordiness and jargon
5. Consider politics and diplomacy when writing your description
6. Describe observed user actions
7. Describe the impact and severity of the problem
8. Be professional and scientific in your description
9. Help the reader sympathize with the user
10. Describe your methodology and background

Instructions

I would like you to help me determine which of these 10 qualities are the most important for a good usability problem description.

- Read each quality and its description
- Rate your impression of the quality for each of the adjectives provided
    - Don't agonize over this - I'm most interested in your first impression
    - The center position indicates neither adjective, or both adjectives equally

- **NOTE**
  - These are qualities of a description of an **individual** usability problem, not a complete usability report
  - Assume the problems were found using usability testing **with end-users**
  - Assume these descriptions are your **only means of communication**, and you cannot supplement with personal communications

---

1. 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.

When a usability practitioner writes a description of a single usability problem, addressing this quality of a problem description is generally ...

required	O	O	O	O	O	O	O	optional
helpful	O	O	O	O	O	O	O	harmful
easy	O	O	O	O	O	O	O	difficult
irrelevant	O	O	O	O	O	O	O	relevant

2. 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.

When a usability practitioner writes a description of a single usability problem, addressing this quality of a problem description is generally ...

required	O	O	O	O	O	O	O	optional
helpful	O	O	O	O	O	O	O	harmful
easy	O	O	O	O	O	O	O	difficult
irrelevant	O	O	O	O	O	O	O	relevant

3. Justify the problem with data from the study, both quantitative and qualitative. Include how many users experienced the problem and how often; task attempts, time and success/failure; and critical incident descriptions. Provide traceability of the problem to observed data.

When a usability practitioner writes a description of a single usability problem, addressing this quality of a problem description is generally ...

required	O	O	O	O	O	O	O	optional
helpful	O	O	O	O	O	O	O	harmful
easy	O	O	O	O	O	O	O	difficult
irrelevant	O	O	O	O	O	O	O	relevant

4. 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 people will appreciate. Avoid so much detail that no one will want to read the description.

When a usability practitioner writes a description of a single usability problem, addressing this quality of a problem description is generally ...

required	O	O	O	O	O	O	O	optional
helpful	O	O	O	O	O	O	O	harmful
easy	O	O	O	O	O	O	O	difficult
irrelevant	O	O	O	O	O	O	O	relevant

5. Consider politics and diplomacy when writing your description. Avoid judging the system, criticizing decisions made by other team members, pointing fingers or assigning blame. Point out good design elements and succesful user interactions. Be practical, avoiding theory and jargon.

When a usability practitioner writes a description of a single usability problem, addressing this quality of a problem description is generally ...

required	O	O	O	O	O	O	O	optional
helpful	O	O	O	O	O	O	O	harmful
easy	O	O	O	O	O	O	O	difficult
irrelevant	O	O	O	O	O	O	O	relevant

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.

When a usability practitioner writes a description of a single usability problem, addressing this quality of a problem description is generally ...

required	O	O	O	O	O	O	O	optional
helpful	O	O	O	O	O	O	O	harmful
easy	O	O	O	O	O	O	O	difficult
irrelevant	O	O	O	O	O	O	O	relevant

7. 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.

When a usability practitioner writes a description of a single usability problem, addressing this quality of a problem description is generally ...

required	O	O	O	O	O	O	O	optional
helpful	O	O	O	O	O	O	O	harmful
easy	O	O	O	O	O	O	O	difficult
irrelevant	O	O	O	O	O	O	O	relevant

8. Be professional and scientific in your description. Use only facts from the study, rather than opinions or guesses. Back your findings with sources beyond the current study (such as an external classification scheme), proven usability design principles, and previous research.

When a usability practitioner writes a description of a single usability problem, addressing this quality of a problem description is generally ...

required	O	O	O	O	O	O	O	optional
helpful	O	O	O	O	O	O	O	harmful
easy	O	O	O	O	O	O	O	difficult
irrelevant	O	O	O	O	O	O	O	relevant

9. Help the reader sympathize with the user by using a problem description that is evocative and story-like. Make sure the description is readable and understandable. Use user-centric language rather than system-centric. Be complete while avoiding excessive detail.

When a usability practitioner writes a description of a single usability problem, addressing this quality of a problem description is generally ...

required	O	O	O	O	O	O	O	optional
helpful	O	O	O	O	O	O	O	harmful
easy	O	O	O	O	O	O	O	difficult
irrelevant	O	O	O	O	O	O	O	relevant

10. Describe your methodology and background. Describe how you found this problem (field study, lab study, expert evaluation, etc.). Describe the limitations of your domain knowledge. Describe the user groups that were affected and the breadth of system components involved.

When a usability practitioner writes a description of a single usability problem, addressing this quality of a problem description is generally ...

required	O	O	O	O	O	O	O	optional
helpful	O	O	O	O	O	O	O	harmful
easy	O	O	O	O	O	O	O	difficult
irrelevant	O	O	O	O	O	O	O	relevant

Submit Section 2

Usability Survey [redacted]

Section 3 of 3: Conclusion

Do you have any other comments you would like to share with me about the survey you have just completed, usability problem descriptions, or about the formative usability process?

Submit Section 3

D.3 Summary of SAS Programs and Outputs

D.3.1 Supplementary ANOVA for left-right order effects

```
proc mixed;
        class Participant LeftRightOrder;
        model rating = Participant LeftRightOrder;
        repeated                                / subject= Participant type=cs;
        by AdjectivePair;
run;
```

	***F*(1, 72)**	***p***
Easy/Difficult	0.01	.92
Helpful/Harmful	0.07	.80
Relevant/Irrelevant	2.10	.15
Required/Optional	1.08	.30

D.3.2 Supplementary ANOVA for effects due to adjective position (1-4)

```
proc mixed;
        class Participant AdjectivePosition;
        model Rating =    Participant AdjectivePosition;
        repeated                                / subject= Participant type=cs;
        lsmeans AdjectivePosition / adjust=tukey;
        by AdjectivePair;
run;
```

	***F*(3, 70)**	***p***
Easy/Difficult	0.61	.61
Helpful/Harmful	0.76	.52
Relevant/Irrelevant	2.10	.11
Required/Optional	4.27	.008

Post-Hoc Comparisons for position of Required/Optional:
Differences of Least Square Means

Position	**Comparison**	**Estimate**	**Standard Error**	***t*(70)**	***p***	**Tukey-Kramer adjusted *p***
1	2	0.29	0.30	0.98	.33	.76
1	3	0.78	0.23	3.35	.001	.007*
1	4	0.14	0.24	0.61	.55	.93
2	3	0.49	0.29	1.67	.10	.35
2	4	-0.15	0.30	-0.49	.63	.96
3	4	-0.63	0.23	-2.69	.009	.04**

* $p < 0.05$. ** $p < 0.01$.

D.3.3 Primary ANOVA for factors of interest

```
/***      Quality (1-10)                                  ***/
/***      NumEvaluations (Quartile1-Quartile4)            ***/
/***      YearsExperience (Quartile1-Quartile4)           ***/
proc mixed;
        class Participant Quality NumEvaluations YearsExperience;
        model Rating =    Quality | NumEvaluations
                                  Quality | YearsExperience;
        lsmeans Quality             / adjust=tukey;
        lsmeans NumEvaluations      / adjust=tukey;
        lsmeans YearsExperience     / adjust=tukey;
        repeated                              / subject= Participant type=cs;
        by AdjectivePair;
run;
```

Effect	**Degrees of Freedom**		**Difficult**		**Helpful**		**Relevant**		**Required**	
	Num.	***Denom.***	***F***	***p***	***F***	***p***	***F***	***p***	***F***	***p***
Quality	9	603	18.52	<.0001	15.02	<.0001	11.36	<.0001	15.61	<.0001
NumEvaluations	3	67	0.16	.93	0.06	.98	0.66	.58	1.04	.38
Quality * NumEvaluations	27	603	1.20	.22	0.74	.82	0.76	.80	0.89	.62
YearsExperience	3	67	1.97	.13	0.21	.89	0.05	.99	0.50	.68
Quality * YearsExperience	27	603	0.82	.73	1.00	.46	0.92	.58	0.95	.54

D.3.4 Correlations with Experience

		Years of Experience		**Number of Evaluations**	
		***r*(72)**	*p*	***r*(72)**	*p*
Easy/Difficult	Provide Backing Data	**.10**	.49	**.07**	.55
	Clarity/Avoid Jargon	**.11**	.33	**.08**	.53
	Impact/Severity	**.06**	.61	**.10**	.40
	Describe Methodology	**.21**	.07	**.15**	.22
	Politics/Diplomacy	**.07**	.54	**.05**	.70
	Problem Cause	**.01**	.95	**.01**	.94
	Professional/Scientific	**.01**	.93	**.26**	.02*
	Provide a Solution	**.02**	.86	**.07**	.54
	Evoke Sympathy	**.05**	.64	**.06**	.60
	Describe User Actions	**.14**	.23	**.14**	.23
Helpful/harmful	Provide Backing Data	**.08**	.50	**.18**	.18
	Clarity/Avoid Jargon	**.19**	.18	**.04**	.72
	Impact/Severity	**.00**	.99	**.03**	.80
	Describe Methodology	**.17**	.51	**.09**	.46
	Politics/Diplomacy	**.21**	.07	**.14**	.23
	Problem Cause	**.13**	.26	**.03**	.82
	Professional/Scientific	**.03**	.78	**.13**	.27
	Provide a Solution	**.02**	.83	**.07**	.58
	Evoke Sympathy	**.28**	.08	**.08**	.55
	Describe User Actions	**.04**	.75	**.07**	.58
Relevant/irrelevant	Provide Backing Data	**.09**	.45	**.12**	.32
	Clarity/Avoid Jargon	**.07**	.56	**.06**	.61
	Impact/Severity	**.16**	.10	**.10**	.40
	Describe Methodology	**.05**	.70	**.03**	.82
	Politics/Diplomacy	**.08**	.48	**.03**	.84
	Problem Cause	**.04**	.73	**.01**	.95
	Professional/Scientific	**.07**	.56	**.22**	.06
	Provide a Solution	**.15**	.94	**.07**	.55
	Evoke Sympathy	**.24**	.04*	**.12**	.37
	Describe User Actions	**.04**	.77	**.11**	.35
Required/Optional	Provide Backing Data	**.08**	.49	**.10**	.40
	Clarity/Avoid Jargon	**.00**	.98	**.05**	.68
	Impact/Severity	**.18**	.34	**.03**	.80
	Describe Methodology	**.06**	.63	**.07**	.53
	Politics/Diplomacy	**.16**	.78	**.15**	.22
	Problem Cause	**.03**	.78	**.03**	.77
	Professional/Scientific	**.03**	.83	**.05**	.67
	Provide a Solution	**.06**	.62	**.02**	.86
	Evoke Sympathy	**.27**	.02*	**.07**	.58
	Describe User Actions	**.05**	.67	**.03**	.79

* $p < 0.05$.

APPENDIX E. Study 4 Participant Materials

E.1 IRB Approval for Study 4

Virginia Tech
VIRGINIA POLYTECHNIC INSTITUTE AND STATE UNIVERSITY

Institutional Review Board

Dr. David M. Moore
IRB (Human Subject) Chair
Assistant Vice President for Research Compliance
CVM Phase II - Duckpond Dr., Blacksburg VA 24061-0442
Office: 540/231-4991; FAX: 540/231-6033
e-mail: moored@vt.edu

30 June 2005

MEMORANDUM

TO: Tonya Smith-Jackson and Miranda Capra
ISE 0118

FROM: David M. Moore

SUBJECT: IRB *Exempt* Approval: # 05-415 "Novice and Expert Descriptions of Usability Problems"

I have reviewed your request to the IRB for exemption for the above referenced project. I concur that the research falls within the Exempt status as defined in 45CFR 46.101(b), based upon criteria 2. Approval is granted effective as of June 30, 2005.

As an Exempt study, and as allowed in federal regulations regarding Exempt studies, it is not necessary to obtained signed consent from participants.

Virginia Tech has an approved Federal Wide Assurance (FWA00000572, exp. 7/20/07) on file with OHRP, and its IRB Registration Number is IRB00000667.

cc: File

E.2 Recruiting Letter

Dear Usability Practitioner:

I am a Ph.D. student in Human Factors at Virginia Tech, studying under Dr. Tonya L. Smith-Jackson in the Department of Industrial and Systems Engineering. My dissertation research focuses on understanding differences among evaluators when conducting usability evaluations.

I am seeking usability practitioners to participate in a study. You will watch a recorded usability session, write up a report with comments about the interface you are evaluating, and fill out a survey. Your total time for this study should be about 2.5 hours. You can participate at home or in your office, and you can spread out your time over the course of a week. You will not be compensated for your time, but I would be happy to send you the results of the study when I complete my dissertation.

If you are interested in participating, would like more information, or would like me to send you the results, please contact me:

Miranda Capra <mcapra@vt.edu>

If you choose to participate, I will send you instructions for participating and a CD with a digital movie of the usability session via the US Postal Service.

Please feel free to share this letter with anyone that might be interested in this study (either in participating or in receiving a copy of the results).

Thank you for your consideration,
Miranda Capra

Miranda Capra <mcapra@vt.edu> http://www.thecapras.org/mcapra/
Ph.D. candidate, Human Factors and Ergonomics http://hfec.vt.edu/
Grado Department of ISE http://www.ise.vt.edu/
Assessment & Cognitive Ergonomics Lab http://ace.ise.vt.edu/

E.3 Study Packet: Cover Letter

Miranda Capra
ISE – Virginia Tech
250 Durham Hall
Blacksburg, VA 24061-0118
mcapra@vt.edu

April 5, 2006

Recipient's Name
Recipient's Address

Dear Recipient,

Thank you for agreeing to participate in this study for my dissertation research. You will find the instructions for the study attached to this letter, and the electronic files you will need are on the included CD. The study should take about 2.5 hours. You do not have to do the study in one sitting, but may spread it out over the course of a week.

Your 4-letter participant ID code is: xxxx

Please try to finish the study by **Monday, August XXXth.** If you need more time, please contact me to arrange for an extension.

Thank you,

Miranda Capra, Ph.D. Candidate, Human Factors
Grado Department of Industrial and Systems Engineering

E.4 Study Packet: Instructions and Usability Report Template

Overview

We are collecting descriptions of usability design issues from usability students and practitioners. We are asking you to conduct a usability evaluation, which involves the following:

- We have provided you with a movie of a usability session (UsabilityMovie.* on the CD)
- We have provided you a template for a usability report (UsabilityReport.* on the CD)
- You watch the movie and write a report with comments based on the movie (~ 2 hours)
    - Write this report as you would for your own usability assessments, including whatever details and images you would typically include in your own write-ups.
- You visit a website, upload your report and fill out a questionnaire (< 30 minutes)
    - http://hci.ise.vt.edu/~mcapra/DS/

You do not have to write the usability report in one sitting – you may spread out your work over the course of a week. Please keep track of the number of hours you spend watching the movie and writing the report. We will ask you for this information in the final questionnaire.

We estimate that the time spent on the evaluation and questionnaire will be about 2.5 hours.

Watching the Movie

The CD has a movie of a usability session, UsabilityMovie.[avi|mov]

PC Users: Please watch UsabilityMovie.avi using the **Morae Movie Player** (MoraePlay.exe), included on the CD. No installation is necessary – this file is the actual movie player

Mac Users: Please watch the QuickTime version (UsabilityMovie.mov) using the QuickTime player.

When you watch the movie, please make sure that you view it at 100% size. Shrinking or expanding the movie to fit it on your desktop will distort the text recorded from the web browser.

The movie was recorded from a computer desktop set to 1024x768. You can fit the entire movie on your screen if you set your desktop to a larger size, such as 1152x864 or 1280x1024.

Writing the report

The next page includes background and instructions for writing the report, and the final page is a printout of the report template.

The CD has an electronic copy of the report template, UsabilityReport.[doc|rtf]. Please make a copy of this document and edit it using your favorite document editor (Microsoft Word, Word Perfect, Open Office, etc.). You need to fill in all the areas highlighted in yellow.

Contact Information

If at any time you have any questions, please contact: Miranda Capra <mcapra@vt.edu>.

Background of the Usability Evaluation

You have been asked to do a small usability evaluation of the Internet Movie Database (IMDb; imdb.com). You have had four participants do the following task:

> **Task:** Name all the movies that both Owen Wilson and Luke Wilson (the actor from Old School) have appeared in together.
> **Answer:** The Wendell Baker Story, Rushmore, The Royal Tenenbaums, Bottle Rocket, and Around the World in 80 Days

User profile: The IMDB has a broad range of users. The site has both occasional visitors and two types of frequent visitors – those who do only basic tasks (such as looking up an actress or movie), and those who do more complex tasks. Visitors may be general movie watchers or movie enthusiasts, independent of their level of experience with the IMDb website.

The ultimate goal of your assessment is to develop a set of improvements to the interface.

Report Goals and Audience

Your current goal is to generate a list of usability problem descriptions that summarizes the data collected during the study both from user interactions and from expert observations during the test.

The people reading the descriptions (e.g. usability practitioners, developers, marketing, product development, management, clients) may or may not have usability training. They may use this list for many purposes, such as deciding which problems to fix in the next product release, designing interface changes, or adding features to the software.

Evaluation Instructions

Please watch the movie of the usability session and comment on the IMDb user interface.

- You may watch the movie as many times as you like.
- Use the provided report template for your comments (UsabilityReport.[doc|rtf])
- Your report should focus on the search feature tested in this task.
- Provide as many comments as you feel are appropriate.

For each comment that you write, please follow these guidelines.

- Please provide a category code for each comment, using the scheme in the table on the next page.
- In the description, include as much detail as you would typically include in your own reports. If you put images in your own reports you may include them in this report.
- Report one usability problem or one positive feature per comment.
  Split comments that are conglomerates of several problems or positive features.

Report Template: Usability Evaluation of www.imdb.com

Please provide a category code for each comment, using the scheme in the table below.

Code	Category	Description
PF	Positive finding	This approach is recommendable and should be preserved
MP	Minor problem	Caused test participants to hesitate for a few seconds
SP	Serious problem	Delayed test participants in their use of the website for 1 to 5 minutes, but eventually they were able to continue. Caused occasional "catastrophes"
CP	Critical problem	Caused frequent catastrophes. A catastrophe is a situation where the website "wins" over the test participant, i.e. a situation where the test participant cannot solve a reasonable task or where the website annoys the test participant considerably
GI	Good Idea	A suggestion from a test participant that could lead to a significant improvement of the user experience.
B	Bug	The website works in a way that's clearly not in accordance with the design specification. This includes spelling errors, dead links, scripting errors, etc

Comments on the Website

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.

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.

E.5 Questionnaire

Usability Study [redacted]

Welcome

Thank you for participating in this study. To complete the study, you will need to upload your usability report and complete a 2-page questionnaire. This should take under 30 minutes.

Please help us identify you by entering the 4-letter identification code from your instruction packet.

If you do not have a code, please contact Miranda Capra <mcapra@vt.edu>

[ ] Submit this 4-letter code (codes are case-sensitive)

If you have any questions or would like to be informed of the results of this study please contact:
Miranda Capra <mcapra@vt.edu>

Faculty Advisor: Dr. Tonya L. Smith-Jackson

Usability Study [redacted]

Participant identification code: AAAA

Step 1:	Upload usability report	incomplete
Step 2:	Problem Description Questionnaire	incomplete
Step 3:	Demographic Questionnaire	incomplete
Step 4:	Optional Questionnaire	incomplete

Step 1: Upload Usability Report

Please select the file that contains your usability report to upload
[ ] Browse… then upload this file

The maximum filesize is 5MB - if your file is larger than this, please contact Miranda Capra <mcapra@vt.edu>

If you have any questions or would like to be informed of the results of this study please contact:
Miranda Capra <mcapra@vt.edu>

Faculty Advisor: Dr. Tonya L. Smith-Jackson

Usability Study [redacted]

Participant identification code: AAAA

Step 1:	Upload usability report	*complete*	*filename: UsabilityReport.doc*
Step 2:	Problem Description Questionnaire	incomplete	
Step 3:	Demographic Questionnaire	incomplete	
Step 4:	Optional Questionnaire	incomplete	

Step 1: Upload Usability Report

You have sucessfully uploaded the following usability report: **UsabilityReport.doc**

If that is your final report, please continue to Step 2
Once you continue to Step 2 you will not be able to return to Step 1 to make changes to your uploaded report.

If you need to upload a new version, please use the form below.

Please select the file that contains your usability report to upload
Browse… then upload this file

The maximum filesize is 5MB - if your file is larger than this, please contact Miranda Capra <mcapra@vt.edu>

If you have any questions or would like to be informed of the results of this study please contact:
Miranda Capra <mcapra@vt.edu>

Faculty Advisor: Dr. Tonya L. Smith-Jackson

Usability Study [redacted]

Participant identification code: AAAA

Step 1:	Upload usability report	*complete filename: UsabilityReport.doc (editing is no longer available)*
Step 2:	Problem Description Questionnaire	incomplete
Step 3:	Demographic Questionnaire	incomplete
Step 4:	Optional Questionnaire	incomplete

Step 2: Problem Description Questionnaire

Imagine the following scenario (similar to the evaluation you just completed):

> You, a usability practitioner, are part of a team assessing the usability of a user interface (website, software, consumer product, etc.) that you may or may not have designed. The ultimate goal of your assessment is to develop a set of improvements to the interface.
>
> You have just completed a formative usability evaluation using usability testing with end-users. Your current goal is to generate a list of usability problem descriptions that summarizes the data collected during the study both from user interactions and from expert observations during the test.
>
> The people reading the descriptions (e.g. usability practitioners, developers, marketing, product development, management, clients) may or may not have usability training. They may use this list for many purposes, such as deciding which problems to fix in the next product release, designing interface changes, or adding features to the software.

Given this scenario, consider the individual usability problem descriptions that you will write. What should you include in these descriptions? What should you avoid? What makes one usability problem description better or worse than another problem description?

Through previous research I have collected the following 10 qualities of a description of a usability problem (in random order):

1. Describe the cause of the problem
2. Describe a solution to the problem
3. Justify the problem with data from the study
4. Be clear and precise while avoiding wordiness and jargon
5. Consider politics and diplomacy when writing your description
6. Describe observed user actions
7. Describe the impact and severity of the problem
8. Be professional and scientific in your description
9. Help the reader sympathize with the user
10. Describe your methodology and background

Instructions

We would like you to help me determine which of these 10 qualities are the most important for a good usability problem description.

- Read each quality and its description
- Rate your impression of the quality for each of the adjectives provided
  - Don't agonize over this - I'm most interested in your first impression
  - The center position indicates neither adjective, or both adjectives equally
- **NOTE**
  - These are qualities of a description of an **individual** usability problem, not a complete usability report
  - Assume the problems were found using usability testing **with end-users**
  - Assume these descriptions are your **only means of communication**, and you cannot supplement with personal communications

---

1. 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.

When I write a description of a single usability problem,
I consider addressing this quality of a problem description ...

required	O	O	O	O	O	O	O	optional
helpful	O	O	O	O	O	O	O	harmful
easy	O	O	O	O	O	O	O	difficult
irrelevant	O	O	O	O	O	O	O	relevant

2. 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.

When I write a description of a single usability problem,
I consider addressing this quality of a problem description ...

required	O	O	O	O	O	O	O	optional
helpful	O	O	O	O	O	O	O	harmful
easy	O	O	O	O	O	O	O	difficult
irrelevant	O	O	O	O	O	O	O	relevant

3. Justify the problem with data from the study, both quantitative and qualitative. Include how many users experienced the problem and how often; task attempts, time and success/failure; and critical incident descriptions. Provide traceability of the problem to observed data.

When I write a description of a single usability problem,
I consider addressing this quality of a problem description ...

required	O	O	O	O	O	O	O	optional
helpful	O	O	O	O	O	O	O	harmful
easy	O	O	O	O	O	O	O	difficult
irrelevant	O	O	O	O	O	O	O	relevant

4. 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 people will appreciate. Avoid so much detail that no one will want to read the description.

When I write a description of a single usability problem,
I consider addressing this quality of a problem description ...

required	O	O	O	O	O	O	O	optional
helpful	O	O	O	O	O	O	O	harmful
easy	O	O	O	O	O	O	O	difficult
irrelevant	O	O	O	O	O	O	O	relevant

5. Consider politics and diplomacy when writing your description. Avoid judging the system, criticizing decisions made by other team members, pointing fingers or assigning blame. Point out good design elements and succesful user interactions. Be practical, avoiding theory and jargon.

When I write a description of a single usability problem,
I consider addressing this quality of a problem description ...

required	O	O	O	O	O	O	O	optional
helpful	O	O	O	O	O	O	O	harmful
easy	O	O	O	O	O	O	O	difficult
irrelevant	O	O	O	O	O	O	O	relevant

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.

When I write a description of a single usability problem,
I consider addressing this quality of a problem description ...

required	O	O	O	O	O	O	O	optional
helpful	O	O	O	O	O	O	O	harmful
easy	O	O	O	O	O	O	O	difficult
irrelevant	O	O	O	O	O	O	O	relevant

7. 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.

When I write a description of a single usability problem,
I consider addressing this quality of a problem description ...

required	O	O	O	O	O	O	O	optional
helpful	O	O	O	O	O	O	O	harmful
easy	O	O	O	O	O	O	O	difficult
irrelevant	O	O	O	O	O	O	O	relevant

8. Be professional and scientific in your description. Use only facts from the study, rather than opinions or guesses. Back your findings with sources beyond the current study (such as an external classification scheme), proven usability design principles, and previous research.

When I write a description of a single usability problem,
I consider addressing this quality of a problem description ...

required	O	O	O	O	O	O	O	optional
helpful	O	O	O	O	O	O	O	harmful
easy	O	O	O	O	O	O	O	difficult
irrelevant	O	O	O	O	O	O	O	relevant

9. Help the reader sympathize with the user by using a problem description that is evocative and story-like. Make sure the description is readable and understandable. Use user-centric language rather than system-centric. Be complete while avoiding excessive detail.

When I write a description of a single usability problem,
I consider addressing this quality of a problem description ...

required	O	O	O	O	O	O	O	optional
helpful	O	O	O	O	O	O	O	harmful
easy	O	O	O	O	O	O	O	difficult
irrelevant	O	O	O	O	O	O	O	relevant

10. Describe your methodology and background. Describe how you found this problem (field study, lab study, expert evaluation, etc.). Describe the limitations of your domain knowledge. Describe the user groups that were affected and the breadth of system components involved.

When I write a description of a single usability problem,
I consider addressing this quality of a problem description ...

required	O	O	O	O	O	O	O	optional
helpful	O	O	O	O	O	O	O	harmful
easy	O	O	O	O	O	O	O	difficult
irrelevant	O	O	O	O	O	O	O	relevant

Submit Problem Description Questionnaire

Usability Study [redacted]

Participant identification code: AAAA

Step 1:	Upload usability report	*complete filename: UsabilityReport.doc (editing is no longer available)*
Step 2:	Problem Description Questionnaire	complete edit
Step 3:	Demographic Questionnaire	incomplete
Step 4:	Optional Questionnaire	incomplete

Step 3: Demographic Questionnaire

1. Approximate number of hours spent watching the movie (decimals are fine) and writing your report:

2. Approximate years of usability experience. Please only count years that you have spent conducting usability evaluations of user interfaces in any context (industry, university, etc.).

3. Approximate number of usability evaluations you have conducted:

4. What usability technique(s) do you commonly use when conducting usability evaluations?
   - Cognitive Walkthrough
   - Expert Review
   - Focus Groups
   - Heuristic Evaluation
   - Heuristic Inspection
   - Laboratory Testing With Users
   - Naturalistic Observation
   - Remote Usability Testing
   - Surveys
   - User Interviews

   Other:

5. Where do you currently work?
   - Industry
   - Government
   - Military
   - University

   Other:

6. What is your job position/title?

7. Please briefly describe any usability-related degrees, certifications, training or experience that you have.

8. When I perform a usbility evaluation, I usually report the results of the evaluation in the following way:

formally	O	O	O	O	O	O	O	informally
verbal communication	O	O	O	O	O	O	O	written report
no solutions/fixes/suggestions	O	O	O	O	O	O	O	detailed solutions/fixes/suggestions
as a specialist brought in for the evaluation	O	O	O	O	O	O	O	as an ongoing member of the product team

9. Approximately how many times a month do you use IMDb.com?
   [ ]

10. Approximately how many times have you used the IMDb.com search feature tested in this usability session?
    [ ]

11. What country are you from?
    [ ]

12. Is English your first language?
    O yes O no

13. Please indicate your agreement with the following statement:
    I speak English as well as someone that only speaks English.

Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
O	O	O	O	O	O

Submit Demographic Questionnaire

If you have any questions or would like to be informed of the results of this study please contact:
Miranda Capra <mcapra@vt.edu>

Faculty Advisor: Dr. Tonya L. Smith-Jackson

Usability Study [redacted]

Participant identification code: AAAA

Step 1:	Upload usability report	*complete filename: UsabilityReport.doc (editing is no longer available)*
Step 2:	Problem Description Questionnaire	complete edit
Step 3:	Demographic Questionnaire	complete edit
Step 4:	Optional Questionnaire	incomplete

Optional Questions

Would you like to be informed of the results of this study?

Please send me a brief summary of your results
Please send me a link to an electronic copy of your dissertation
Please let me know when these results are published

If you are a student or recent graduate, what usability experience do you have, and what classes have you taken in Usability and Human-Computer interaction?

Please tell us a little bit about how you work. How do you fit into your company and the product design pocess? Are you a consultant or in-house or something else? What kind of studies do you run?

Do you work alone or are you part of a group?

How do you communicate the results of your studies? Do you write a formal report, and do you write it alone or with a group?

When you describe usability problems you find in a study, do you use any special reporting template or bug tracking system? What do you use, and what standard information do you include?

We have found that usability practitioners vary greatly in whether or not they think that a usability report like this one should recommend solutions to the problems found. Do you think that the results of a usability study should suggest solutions to the problems described? Why or why not? If so, how detailed a solution do you recommend?

Do you have any other comments for us about the usability study you watched, the report you wrote, this questionnaire, or anything else about this study?

Submit These Comments

If you have any questions or would like to be informed of the results of this study please contact:
Miranda Capra <mcapra@vt.edu>

Faculty Advisor: Dr. Tonya L. Smith-Jackson

Usability Study [redacted]

Participant identification code: AAAA

Step 1:	Upload usability report	*complete*	*filename: UsabilityReport.doc (editing is no longer available)*
Step 2:	Problem Description Questionnaire	complete	edit
Step 3:	Demographic Questionnaire	complete	edit
Step 4:	Optional Questionnaire	complete	edit

Questionnaire Complete

You have now completed the questionnaire. You may change your response to Parts 2-4 using the "edit" buttons above.

If you are finished, please click the "Done - submit my questionnaire" button below. Once you submit your questionnaire you will no longer be able to edit your responses to the questionnaire.

Done – submit my questionnaire

If you have any questions or would like to be informed of the results of this study please contact:
Miranda Capra <mcapra@vt.edu>

Faculty Advisor: Dr. Tonya L. Smith-Jackson

Usability Study [redacted]

Participant identification code: AAAA

Thank You

Thank you for participating in this study.

You have uploaded your usability report and completed the online questionnaires. You can no longer edit any of your responses.

If you have any questions or would like to be informed of the results of this study please contact:
Miranda Capra <mcapra@vt.edu>

This page intentionally left blank.

APPENDIX F. Study 4 Judging Materials

F.1 Judging Instructions – Matching Problems

Judging Goals

There are four main goals of this process.

1. Create a comprehensive list of every problem mentioned by any evaluator
2. Identify every practitioner that identified/discovered/discussed each problem
3. Identify problems that contain errors or misunderstandings, or are very vague
4. Decide which problems are "real" problems (serious or critical) and which ones are not (minor/superficial)

Judging Instructions

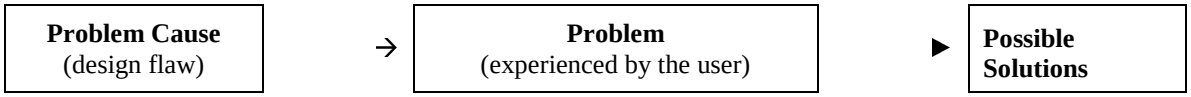
Your login is your first name (Jon, Laurian, Pardha). Yes, you can see each other's work if you want to, but please don't. Also, please minimize talking to each other about specific evaluators or comments, problems you added to the master list, or anything else that might bias each other's judgements.

For each of the 507 usability problem descriptions, decide if the problems discussed are the same as any problems in the master problem list

- Man y problem descriptions will describe multiple problems in the master list.
- Please glance over the printed report when you open a new evaluator. Sometimes I had to approximate the formatting of their report in the website version. Sometimes context can resolve vagueness.
- You can ignore positive comments – I am not cataloging those
- When you find problems not listed in the master list of problems, add them to your copy of the master list
- Keep an eye open for additional comments about the experimental protocol (group j) and common misunderstandings by the evaluators (group k)
- Feel free to take notes about problems in the master list, such as additional causes/solutions or your questions.
- If you think a problem in the master list is really two problems, create two problems and use them instead

What is a problem?

A problem is something experienced by the user, as shown in this diagram. One problem can have multple causes, and frequently it's hard to tell which one it really was. For example, did the user miss the instructions entirely, or did the user see the intructions but not understand them? Different causes can imply slightly different solutions.



Sometimes an evaluator describes a solution, but no problem. This is particularly true in the case of predicted problems, as opposed to observed. In this case you'll need to determine what problem is inferred by the solution.

When are two problems the same?

Let's say that you have solution A and solution B, and you're trying to decide if they solve the same problem. First, assume that the project team will make only the change discussed in the problem description and no more. Second, solution A and solution B probably solve the same problem(s) if

- making fix A solves all the problem(s) for B **AND** making fix B solves all the problem(s) for A

Here's an example:

- **Problem A:** users don't understand the difference between Luke Wilson I and II.
  Cause/Solution: I and II have little meaning. Add a photo or most popular movie title
- **Problem B:** users think Luke Wilson I and II are Luke Wilson Jr and Sr.
  Cause/Solution: I and II have other meanings. Use arabic numerals (1 and 2) instead of roman (I and II)

A and B really aren't the same, because making change B will not fix A, even though making change A will fix B.

- **Problem C:** User could not find the "credited alongside" search, even when user was on the right page.
- *Cause/Solution:* Add a clear shortcut above the fold.
- *Cause/Solution:* Move the "credited alongside" search above the fold.

This is one problem, because either fix makes the other fix unecessary.

Feel free to ask me <mcapra@vt.edu> questions about the process or issues that come up – if I think my answer might bias you too much, I can always decline to anwer.

F.2 Judging Form – Matching Problems

Usability Problem Description Matching and Rating System

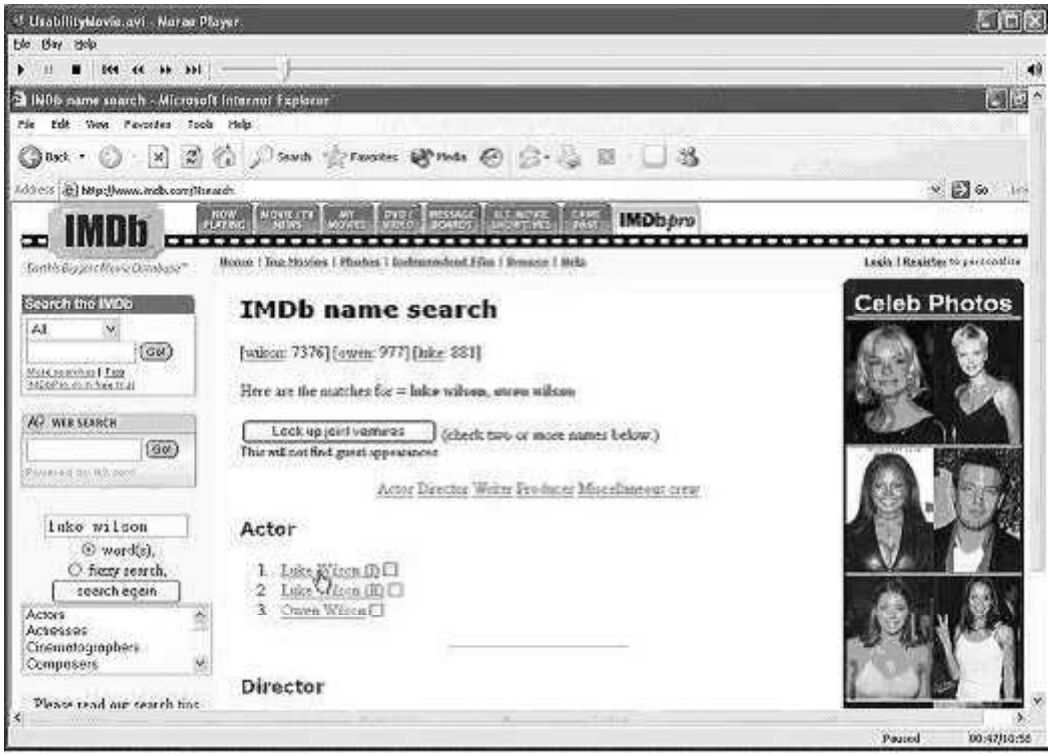
Legend: done, started, on hold, not started
System will start at first one that is not done, skipping stuff that is on hold. Feel free to edit stuff that is done .
Master Problems: ALL aa ab ac ad ae af ag ah ba bb bc bd be ca da db ea eb ec ed ee fa fb fc fd fe ff ga gb gc ha hb hc ia ja jb jc ka kb

Add a new problem to the Master List

Evaluators:
eZYt Qthm czLR XHXd YYYL XKsh wrdn AySZ KfpB BfBa WHFy AZXR YBYn qpyq Qjdr
iTgJ Jifa eKpB ZfrN KycZ JMbR cwnT MGSE NZdb Bwje KqTn YWmq yXht tHic mqfm
LBjg ewjn deZL AwEr emsR tZSp nSeM TTGG dthe sAcp xXYm LZbQ fxby ThNE hxXn
XHXd comments: ALL 1 2 3 4 5 6 7 8 9

---

eZYt Comment #1:
Open this comment in a separate window.



Links to Luke Wilson's bibliography on the above page are more salient than the check boxes. This leads users to believe clicking is the way to select rather than checking the box.

This a purely positive or neutral comment, without any mention of either problems or solutions.

How vague is this problem?
O I can't tell what problem is described
O Most people that have not seen the movie will not understand what problem is described
O Some people the have not seen the movie will not understand what problem is described
O The description is not vague
Please describe anything that is vague:

If the comment is positive, neutral, or so vague you can't tell what is described, of you already filled out the rest of the form, you may submit the form and skip the rest of the page

This comment contains factual errors.
Please describe anything the evaluator misunderstood or did not know:

Please list all master problems that match this comment (aa ab ac):

You can write comments in the comment box below, such as any doubts, two that seem equally likely, or a runner-up you almost chose.
If you find a new problem not already in the list, you can add a new problem to the Master List
Clicking this button will save everything you have entered into this form so far. You can come back to this form later.

Do you have any comments about this comment? (optional)

Do you have any comments about this evaluator? (optional)

Do you have any comments about this study? (optional)

submit the form

If you have any questions or discover a bug, or if you have a feature that would really help you get through this faster, drop me an email:

Miranda Capra <mcapra@vt.edu>

F.3 Judging Form – Which Master Problems Are Real?

Judging which master problems are "real" and "not real"

For every problem in the list, I need you to make two decisions:

1. What is the severity of this problem (according to the written problem description)? Minor, serious or critical?

   - MP Minor Problems: Caused test participants to hesitate for a few seconds
   - SP Serious Problems: Delayed test participants in their use of the website for 1 to 5 minutes, but eventually they were able to continue. Caused occasional "catastrophes"
   - CP Critical Problems: Caused frequent catastrophes. A catastrophe is a situation where the website "wins" over the test participant, i.e. a situation where the test participant cannot solve a reasonable task or where the website annoys the test participant considerably

2. Is this problem "correct" or is it "wrong"?

   - "Wrong" problems are where the evaluator misunderstood what was going on, or gave a bad suggestion

Based on these judgments, "real" problems are serious/critical AND correct, "not real" problems are minor OR wrong.

Remember that a problem is something experienced **by the user**, so "bugs" (system glitches) are problems!

	Severity [MP SP CP] (pick one)	**Correct or wrong?** [C W] (pick one). If wrong, briefly note why (e.g. evaluator mistake, bad idea)		**Severity** [MP SP CP] (pick one)	**Correct or wrong?** [C W] (pick one). If wrong, briefly note why (e.g. evaluator mistake, bad idea)
A aa				**H** ha	
ab				hb	
ac				hc	
ad				**I** ia	
ae				**J** ja	
af				jb	
ag				jc	
ah				**K** ka	
B ba				kb	
bb				**L** la	
bc				lb	
bd				lc	
be				ld	
C ca				le	
D da				lf	
db				lg	
E ea				lh	
eb				li	
ec				lj	
ed				lk	
F fa				ll	
fb				lm	
fc				ln	
fd				lo	
fe				lp	
ff					
G ga					
gb					
gc					

F.4 Judging Form – Rating Reports

Report Rating Form for Evaluator: _____

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 people will appreciate. Avoid so much detail that no one will want to read the description.	According to the guideline at the left, **this report is clear and precise**					
	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
	○	○	○	○	○	○
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.	According to the guideline at the left, **this report describes the impact and severity of problems**					
	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
	○	○	○	○	○	○
Justify the problem with data from the study, both quantitative and qualitative. Include how many users experienced the problem and how often; task attempts, time and success/failure; and critical incident descriptions. Provide traceability of the problem to observed data.	According to the guideline at the left, **this report justifies problems with data**					
	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
	○	○	○	○	○	○
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.	According to the guideline at the left, **this report describes the cause of the problems**					
	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
	○	○	○	○	○	○
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.	According to the guideline at the left, **this report describes observed user actions**					
	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
	○	○	○	○	○	○
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.	According to the guideline at the left, **this report describes solutions to problems**					
	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
	○	○	○	○	○	○

This report describes solutions or changes, but does not describe the problems that are being fixed

Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
○	○	○	○	○	○

Any other notes about this report:

APPENDIX G. Study 4 Final Master Problem List (MPL)

G.1 Group a: Top-left search box and results (IMDb Name Search #1)

Problem aa (minor):

The user hesitated between choosing the top-left search box and the A9 search box immediately underneath.

Possible Causes/Solutions:

- There's not enough information about what each box does - add more info to help distinguish the two.

Problem ab (severe):

The user did not know that the search box would not accept exact string (""), AND or + searches. (Similar to next problem.)

Possible Causes/Solutions:

- The top-left search box should give some examples of how to use it.
- Provide a link to an advanced search ("more searches" does not lead to an advanced search, just different simple searches)

Problem ac (severe):

Even after doing a complex search, the user did not know that the search box would not accept exact string (""), AND or + searches. (Similar to previous problem.)

Possible Causes/Solutions:

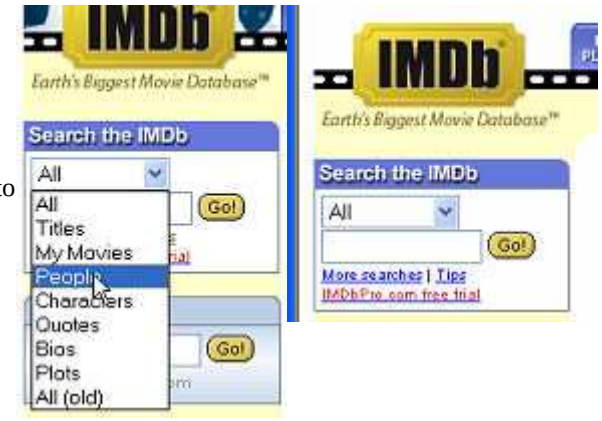
- Recognize that such a search has been done and provide specific search tips
- Recognize that such a search has been done and provide a link to the proper search page
- Users expect search engines to do this type of search, so this search engine should support this functionality
- "Using whole word searching" is probably meant to indicate this, but probably no one reads, or understands that this is what it means. Reword and place in a better location.

Problem ad (minor):

The user did not notice the pull-down menu in the top-left search box.

Possible Causes/Solutions:

- Perhaps the user missed it - make it easier to notice



Problem ae (minor):

The user was confused about the options in the pulldown box (All, People, Characters) and kept trying different ones, hoping one would do what he wanted. This is just about design of this pull-down box.

Possible Causes/Solutions:

- The terms are confusing - provide more information or pick better names

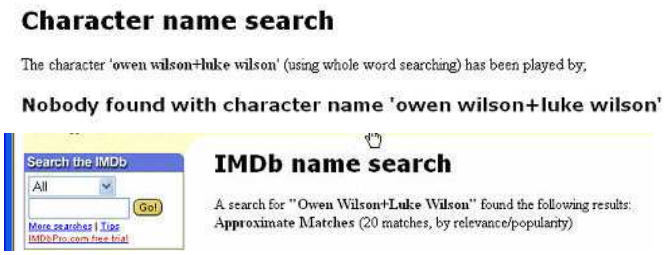
Problem af (minor):

The user could not see his entire search string before submitting.

Possible Causes/Solutions:

- The text entry box is too small - make it wider.

Problem ag (severe):



After submitting a search in the top-left box, the user was confused about the results. The user could not tell exactly what search had been done to compare the results to his task goal. This is just about the display of the search results (feedback and information display issues). (Similar to next problem.)

Possible Causes/Solutions:

- System should display exactly what search was done, such as the exact option selected in the pull-down box.

Problem ah (minor):

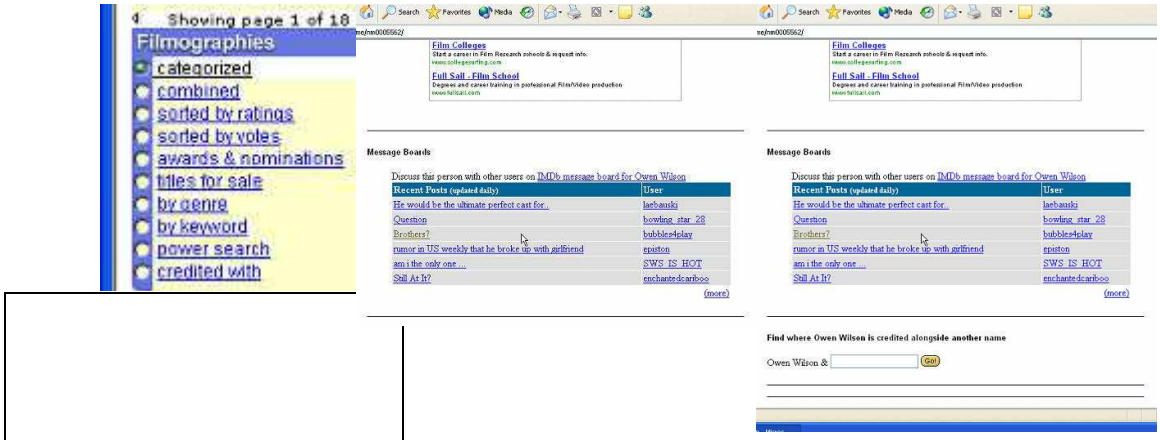
The user had to re-type the search query into the search box to try the same search with a different pull-down option. (Similar to previous problem.)

Possible Causes/Solutions:

- Redisplay the search form with the previous values and allow the user to edit and resubmit. That's the obvious next step - make it easy for the user.

G.2 Group b: Actor Pages (Owen Wilson, Luke Wilson) including "Filmography" box and "credited alongside" search

Problem ba (severe):



User could not find the "credited alongside" search box, even though user was already on the correct page.

Possible Causes/Solutions:

- The page has an overwhelming amount of information that is poorly organized. Restructure the page to have more clearly defined sections or fewer sections
- There were several false cognitive affordances that looked like the bottom of the page and kept the user from scrolling all the way down (horizontal lines, advertising box, message boards). Redesign the page to have better organized and designed sections.
- Add a shortcut to the "credited alongside" search somewhere very prominent and above the fold, or make the "credited with" link on the left side under "Filmography" more prominent"
- Move the "credited alongside" search box above the fold
- Move the "credited alongside" search box before the message boards

Problem bb (deleted in favor of problems la and lp):

The user got lost in the website and had a hard time understanding exactly what he was looking at or where he was. This is about the user getting lost in the website in general – if the user is on the Owen Wilson page but is having trouble finding the "credited alongside" earch box, use the previous problem in the list.

Possible Causes/Solutions:

- The "Filmography" section down the left side of the page describes the current location, but it is too far away from the main section of the page and users might not notice it. Increase its connection to the rest of the page through proximity or visual grouping (lines/colors)
- The little dots next to each item in the "Filmography" section that turn green to indicate your current location are too small to be noticed. Make a more prominent indicator of your current location.
- The "Owen Wilson" and "Luke Wilson" pages have no indication in the main area that you are on the "categorized filmography" page.

Problem bc (minor):

The user looking for the "credited alongside" box did not click on the "credited with" link, even though the user clicked on other links in the "Filmography" section.

Possible Causes/Solutions:

- Use words for "credited with" that are more meaningful to the user.

Problem bd (severe):

User thought the "combined" link under Filmographies would be useful for a joint search.

Possible Causes/Solutions:

- Rename the "combined" link to be more meaningful to the user and more descriptive of what it does.

Problem be (minor):

User thought the "keyword" link under Filmographies would be useful for a joint search.

Possible Causes/Solutions:

- The user thinks a "keyword" search might be a special of kind of "and" search, and that Luke Wilson might be a keyword for Owen Wilson. However, IMDb probably means "keyword" to be special words tracked in their database, but not people. Use user-centric terms and hide system internals from the user.
- Rename the "keyword" link to be more meaningful to the user and more descriptive of what it does.

G.3 Group c: Keywords page

Problem ca (severe):

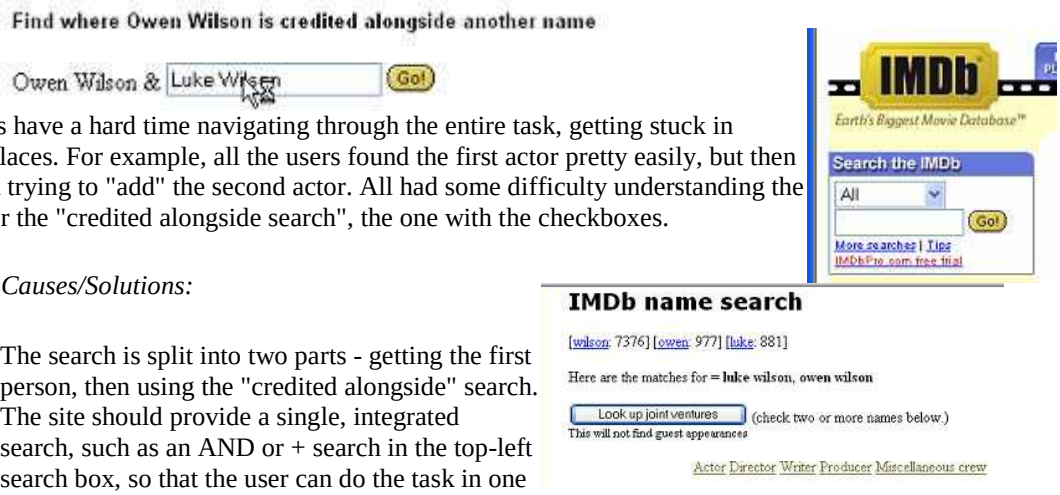
User was really confused by the "Keywords" page.

Possible Causes/Solutions:

- It has lots of information - simplify or organize better
- It has no description of what it is or how it is organized - add a description

G.4 Group d: Issues Involving Multiple Pages

Problem da (severe):

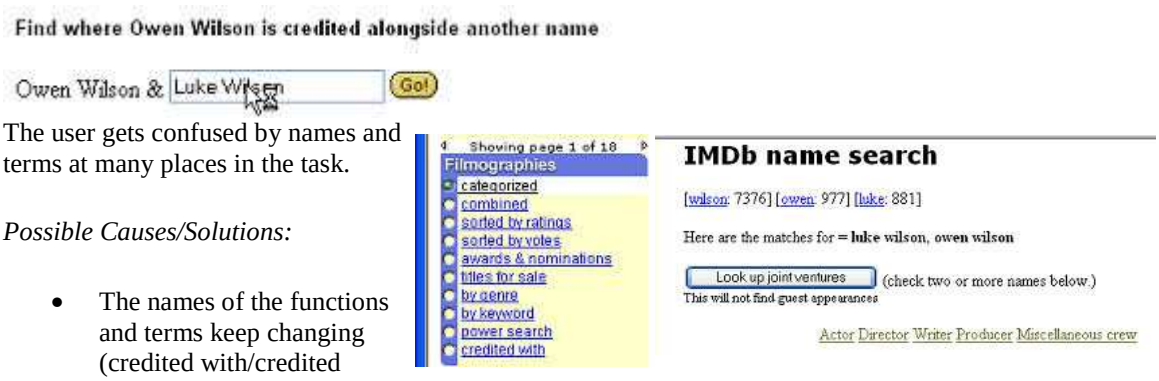


The users have a hard time navigating through the entire task, getting stuck in several places. For example, all the users found the first actor pretty easily, but then got stuck trying to "add" the second actor. All had some difficulty understanding the page after the "credited alongside search", the one with the checkboxes.

Possible Causes/Solutions:

- The search is split into two parts - getting the first person, then using the "credited alongside" search. The site should provide a single, integrated search, such as an AND or + search in the top-left search box, so that the user can do the task in one step

Problem db (severe):



The user gets confused by names and terms at many places in the task.

Possible Causes/Solutions:

- The names of the functions and terms keep changing (credited with/credited alongside -> IMDb name search -> joint ventures). Use consistent terminology.

G.5 Group e: First results page for "credited alongside" search (IMDb Name Search #2)

Problem ea (severe):

When users saw the results of the "credited alongside" search, they had no idea what to do with this page. They hesitated for a while, scrolling up and down the page. Note: for issues about accidentally checking the writer boxes, see the checkboxes section.

Possible Causes/Solutions:

- The search results don't match the user's expection of a list of movies. Make some assumptions about the best initial search. Present a list of movies and then let the user refine their query. This is how search engines work, and users expect it now.
- The page is complicated. Provide instructions about what this page is and what to do, and use a more logical layout, with the page items presented in the order they are used.
- The page says it lists "matches", but it means matches to the people, whereas the user is expecting movies. Describe what the page actually returns - occupations/roles.
- The page is spread out vertically, and it's hard to see the whole page at once. Use a more compact layout, like a table, so that the entire form is above the fold.

Problem eb (minor):

User almost clicked on the "[wilson: 7376]" link at the top of the page.

Possible Causes/Solutions:

- This information is irrelevant to the user's current task - consider removing this information.
- Provide more information about what this link means

Problem ec (minor):

User did not know what the "look up joint ventures" button would do.

Possible Causes/Solutions:

- The wording does not match the user's view of the task. User more descriptive and user-centric wording.
- The wording is inconsistent with previous portions of the task. Use consistent terminology.

Problem ed (minor):

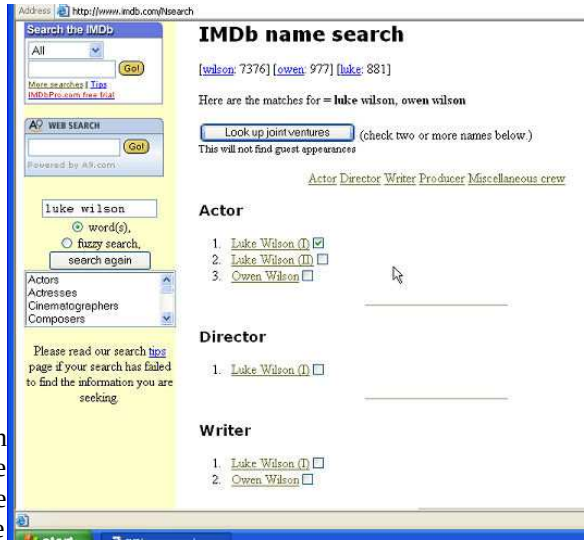
User was unsure about what “ventures” are. User spoke about "venues" instead of "ventures."

Possible Causes/Solutions:

- What exactly are "ventures"? Pick a word that is more descriptive.
- "Venue" is a common movie-related term. Pick a word that is less similar to venue.

G.6 Group f: First results page for "credited alongside" search (Issues related to checkboxes)

Problem fa (severe):



The user did not click on any checkboxes before clicking on the "look up joint ventures" button.

Possible Causes/Solutions:

- The button is at the top of the page. Stuff should be in the order that you do it. If you have to check boxes first, don't put the button at the top of the page.
- The button is before the instructions and the instructions are in parentheses, which implies that they are optional. Make the instructions more prominent.
- The instructions say to "check" boxes, which may be unclear when none of the boxes have checks yet, and which is an American phrase (others use "tick" or "mark"). Rephrase to be more clear.
- In cultures that read left-right, the checkboxes should be on the left side of their labels.

Problem fb (severe):

It took a while for the user to understand what the checkboxes were for.

Possible Causes/Solutions:

- The purpose of the checkboxes is not indicated. Make their purpose clear. For example, put them all in a column and label the column, or add instructions at the top of the page.

Problem fc (severe):

The user only clicked one checkbox before clicking the "look up joint ventures" button.

Possible Causes/Solutions:

- The users had already selected one user they wanted. The system should remember this and not make the user re-do work.

Problem fd (severe):

The users did not understand the occupations/roles. The user did not realize that "actor director writer" were different occupations/roles that applied to each person. One user checked the "writer" boxes instead of "actor" boxes.

Possible Causes/Solutions:

- There is no explanation of what the occupations/roles are or why they are there. Add instructions.
- Most moviegoers don't think about actors that also direct and write. The explanation that these are occupations/roles is all the way at the bottom of the page and most people probably missed them. Explain what these occupations/roles are and why you have to pick them near the top of the page.

Problem fe (minor):

One user fogot to uncheck the "writer" boxes before checking the director boxes.

Possible Causes/Solutions:

- It's hard to tell which boxes are and are not checked. Redesign the page so that all the boxes are lined up in columns.
- The user had to hit the "back" button to revise the search after seeing the results, and may not have realized that the boxes were still checked. Always re-display the search form with the previous search parameters so that the user doesn't have to use the "back" button, and the results are on the same page as the search form to revise.
- User thought the site would do an OR search, but it did an AND search. Move the selection box for this above the fold and rephrase to be more clear.

Problem ff (minor):

The user clicked on a name instead of checking a box.

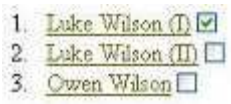
Possible Causes/Solutions:

- In cultures that read left-right, the checkboxes should be on the left side of their labels.
- There's no indication of the different results that happen after clicking on the name versus the checkboxes. Provide instructions about what the checkboxes mean.
- IMDb.com has a site-wide standard that names are linked to the person's homepage. However, new users may not be aware of this, especially in this tasks's context. Redesign page to make the meaning clear.

G.7 Group g: First results page for "credited alongside" search (Issues related to I, II)

Problem ga (minor):

The user found the correct Luke Wilson and used the "credited alongside" box on that Luke Wilson's page. However, the search results then asked the user to choose between Luke Wilson I and II.

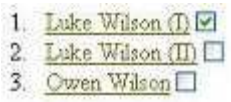


Possible Causes/Solutions:

- The user had already selected the correct Luke Wilson. The system should remember this and not make the user re-do work.

Problem gb (minor):

User thought I/II meant jr/sr. (Similar to next problem.)

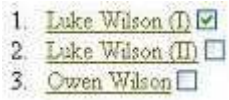


Possible Causes/Solutions:

- Currently they indicate the order the people were added to the database, which has no meaning to users. Hide system internals from the user.
- Use arabic numerals (1/2) instead of roman (I/II).

Problem gc (minor):

User was unsure what the difference was between Luke Wilson I and II. (Similar to previous problem.)



Possible Causes/Solutions:

- Provide additional information, such as most popular movie, photo or balloon/rollover informaiton.

G.8 Group h: Joint Ventures Search Results

Problem ha (minor):

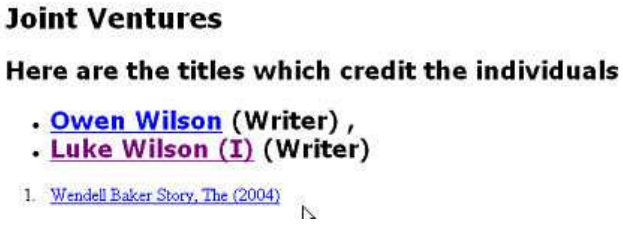
The results were not what the user wanted, but the user had to click the "back" button to revise the search.

Possible Causes/Solutions:

- Redisplay the search form with the previous values and allow the user to edit and resubmit. That's the obvious next step - make it easy for the user.

Problem hb (severe):

User had a hard time understanding why the search with the two writer boxes checked didn't produce the desired result. This is just about understanding the search results - look in previous sections for issues about conducting the correct search.



Possible Causes/Solutions:

- User did not understand the difference between the writer and actor occupation/roles. Both this page and the previous page should have explanations of these occupations/roles.
- User did not realize that the user had done a search for writer. Put up an easier-to-understand display of exactly what search was done, such a table of people and roles.

Problem hc (minor):

User was unsure if the search results were the list of movies that he was looking for.

Possible Causes/Solutions:

- The final results page emphasizes the search parameters too much, and the search results not enough. The page should be redesigned to emphasize the results more (i.e. the list of matching movies).
- The movies are not actually labeled as “movies” – provide a label for any information displayed.
- The results include information that is irrelevant to the user's task. Consider removing the ratings and votes (e.g. 7.3/10 (401 votes)).

G.9 Group l: Problems Added By Judges

Master problem #la description: (severe)

Problem:
Users have a hard time finding features they want to use in the site, and feel overloaded by the amount of information/features available. This is an overall problem with the entire site.

Possible Causes/Solutions:

- The site is very busy and cluttered, with many features and lots of text.
- The page layout is poor, with few aids to help users sort through all the links.
    - A more focused layout would eliminate some of the surrounding text and links.
- There are 3 types of search on the left including:
    - a fuzzy search (many won't know what that means)
    - a Web search (which probably isn't really needed here or could be put elsewhere)
    - search with the dropdown menu
- Reorganize/consolidate the search options to Simple Search and Advanced Search. Move web search away to a different part of the page.
- The actual solution will vary by page

Master problem #lc description: (minor)

Problem:
The error message "need 2 or more names" does not give the user enough information about why the search did not work (feedback), and does not give the user enough direction as to what they should do next (feed-forward).

Possible Causes/Solutions:

- Change this statement to say something such as "Select 2 or more names from the list to find joint ventures. Return to look up page."

Master problem #lf description: (minor)

Problem:
Users use the "back" button to go back to the page on which they first used the "search the IMDb" search box instead of using the search box that was available on the current page.

Possible Causes/Solutions:

- User doesn't understand that the identical search box is available on every page
- User wants to return to the homepage to start a new task instead of picking up from the current location

Master problem #lg description: (severe)

Problem:
Users have a hard time revising their searches when their first try does not work.

Possible Causes/Solutions:

- The searches do not explain how they work, particularly whether entering/selecting mutiple items will do an "and" or an "or" search
- On the search form, provide more information about how it works
- On the search results, provide more information about what search was actually done

Master problem #lh description: (minor)

Problem:
The Go button is enabled when the search field is empty. Pressing the Enter key or clicking on the Go button, when the search field is empty, does not start a search (which is the desired behavior). If the search function is not enabled, because there is no search string, then the Go button could indicate the function's state by toggling between an enabled and disabled state based on the contents of the search field.

Possible Causes/Solutions:

Master problem #li description: (minor)

Problem:
One participant encountered a lengthy delay before a search results page appeared. While the system was searching the system displayed a blank page. There was no feedback.

Possible Causes/Solutions:

- Displaying an "I'm working" page would let the participant know that the search terms had been accepted and that the system was searching the database.

Master problem #lj description: (minor)

Problem:
Numbers would be easier to read if punctuated/formatted in the normal way, such as "26,969" for "26969," and "19,633" for "19633."

Possible Causes/Solutions:

Master problem #ll description: (minor)

Problem:
From a single person's page (e.g., Owen Wilson), you can click the "Find where Owen Wilson is credited along side another name" link, even if you haven't filled in a name (i.e., the text box has no text in it). This leads to the unusual case that you are on the "Joint Venture Search" page, but have info only on one person.

Possible Causes/Solutions:

- Need to check to make sure that the text box has a valid entry. If not provide an error message.

Master problem #lo description: (severe)

Problem:
It is not clear what the "Match names with any occupation" checkbox at the bottom of the "IMDb Name Search" does or why it is only at the bottom and not repeated at the top.

Possible Causes/Solutions:

Master problem #lp description: (severe)

Problem:
The "Filmography" bar on the left-hand side of the page acts to filter the results of a query. Generally, a bar on the left-hand side of the page is a navigation bar.

Possible Causes/Solutions:

- Remove the links in the "Filmography" bar and integrate them in the content view

Create a separate page with the contents of the filmography bar and have a link to the page in the content area titled "Customize Results"

This page intentionally left blank.

APPENDIX H. Study 4 Detailed Analysis Outputs

H.1 Student vs. Practitioner Report Comments MANOVA

```
proc glm;
      title 'MANOVA across all variables in Table 3.6';
      class Type Evaluator;
      model Comments-All
            Comments-Critical
            Comments-Serious
            Comments-Minor
            Comments-Positive
            Comments-GoodIdea
            Comments-Bug
            Comments-Other
            Words
            WordsPerComment
            ImagesOrTables
            Hours
            = EvaluatorType/ SS3;
      manova h=type;
run;
```

MANOVA Test Criteria and Exact F Statistics for the
Hypothesis of No Overall **EvaluatorType** Effect
H = Type III SSCP Matrix for **EvaluatorType**
E = Error SSCP Matrix

S=1 M=4.5 N=15

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.73822585	1.03	11	32	0.4430
Pillai's Trace	0.26177415	1.03	11	32	0.4430
Hotelling-Lawley Trace	0.35459900	1.03	11	32	0.4430
Roy's Greatest Root	0.35459900	1.03	11	32	0.4430

H.2 Judges: Master Problem Severity Judgements

	A	B	C	3 agree	2/3 Vote	Why?
aa	M	M	M	M	Not	Three
ab	**S**	**M**	**S**		**Real**	**AC**
ac	**C**	**M**	**S**		**Real**	**AC**
ad	M	M	M	M	Not	Three
ae	M	M	S		Not	AB
af	M	M	M	M	Not	Three
ag	**S**	**S**	**S**	**S**	**Real**	**Three**
ah	M	M	M	M	Not	Three
ba	**S**	**C**	**C**	**1S, 2C**	**Real**	**Three**
bc	M	C	M		Not**	AC
bd	**M**	**S**	**S**		**Real**	**BC**
be	M	S	M		Not	AC
ca	**S**	**S**	**M**		**Real**	**AB**
da	**S**	**C**	**S**	**2S, 1C**	**Real**	**Three**
db	**S**	**M**	**S**		**Real**	**AC**
ea	**S**	**S**	**S**	**S**	**Real**	**Three**
eb	M	M	M	M	Not	Three
ec	M	M	M	M	Not	Three
ed	M	M	M	M	Not	Three
fa	**S**	**M**	**S**		**Real**	**AC**
fb	**S**	**M**	**S**		**Real**	**AC**
fc	**S**	**C**	**M**		**Real**	**AB**
fd	**S**	**M**	**S**		**Real**	**AC**
fe	S	M	M		Not	BC
ff	M	M	S		Not	AB
ga	M	M	S		Not	AB
gb	M	M	S		Not	AB
gc	M	M	S		Not	AB
ha	M	M	M	M	Not	Three
hb	**S**	**M**	**S**		**Real**	**AC**
hc	M	M	S		Not	AB
la	**S**	**C**	**S**	**2S, 1C**	**Real**	**Three**
lc	M	S	M		Not	AC
lf	M	M	M	M	Not	Three
lg	**S**	**C**	**S**	**2S, 1C**	**Real**	**Three**
lh	M	M	M	M	Not	Three
li	M	M	S		Not	AB
lj	M	M	M	M	Not	Three
ll	M	M	S		Not	AB
lo	**M**	**S**	**S**		**Real**	**BC**
lp	**S**	**M**	**S**		**Real**	**AC**

** Note: two minor votes and one critical vote. M = minor, S = serious, C = critical.

H.3 Judges: Reliability of Master Problem Severity Judgements

McNemar Change Test

```
proc freq;
      tables A*B A*C A*Group B*C B*Group C*Group;
      exact mcnem;
      ods output McNemarsTest;
run;
```

Comparison	$\chi^2(1)$	Exact *p*
A * B	1.14	0.42
A * C	3.77	0.09
A * Group	0.33	1.00
B * C	5.76	0.03
B * Group	2.27	0.23
C * Group	3.60	0.11

Tetrachoric Correlation

```
proc freq data=WORK.RealNot_Wide;
      tables A*B A*C A*Group B*C B*Group C*Group / RELRISK PLCORR;
      ods output measures;
run;
```

	SAS Output		Excel Calculations	
Comparison	***r****	**ASE**	***z***	***p*-value**
A * B	.44	.22	1.97	.02
A * C	.62	.18	3.45	.00
A * Group	.97	.03	33.96	.00
B * C	.07	.26	0.27	.39
B * Group	.68	.17	4.04	.00
C * Group	.78	.13	6.04	.00

Note: $z = r^*$ / ASE

H.4 Hypothesis 1a: Practitioners in Study 3 vs. Study 4

Practitioners in Study 3 vs. Study 4: Preliminary MANOVA

```
proc glm;
      class ID Guideline Study;
      model difficult required relevant helpful
            = Study|Guideline / SS3;
      manova h=Study|Guideline;
run;
```

MANOVA Test Criteria and F Approximations for the Hypothesis of No Overall **Study** Effect
H = Type III SSCP Matrix for **Study**
E = Error SSCP Matrix

S=2 M=0.5 N=572.5

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.86212545	22.08	8	2294	<.0001
Pillai's Trace	0.13909394	21.45	8	2296	<.0001
Hotelling-Lawley Trace	0.15850961	22.71	8	1636.2	<.0001
Roy's Greatest Root	0.14901818	42.77	4	1148	<.0001

MANOVA Test Criteria and F Approximations for the Hypothesis of No Overall **Guideline** Effect
H = Type III SSCP Matrix for **Guideline**
E = Error SSCP Matrix

S=4 M=2 N=572.5

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.78381672	8.02	36	4300.1	<.0001
Pillai's Trace	0.23249515	7.89	36	4600	<.0001
Hotelling-Lawley Trace	0.25556509	8.13	36	3101.8	<.0001
Roy's Greatest Root	0.14425856	18.43	9	1150	<.0001

MANOVA Test Criteria and F Approximations for the Hypothesis of No Overall **Guideline*Study** Effect
H = Type III SSCP Matrix for **Guideline*Study**
E = Error SSCP Matrix

S=4 M=6.5 N=572.5

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.91270122	1.47	72	4512.7	0.0060
Pillai's Trace	0.09003593	1.47	72	4600	0.0062
Hotelling-Lawley Trace	0.09268581	1.47	72	3674.4	0.0060
Roy's Greatest Root	0.03943948	2.52	18	1150	0.0004

Practitioners in Study 3 vs. Study 4: Four ANOVAs

```
/* Note: same dataset as the MANOVA, but in tall format */
/* to enable the "by Adjective" statement               */
proc mixed;
      class ID Guideline Study;
      model Rating = Study|Guideline;
      lsmeans Study*Guideline / adjust=tukey slice=Guideline;
      repeated / subject=ID type=cs;
      by Adjective;
run;
```

Adjective	Effect	NumDF	DenDF	FValue	ProbF
difficult	Study	1	93	0.89	.35
	Guideline	9	837	12.31	<.0001*
	Guideline*Study	9	837	2.38	.01*
helpful	Study	1	93	0.4	.53
	Guideline	9	837	13.01	<.0001*
	Guideline*Study	9	837	1.23	.27
relevant	Study	1	93	0.82	.37
	Guideline	9	837	10.86	<.0001*
	Guideline*Study	9	837	0.93	.50
required	Study	1	93	27.47	<.0001*
	Guideline	9	837	12.61	<.0001*
	Guideline*Study	9	837	0.35	.96

* $p < .05$.

Post-hoc tests for Guideline*Study for *difficult* using effect slices by Guideline, DF=1, 837

Guideline	FValue	ProbF
Backing Data	1.57	.21
Clarity/Jargon	2.37	.12
Impact/Severity	0.88	.35
Methodology	0.42	.52
Politics/Diplomacy	11.36	.00*
Problem Cause	0.33	.57
Professional/Scientific	0.26	.61
Describe a Solution	0.21	.64
Evoke Sympathy	0.02	.88
User Actions	1.26	.26

* $p < .05$.

H.5 Hypothesis 1b: Study 3 Practitioners vs. Study 4 Students

Study 3 Practitioners vs. Study 4 Students : Preliminary MANOVA

```
proc glm data=Ratings3n4Wide;
      class ID Guideline Study;
      model difficult required relevant helpful = Study|Guideline /
SS3;
      manova h=Study|Guideline;
run;
```

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall Study Effect
H = Type III SSCP Matrix for Study
E = Error SSCP Matrix

S=1 M=1 N=207.5

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.91504556	9.68	4	417	<.0001
Pillai's Trace	0.08495444	9.68	4	417	<.0001
Hotelling-Lawley Trace	0.09284176	9.68	4	417	<.0001
Roy's Greatest Root	0.09284176	9.68	4	417	<.0001

MANOVA Test Criteria and F Approximations for the Hypothesis of No Overall Guideline Effect
H = Type III SSCP Matrix for Guideline
E = Error SSCP Matrix

S=4 M=2 N=207.5

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.69527744	4.43	36	1564.4	<.0001
Pillai's Trace	0.33855036	4.31	36	1680	<.0001
Hotelling-Lawley Trace	0.39156021	4.52	36	1120.4	<.0001
Roy's Greatest Root	0.23087990	10.77	9	420	<.0001

MANOVA Test Criteria and F Approximations for the Hypothesis of No Overall Guideline*Study Effect
H = Type III SSCP Matrix for Guideline*Study
E = Error SSCP Matrix

S=4 M=2 N=207.5

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.90197376	1.21	36	1564.4	0.1814
Pillai's Trace	0.10112445	1.21	36	1680	0.1837
Hotelling-Lawley Trace	0.10528267	1.22	36	1120.4	0.1800
Roy's Greatest Root	0.05768015	2.69	9	420	0.0047

Study 3 Practitioners vs. Study 4 Students : Four ANOVAs

```
/* Note: same dataset as the MANOVA, but in tall format */
proc mixed data=Ratings3n4;
      class ID Guideline Study;
      model Rating = Study|Guideline;
      lsmeans Study*Guideline / adjust=tukey slice=Guideline;
      repeated / subject=ID type=cs;
      by Adjective;
run;
```

Adjective	Effect	NumDF	DenDF	FValue	ProbF
difficult	Study	1	95	1.49	.23
	Guideline	9	855	11.82	<.0001*
	Guideline*Study	9	855	2.21	.02*
helpful	Study	1	95	2.78	.10
	Guideline	9	855	13.15	<.0001*
	Guideline*Study	9	855	3.15	.00*
relevant	Study	1	95	1.85	.18
	Guideline	9	855	10.74	<.0001*
	Guideline*Study	9	855	1.26	.25
required	Study	1	95	2.07	.15
	Guideline	9	855	8.9	<.0001*
	Guideline*Study	9	855	1.06	.39

* $p < .05$.

Post-hoc tests for Guideline*Study for *difficult* and *helpful* using effect slices by Guideline, DF=1, 855

Adjective	QualName	FValue	ProbF
difficult	Backing Data	0.54	.46
	Clarity/Jargon	0.3	.59
	Impact/Severity	1.15	.28
	Methodology	11.97	.00*
	Politics/Diplomacy	0.45	.50
	Problem Cause	0.98	.32
	Professional/Scientific	2.91	.09
	Describe a Solution	0.99	.32
	Evoke Sympathy	0	.99
	User Actions	0.7	.40
helpful	Backing Data	0.31	.58
	Clarity/Jargon	1.43	.23
	Impact/Severity	3.72	.05
	Methodology	2.37	.12
	Politics/Diplomacy	11.48	.00*
	Problem Cause	0.46	.50
	Professional/Scientific	9.4	.00*
	Describe a Solution	1.52	.22
	Evoke Sympathy	0.02	.89
	User Actions	0	.95

* $p < .05$.

H.6 Hypothesis 1d: Opinion vs. Behavior

```
proc corr PEARSON NOSIMPLE;
      var Opinion Judgement;
      by Guideline Adjective;
run;
```

Guideline	Adjective	*r*	*p*
Backing Data	difficult	0.31	0.17
Clarity/Jargon	difficult	0.16	0.48
Impact/Severity	difficult	-0.08	0.74
Problem Cause	difficult	0.14	0.53
Solution	**difficult**	**-0.51**	**0.02***
User Actions	difficult	-0.02	0.92
Backing Data	helpful	-0.19	0.42
Clarity/Jargon	helpful	0.31	0.17
Impact/Severity	helpful	0.25	0.27
Problem Cause	helpful	-0.02	0.94
Solution	**helpful**	**0.58**	**0.01***
User Actions	helpful	0.16	0.48
Backing Data	relevant	0.03	0.90
Clarity/Jargon	relevant	0.27	0.24
Impact/Severity	relevant	0.35	0.12
Problem Cause	relevant	-0.03	0.89
Solution	**relevant**	**0.57**	**0.01***
User Actions	relevant	0.20	0.39
Backing Data	required	-0.11	0.63
Clarity/Jargon	required	-0.06	0.79
Impact/Severity	required	0.24	0.30
Problem Cause	required	-0.11	0.64
Solution	**required**	**0.67**	**0.00***
User Actions	required	-0.03	0.89

* $p < .05$

H.7 Hypothesis 2a: Problem Discovery

Problem Discovery: Thoroughness

```
proc mixed;
      class ID Measure Evaluator;
      model Thoroughness = Evaluator Measure Evaluator*Measure;
      repeated / subject=ID Measure=cs;
      lsmeans Measure / adjust=tukey;
run;
```

Measure 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
Evaluator	1	42	0.31	0.5778
Measure	2	84	61.47	<.0001
Evaluator*Measure	2	84	1.14	0.3250

Post-hoc tests for Measure, DF = 84, Adjustment = Tukey-Kramer

Measure	_Measure	Estimate	StdErr	tValue	Probt	Adjp
All	Severe	-0.1151	0.01344	-8.56	<.0001	<.0001
All	SMS	0.02446	0.01344	1.82	0.0724	0.1694
Severe	SMS	0.1396	0.01344	10.38	<.0001	<.0001

Problem Discovery: Reliability

```
proc mixed;
      class Evaluator Measure;
      model Jaccard = Evaluator Measure Evaluator*Measure;
run;
```

Post-hoc tests for Measure*Evaluator, DF=5667, Adjustment=Tukey-Kramer

Evaluator	Measure	_Evaluator	_Measure	Estimate	StdErr	tValue	Probt	Adjp
between	All	between	SMS	0.1096	0.007648	14.33	<.0001	<.0001
between	All	between	Sev	-0.1105	0.007648	-14.45	<.0001	<.0001
between	SMS	between	Sev	-0.2201	0.007648	-28.78	<.0001	<.0001
between	All	practitioner	All	-0.00634	0.009824	-0.65	0.5186	0.9993
between	All	practitioner	SMS	0.02485	0.009824	2.53	0.0114	0.2179
between	All	practitioner	Sev	-0.141	0.009824	-14.35	<.0001	<.0001
between	Sev	practitioner	All	0.1042	0.009824	10.6	<.0001	<.0001
between	Sev	practitioner	SMS	0.1354	0.009824	13.78	<.0001	<.0001
between	Sev	practitioner	Sev	-0.03049	0.009824	-3.1	0.0019	0.0499
between	SMS	practitioner	All	-0.1159	0.009824	-11.8	<.0001	<.0001
between	SMS	practitioner	SMS	-0.08473	0.009824	-8.62	<.0001	<.0001
between	SMS	practitioner	Sev	-0.2506	0.009824	-25.51	<.0001	<.0001
between	All	student	All	-0.00895	0.009224	-0.97	0.3319	0.9885
between	All	student	SMS	0.1391	0.009224	15.08	<.0001	<.0001
between	All	student	Sev	-0.08778	0.009224	-9.52	<.0001	<.0001
between	Sev	student	All	0.1016	0.009224	11.01	<.0001	<.0001
between	Sev	student	SMS	0.2497	0.009224	27.07	<.0001	<.0001
between	Sev	student	Sev	0.02275	0.009224	2.47	0.0137	0.2489
between	SMS	student	All	-0.1185	0.009224	-12.85	<.0001	<.0001
between	SMS	student	SMS	0.02955	0.009224	3.2	0.0014	0.0368
between	SMS	student	Sev	-0.1974	0.009224	-21.4	<.0001	<.0001
practitioner	All	practitioner	SMS	0.03119	0.0116	2.69	0.0072	0.1517
practitioner	All	practitioner	Sev	-0.1347	0.0116	-11.61	<.0001	<.0001
practitioner	SMS	practitioner	Sev	-0.1659	0.0116	-14.3	<.0001	<.0001
practitioner	All	student	All	-0.00261	0.0111	-0.24	0.8141	1
practitioner	All	student	SMS	0.1455	0.0111	13.11	<.0001	<.0001
practitioner	All	student	Sev	-0.08144	0.0111	-7.34	<.0001	<.0001
practitioner	Sev	student	All	0.1321	0.0111	11.9	<.0001	<.0001
practitioner	Sev	student	SMS	0.2801	0.0111	25.25	<.0001	<.0001
practitioner	Sev	student	Sev	0.05324	0.0111	4.8	<.0001	<.0001
practitioner	SMS	student	All	-0.0338	0.0111	-3.05	0.0023	0.059
practitioner	SMS	student	SMS	0.1143	0.0111	10.3	<.0001	<.0001
practitioner	SMS	student	Sev	-0.1126	0.0111	-10.15	<.0001	<.0001
student	All	student	SMS	0.1481	0.01057	14.01	<.0001	<.0001
student	All	student	Sev	-0.07883	0.01057	-7.46	<.0001	<.0001
student	SMS	student	Sev	-0.2269	0.01057	-21.47	<.0001	<.0001

H.8 Hypotheses 3a, 3b: Differences in Following Guidelines

```
proc mixed;
      class Judge Guideline ID group;
      model Rating =
            Judge Guideline Group
            Judge*Guideline Judge*Group Guideline*Group
            Judge*Guideline*Group;
      repeated / subject=ID type=cs;
      lsmeans Group*Guideline / slice=Guideline;
      lsmeans Judge / adjust=tukey;
      lsmeans Judge*Guideline / slice=Guideline;
run;
```

Effect	NumDF	DenDF	FValue	ProbF
Judge	2	84	14.01	<.0001*
Guideline	5	210	30.34	<.0001*
Group	1	42	7.27	.01*
Judge*Guideline	10	420	5.37	<.0001*
Judge*Group	2	84	1.65	.20
Guideline*Group	5	210	3.13	.01*
Judge*Guideline*Group	10	420	0.7	.72

* $p < .05$.

Post-hoc tests for Group*Guideline using effect slices by Guideline, DF=1, 210

Guideline	FValue	ProbF
Backing Data	11.53	0.00
Clarity/Jargon	0.01	0.93
Impact/Severity	7.36	0.01
Problem Cause	1.75	0.19
Provide a Solution	9.05	0.00
User Actions	0.05	0.82

Post-hoc tests for Judge, DF=84, Adjustment=Tukey-Kramer

Judge	_Judge	Estimate	StdErr	tValue	Probt	Adjp
A	B	0.43	0.10	4.23	<.0001	.00
A	C	-0.07	0.10	-0.64	.52	.80
B	C	-0.49	0.10	-4.87	<.0001	<.0001

Post-hoc tests for Judge*Guideline using effect slices by Guideline, DF=2, 420

Guideline	FValue	ProbF
Backing Data	6.92	.00*
Clarity/Jargon	3.27	.04*
Impact/Severity	0.24	.79
Problem Cause	1.55	.21
Provide a Solution	5.74	.00*
User Actions	23.16	<.0001*

* $p < .05$.

Post-hoc tests for Judge*Guideline using lsmeans for *Backing Data*, *Clarity/Jargon*, *Provide a Solution*, and *User Actions* DF=420

Judge	_Judge	Guideline	tValue	Probt	Adjp
A	B	Backing Data	3.7	.00	.03*
A	C	Backing Data	2.17	.03	.77
B	C	Backing Data	-1.53	.13	.99
A	B	Clarity Jarg	-0.66	.51	1
A	C	Clarity Jarg	-2.47	.01	.56
B	C	Clarity Jarg	-1.81	.07	.94
A	B	Solutions	3.04	.00	.19
A	C	Solutions	0.22	.83	1.00
B	C	Solutions	-2.82	.01	.31
A	B	User Actions	6.19	<.0001	<.0001*
A	C	User Actions	0.64	.52	1
B	C	User Actions	-5.55	<.0001	<.0001*

```
proc corr;
      var   Hours       YearsExp    NumEvals
            ValidityAll
            ThorAll     ThorSevere  ThorSMS
            ClarityJargon     ImpactSeverity
            BackingData       ProblemCause
            UserActions       Solution;
run;
```

r ***p***	**Clarity Jargon**	**Impact Severity**	**Backing Data**	**Problem Cause**	**User Actions**	**Solution**
Hours	.10	**.36**	.24	.10	.28	-.14
	.52	**.02***	.12	.52	.06	.36
YearsExp	.18	**.30**	.12	.13	-.19	.24
	.25	**.05***	.45	.41	.21	.11
NumEvals	.20	.17	.12	.01	-.02	-.01
	.20	.28	.43	.94	.91	.95
ValidityAll	.06	.19	.12	.08	.01	.21
	.72	.23	.43	.59	.92	.18
ThorAll	-.02	.03	.12	.06	.26	-.04
	.88	.84	.45	.70	.08	.79
ThorSevere	.02	.17	.17	.13	.24	.08
	.91	.26	.26	.39	.12	.58
ThorSMS	-.21	.19	.24	**.32**	**.34**	.05
	.16	.22	.12	**.03***	**.02***	.75

* $p < .05$.

H.9 Differences in Opinion Between Study 4 Students, Practitioners

```
proc mixed data=Ratings3n4Tall;
      class ID Guideline Group;
      model Rating = Group|Guideline / OUTP=MirOutP;
      lsmeans Group*Guideline / adjust=tukey slice=Guideline;
      repeated / subject=ID type=cs;
      by Adjective;
run;
```

Adjective	Effect	NumDF	DenDF	FValue	ProbF
difficult	Group	1	42	3.59	.07
	Guideline	9	378	5.06	<.0001*
	Guideline*Group	9	378	2.17	.02
helpful	Group	1	42	0.61	.44
	Guideline	9	378	9.36	<.0001*
	Guideline*Group	9	378	0.94	.49
relevant	Group	1	42	0.15	.70
	Guideline	9	378	7.31	<.0001*
	Guideline*Group	9	378	0.71	.70
required	Group	1	42	11.25	.00*
	Guideline	9	378	4.62	<.0001*
	Guideline*Group	9	378	0.52	.86

* $p < .05$.

Post-hoc tests for Guideline*Group for *required* using effect slices by Guideline, DF=1, 378

Guideline	FValue	ProbF
Backing Data	2.7	.10
Clarity/Jargon	6.72	.01*
Impact/Severity	2.05	.15
Methodology	0.05	.82
Politics/Diplomacy	1.38	.24
Problem Cause	3.04	.08
Professional/Scientific	4.6	.03*
Describe a Solution	2.38	.12
Evoke Sympathy	0.41	.52
User Actions	2.57	.11

* $p < .05$.