Behaviour & Information Technology, vol. 33, no. 2 (2014), pp. 143-161
Preprint version

What You Get Is What *You* See: Revisiting the Evaluator Effect in Usability Tests

Morten Hertzum

Computer Science, Roskilde University, Roskilde, Denmark, mhz@ruc.dk

Rolf Molich

DialogDesign, Stenløse, Denmark, molich@dialogdesign.dk

Niels Ebbe Jacobsen

Danish Consumer Council, Copenhagen, Denmark, niels.ebbe.jacobsen@gmail.com

Abstract. Usability evaluation is essential to user-centred design, yet evaluators who analyse the same usability test sessions have been found to identify substantially different sets of usability problems. We revisit this evaluator effect by having 19 experienced usability professionals analyse video-recorded test sessions with five users. Nine participants analysed moderated sessions; ten participants analysed unmoderated sessions. For the moderated sessions, participants reported an average of 33% of the problems reported by all nine of these participants and 50% of the subset of problems reported as critical or serious by at least one participant. For the unmoderated sessions, the percentages were 32% and 40%. Thus, the evaluator effect was similar for moderated and unmoderated sessions, and it was substantial for the full set of problems and still present for the most severe problems. In addition, participants disagreed in their severity ratings. As much as 24% (moderated) and 30% (unmoderated) of the problems reported by multiple participants were rated as critical by one participant and minor by another. The majority of the participants perceived an evaluator effect when merging their individual findings into group evaluations. We discuss reasons for the evaluator effect and recommend ways of managing it.

Keywords: usability evaluation, usability test, thinking-aloud test, evaluator effect, problem detection, severity rating

1 INTRODUCTION

Evaluation is essential to the design of usable systems. This was recognised early by, for example, Lewis (1982) and has recently been reiterated by Siegel and Dray (2011). To conduct evaluations usability professionals need reliable and robust usability evaluation methods. A number of methods have been developed, including cognitive walkthrough (Wharton, Rieman, Lewis, & Polson, 1994), constructive interaction (O'Malley, Draper, & Riley, 1984), heuristic evaluation (Nielsen & Molich, 1990), metaphors of human thinking (Frøkjær & Hornbæk, 2008), and usability tests (Dumas & Redish, 1999). The usability test has long had a prominent position among these methods in that it is by some considered the single most important usability evaluation method (Gulliksen, Boivie, Persson, Hektor, & Herulf, 2004; Nielsen, 1993) and has been used as a yardstick for other usability evaluation methods (Bailey, Allan, & Raiello, 1992; John & Marks, 1997). The prominent position of the usability test warrants careful scrutiny of this method to understand its strengths and learn to stay within, or compensate for, its limitations.

This study scrutinises the usability test by revisiting the evaluator effect, which was first reported by Jacobsen et al. (1998a, 1998b). The evaluator effect is the observation that usability evaluators who analyse the same usability test sessions identify substantially different sets of usability problems. That is, you get what *you* see in the double sense that as a usability evaluator you report the set of problems for which you have seen evidence, but simultaneously other evaluators see the same test sessions as evidence of different sets of problems. The implications of this observation have been put strongly by Lewis (2001, p. 346), who wrote that the evaluator effect raises "the possibility that usability practitioners might be engaging in self-deception regarding the reliability of their problem-discovery methods". At the same time, the evaluator effect must be reconciled with the widespread experience of real improvement achieved through the use of usability tests. For example, Bailey (1993) found a reduction in the number of serious errors experienced by users from the first to the last iteration of evaluation and redesign.

The aim of this study is threefold. We aim to investigate:

- *Whether the evaluator effect still exists*. Previous studies of the evaluator effect suffer from few evaluators (Vermeeren, van Kesteren, & Bekker, 2003), inexperienced evaluators (Hornbæk & Frøkjær, 2008), and differences in the data analysed by the evaluators (Molich et al., 1998). We address all three of these issues and, especially, control the data analysed by the evaluators in order to improve the internal validity of our study and thereby ensure that differences in the problems reported by different evaluators are evidence of an evaluator effect.
- *Whether moderation of the test sessions affects the evaluator effect*. Previous studies have investigated the evaluator effect for tests with a human moderator administering the sessions, but lately unmoderated test sessions, in which users are unsupervised, have become increasingly popular. The absence of prompting in unmoderated sessions may make the user's verbalisations less informative about their experience, thereby increasing evaluators' uncertainty and disagreements.
- *Whether evaluators perceive the presence of an evaluator effect when they are involved in a test*. Previous work on usability inspections have indicated that the evaluator effect may be contested by evaluators, who tend to perceive a high level of agreement even when an independent analysis of the data indicates a substantial evaluator effect (Hertzum, Jacobsen, & Molich, 2002). This point is practically important because steps to manage an evaluator effect will likely be taken only if evaluators perceive it.

To investigate these issues we conduct an empirical study in which 19 experienced usability specialists analyse video-recorded test sessions. They report the usability problems they individually identify, and meet to merge their individual evaluations into group evaluations. On this basis, we analyse the evaluator effect and discuss its causes.

2 RELATED WORK

In this study we focus on the evaluator effect in usability tests that involve users who verbalise their thoughts, but an evaluator effect has also been found for other usability evaluation methods such as cognitive walkthrough and heuristic evaluation (Hertzum & Jacobsen, 2003). In the following, we briefly describe the kind of usability tests targeted in this study, review the evaluator effect, and draw attention to the related Rashomon effect.

2.1 Usability Tests

After Lewis (1982) introduced thinking aloud for use in usability tests, numerous variants of this usability evaluation method have been employed. Today, multiple practitioner's guides provide descriptions of different variants of the method (e.g., Dumas & Loring, 2008; Dumas & Redish, 1999; Rubin & Chisnell, 2008). While there is no single accepted procedure for usability specialists to follow, Clemmensen et al. (2009) proposed a simplified model of usability tests to point out the main elements shared by the variants of the method. The model comprises four elements:

- Users interact with the system in order to solve a set of tasks prepared ahead of the test.
- Users verbalise their thoughts while solving the tasks. To prompt the verbalisation users are reminded to keep talking or are asked questions about their behaviour.
- An evaluator observes the users' behaviour and listens in on their thoughts. On this basis the evaluator identifies and reports usability problems.
- The evaluation takes place in the context of an overall relationship between users and evaluator. To obtain reliable evaluation results the users must feel at ease.

The evaluator effect concerns the third of these elements. The issue of moderated or unmoderated test sessions concerns the second element, particularly whether prompting is possible. The first and last elements are held constant in our empirical study.

2.2 The Evaluator Effect

Hertzum and Jacobsen (2003) collected evidence of the evaluator effect for multiple usability evaluation methods, for novice and experienced evaluators, for simple and complex systems, for minor and severe problems, and for problem detection and severity assessment. They proposed that the principal reason for the evaluator effect is that usability evaluation is an interpretive activity in which evaluators need to exercise judgement in transitioning from a sequence of user-system interactions to a list of usability problems. It is unsurprising that such judgements do not produce the exact same results when performed by different evaluators. What may be surprising is the magnitude of the evaluator effect. In multiple studies the number of problems detected by only a single evaluator has clearly exceeded the number of problems shared by all evaluators (e.g., Jacobsen et al., 1998b; Kessner, Wood, Dillon, & West, 2001; Molich & Dumas, 2008).

Common measures of the evaluator effect are the detection rate and the any-two agreement. The detection rate is the average number of problems detected by a single evaluator in percent of the total number of problems detected by all evaluators. However, this measure inflates the agreement among evaluators, especially in studies with few evaluators (Hertzum & Jacobsen, 2003). This inflation is avoided by the any-two agreement, thereby allowing for direct comparisons of any-two agreements from studies with different numbers of evaluators. The any-two agreement is the number of problems shared by a pair of evaluators in percent of the number of problems they collectively detect, averaged over all pairs of evaluators.

Jacobsen et al. (1998b) reported a detection rate of 52% and an any-two agreement of 42% from a usability test in which four evaluators with varying experience analysed four video-recorded test sessions. In addition, the evaluators differed substantially in their ratings of problem severity (Jacobsen et al., 1998a). The evaluator effect found in this study was substantial, even though the evaluators received a list with nine criteria that defined when an observation should be recorded as a usability problem. They had also access to the system to enable them to try out issues unclear in the video recordings. A couple of studies (Molich et al., 1998; Molich, Ede, Kaasgaard, & Karyukin, 2004) have reported even less agreement in the problems identified by different evaluators, but in these studies the evaluators conducted their own test sessions, rather than analysed the same test sessions.

Attempts at learning to manage the evaluator effect have focused on identifying the ways in which interpretation becomes an element in usability evaluation. One reason is that usability evaluation methods provide vague evaluation procedures, vague problem criteria, and vague support for determining the goals against which to test a system (Hertzum & Jacobsen, 2003). Another reason is that evaluators may have little domain knowledge. Specifically for usability work in complex domains, Chilana et al. (2010) showed that evaluators are often dependent on partnering with domain experts in their interpretation of whether an observation constitutes a usability problem. A third reason is the strength of the evidence available for deciding whether a usability problem exists. For example, Vermeeren et al. (2003) found that the most frequent cause of disagreement between the evaluators in their study was usability problems based solely on verbalisations or facial expressions.

Vermeeren et al. (2003) also identified four other causes of disagreement: inaccuracy in determining user intentions, different thresholds for the amount of inefficiency that triggers a usability problem, uncertainty about whether a candidate usability problem is an artefact of the evaluation, and errors in logging interaction data or hearing user verbalisations. In their study several systems were evaluated by two evaluators with resulting any-two agreements of 53%, 64%, 64%, 69%, and 81%. No information was provided about the evaluators.

Hornbæk and Frøkjær (2008) reported an any-two agreement of 42% in a study with 50 computer science students as evaluators. The main result of this study was, however, that a substantial source of the disagreements was variability in the matching process through which the evaluators merged their individual findings into group findings. This result led the authors to question previous assertions about the evaluator effect, arguing that the matching process used to calculate the evaluator effect may be as variable as the evaluator effect itself. Notably, Hornbæk and Frøkjær did not argue against the presence of an evaluator effect; they qualified the evidence about its size and implied that it might be inflated in previous studies.

The size of the evaluator effect has also been contested by the usability professionals who conduct evaluations. In a study of usability inspections conducted by 11 experienced usability professionals, Hertzum et al. (2002) found that these evaluators reported disparate usability observations yet perceived a high level of agreement about the usability issues in the system. This finding suggests a hierarchy with detailed usability observations that provide supporting evidence for higher-level usability issues. Gorlenko and Englefied (2006) made a similar distinction and proposed that developers tend to focus on detailed usability observations, which may help fix individual errors, whereas client executives tend to focus on high-level usability issues. The size of the evaluator effect perceived by these two groups of practitioners is therefore likely to differ.

Whereas the studies mentioned above concern usability tests in which the users and a human moderator were collocated (though possibly separated by a one-way mirror), usability evaluation may also be done remotely. Remote usability evaluation may be synchronous with a live audio or video connection between moderator and user (McFadden, Hager, Elie, & Blackwell, 2002) or asynchronous with no moderation of the user sessions (Bruun, Gull, Hofmeister, & Stage, 2009). That is, unmoderated usability tests are an instance of asynchronous remote usability testing. In the dominant approaches to asynchronous remote evaluation the identification of usability issues is left to the users, rather than performed by evaluators. Bruun et al. (2009) had three evaluators analyse and merge the usability issues reported by the users and found any-two agreements of 53%, 61%, and 69% among the evaluators for three asynchronous remote evaluation methods. It must be assumed that starting from identified usability issues likely leads to higher any-two agreements than an analysis of videos of user sessions. Nelson and Stavrou (2011) made remote asynchronous usability tests by having users solve tasks and rate usability and by analysing logs of users' clicks and key presses. In this setup, usability issues were identified by evaluators, rather than users, though on the basis of restricted data. The authors, however, did not investigate the evaluator effect.

2.3 The Rashomon Effect

In the social sciences an effect similar to the evaluator effect is known as the Rashomon effect, named after Kurosawa's film from 1950. The film presents four contradictory accounts of an event and leaves the viewer wondering which of the four accounts is true and whether a single true account really exists. The former question is misguided from an interpretivist point of view; the latter question is discomforting from a positivist point of view (Heider, 1988; Roth & Mehta, 2002). Research on the Rashomon effect investigates how contradictory accounts come about and how they may be reconciled. A central issue in this research is how to combine positivist and interpretivist points of view on whether a shared reality exists. Heider (1988, p. 74) proposed the middle ground that "there is a shared reality, true, but differing truths may indeed be said about it." The Rashomon effect is a reminder that contradictory accounts are quite common and that the evaluator effect is but one instance of such contradictory accounts.

3 METHOD

To revisit the evaluator effect we conducted an empirical study in which usability specialists evaluated a website by analysing video-recorded test sessions. The study was the ninth in the series of Comparative Usability Evaluation (CUE) studies. Supplementary material about the study, including the participants' test reports, is available at the CUE website, www.dialogdesign.dk/CUE.html.

3.1 Participants

Nineteen usability specialists participated in the study as evaluators. The participants, who are listed in the Acknowledgements section, comprised 14 industrial usability professionals and 5 usability researchers affiliated with universities and research institutions. Participants had an average of 17 years of experience in doing usability evaluations and had conducted an average of over 100 usability tests, see Table 1. With this background, the participants were deemed to be experienced usability evaluators.

The participants were recruited through the authors' professional networks. Twelve of the participants had participated in previous CUE studies and one of the industrial usability professionals was a representative of the company whose website was evaluated in this study. The second and third authors participated in the study as evaluators.

3.2 Website and Test Tasks

We chose the U-Haul website (www.uhaul.com) for the evaluation because it was an example of a comprehensive yet common type of e-commerce site, because it offered a service understandable to evaluators without special domain knowledge, and because U-Haul was willing to provide supplementary information about the website and its use. At the U-Haul website users can rent moving trucks, self-storage units, and equipment such as moving boxes and dollies. On the basis of an informal exploration of the website and communication with U-Haul, a scenario and seven tasks were devised for the test. The scenario read:

Your friends Mike and Anna are about to move from Pittsburgh, PA, to Denver, CO. They have an apartment in Pittsburgh consisting of a living room, a bedroom, a kitchen, and a bathroom. They want to find the cheapest service for the move to Colorado. They expect to make the move themselves with some help from a few friends.

The seven test tasks are listed in Table 2.

3.3 Videos of Test Sessions

The participants performed their evaluation of the U-Haul website by watching video recordings of five test sessions. In each session a different user solved the test tasks. We chose video-recorded test sessions to ensure that the participants got the same input for their evaluations. This way the differences between the findings reported by different participants can be ascribed to the evaluator effect and not to, for example, differences in the setup of the usability test or the background of the users enrolled in the test. Two sets of video-recorded test sessions were conducted: moderated and unmoderated. A participant evaluated the website on the basis of either one or the other set.

The *moderated* test sessions were run by a usability professional from Fidelity Investments, who volunteered to conduct the sessions. In these sessions a moderator instructed the users before they started solving the tasks and probed the users for information while they were solving the tasks. Before the user started solving the tasks, the moderator left the test room and entered a control room from which the moderator could see the user but not vice versa. For the remainder of the session the user and moderator communicated via an audio link between the two rooms. Users were instructed to verbalise their thoughts and did so with little need for reminders. Users had the descriptions of the tasks available on paper for easy reference.

The video recordings of the moderated sessions showed the screen with the U-Haul website and an inset in the lower right corner with the user's face. In addition, the audio of the recordings gave the user's verbalisations and the moderator's instructions and probing questions. The five moderated sessions lasted an average of 37 minutes (range: 27-48 minutes).

The *unmoderated* test sessions were conducted using the crowdsourcing service UserTesting.com. For these sessions the scenario, the tasks, a link to the U-Haul website, and a user profile were uploaded to UserTesting.com. On the basis of the user profile, which comprised gender, age, country, and web expertise, matching users from the UserTesting.com database were invited to the test and, if they accepted, performed the test session online. Users in the database had been screened for their ability to verbalise their thoughts and they were instructed to verbalise during test sessions. Indeed, the users in the unmoderated sessions kept up a running commentary while they solved the tasks. The users had the U-Haul website and the tasks continuously available on screen. When they felt they had completed a task, they proceeded to the next task; there was no human moderator in the unmoderated sessions.

The video recordings of the unmoderated sessions showed the screen with the U-Haul website and an inset at the upper right with the current task. And, the video recordings also gave the user's verbalisations. The five unmoderated sessions lasted an average of 33 minutes (range: 22-42 minutes).

Both moderated and unmoderated test sessions were conducted in late March 2011 and, thus, evaluated the U-Haul website as of that time.

3.4 Procedure

Participation in the study comprised three activities: an individual evaluation of the website, a group evaluation that consisted of merging individual evaluations, and a plenum discussion.

The *individual evaluation* involved analysing a set of five video-recorded test sessions and documenting the findings in a written report. Nine participants received the moderated sessions; ten participants received the unmoderated sessions. Participants could pause, replay, and revisit the video recordings as many times as they wished and they were encouraged to use their normal procedures for analysing the test sessions. Participants were, however, requested to document the usability issues on the website in a standardized format. This format prescribed that each usability finding was documented by recording:

- A unique identification of the finding
- A textual description of the finding
- A rating of the finding using the categories in Table 3
- The location(s) in the video recordings of the events on which the finding was based

Participants were also requested to describe the way in which they analysed the test sessions, their personal familiarity with truck and storage rental, and the time they spent performing their evaluation.

In total, the material provided to the participants comprised written instructions, the test tasks, a template for documenting usability findings, five video-recorded test sessions, and background information about the users. After receiving this material the participants had a period of six weeks to conduct and document their evaluation. The authors did not observe or interact with participants during this period. Participants spent an average of 22 hours (range: 11 - 56 hours) analysing the test sessions and documenting their individual evaluation.

The *group evaluation* lasted 3 hours and was conducted during the first half of a workshop that convened all but one of the participants for a full day in June 2011. For the group evaluation the participants were divided into four groups of 4-5 persons. Two groups consisted of the participants who had analysed moderated sessions; the two other groups consisted of the participants who had analysed unmoderated sessions. The groups received printed sheets with the usability findings that had been (a) reported by the members of the group in their individual evaluations and (b) rated as critical, serious, or a bug in these individual evaluations. We restricted the group evaluation to a subset of the

reported findings to make the number of findings manageable to the groups. The four groups received 88, 80, 58, and 53 findings, each printed on a separate sheet of paper. The groups were requested to walk through these findings, identify those that reported the same problem, and agree on a rating of each problem as either critical, serious, minor, or not a usability problem after all. In other words, the groups were requested to merge their individual evaluations into a group evaluation.

The second half of the workshop was a *plenum discussion* of the participants' experiences from the individual and group evaluations. Participants discussed whether the group evaluations added new insights and whether the participants had perceived an evaluator effect. This discussion was followed by a more general discussion of the evaluator effect and its causes.

3.5 Data Pre-processing

The 19 participants reported a total of 836 usability findings from their individual evaluations. These findings were analysed to identify the findings that reported the same usability issue and, thereby, produce a list of the different usability issues collectively reported by the participants. Two reported findings were considered instances of the same usability issue if the same revision of the U-Haul website would remedy both findings and different if a revision would remedy one of the findings but not the other (Molich & Dumas, 2008). The analysis consisted of four steps:

Initially, the first author grouped all 836 findings into usability issues and the second author, independently, grouped the 404 findings that evaluators had rated as critical, serious, bugs, or positive. This process involved splitting a small number of findings that reported multiple issues into one finding for each issue, thereby increasing the total number of findings to 860. Out of the 404 findings grouped by both authors, they agreed on their grouping of 74%.

Second, the two authors resolved all disagreements through a discussion that continued until consensus was reached. This grouping and consensus process resulted in a list of 227 usability issues, each reported by one or more participants.

Third, to crosscheck this list the third author, independently, allocated each of the 860 findings to one of the 227 usability issues. The Kappa value of the agreement between this allocation of the findings and that produced in the second step was 0.73, which indicates "substantial" agreement (Landis & Koch, 1977) and is well above the recommended minimum of 0.60 (Lazar, Feng, & Hochheiser, 2010).

Fourth, the two first authors resolved all remaining disagreements through a discussion that continued until consensus was reached. To avoid overestimating the evaluator effect we applied the conservative rule of resolving disagreements by combining findings into fewer usability issues each reported by more participants. The application of this rule led to the inclusion of 43 issues in already existing issues.

The grouping process resulted in a group of 55 findings that were too unclear to be comprehensible and another group of 6 findings that concerned a moderator error rather than a problem with the U-Haul website. These 61 findings were excluded from our further analysis. The remaining 799 findings were included in our analysis. They comprised 182 usability issues.

4 RESULTS

In the following we analyse the data from the individual evaluations, the group evaluations, and the plenum discussion. Figure 1 gives an overview of the data entering into these analyses.

4.1 Individual Evaluations

The 799 findings were rated by the participants as critical (76), serious (159), minor (298), bugs (33), ideas (105), and positive (128). While all findings rated as critical, serious, minor, or bugs were problems and all findings rated as positive were positive issues, the findings rated as ideas contained both problems (99) and positive issues (6). The 182 usability issues comprised 134 usability problems

and 48 positive issues. A matrix showing the participants who reported the different usability issues is included in the Appendix. The average number of usability problems reported by a participant was 32.9 (range: 6-47) for the participants analysing moderated sessions and 23.1 (range: 5-35) for the participants analysing unmoderated sessions. For example, 16 participants reported the problem that users got confused when storage units appeared to be free of charge. One of the findings reporting this problem was: *On one occasion, the storage units were shown at $0 per month. "At this point I would go to another web site, or pick up the phone and call them."* Different participants rated this problem as critical (3), serious (1), minor (3), or a bug (9).

The nine participants analysing moderated sessions reported substantially different sets of problems and even more different sets of positive issues, as evidenced by their detection rates and any-two agreements, see Table 4. Participants reported an average of 22% of the positive issues and 33% of all problems. To investigate whether the low problem detection rate was mainly caused by differences in the minor problems reported, Table 4 also shows detection rates and any-two agreements for three subsets of severe problems. The three subsets consisted of the problems reported as critical or serious by at least one of the nine participants, the problems experienced by at least four of the five users, and the problems reported as critical by at least two of the participants. We note, that the subset of problems rated as critical by at least two participants may lead to somewhat inflated detection rates and any-two agreements. The reason is that the problems in this subset were, by definition, reported by at least two participants. The number of users experiencing a problem is a frequently used indicator of problem severity (Hertzum, 2006); it was determined on the basis of the nine participants' reporting of the videos (i.e., users) providing evidence of each problem. For these increasingly smaller subsets of more severe problems, the any-two agreement reached 53% for the most select subset. That is, within this subset, an average of 53% of the problems reported by a pair of participants was reported by both participants in the pair. Only one of the nine problems in this subset was reported by all nine participants, but all nine problems were reported by at least five participants.

Detection rates and any-two agreements were also moderate for the ten participants analysing unmoderated sessions, see Table 5. Participants reported an average of 20% of the positive issues and 32% of all problems. For the unmoderated sessions we defined subsets of severe problems similar to those for the moderated sessions. The any-two agreement reached 69% for the subset of nine problems rated critical by two or more participants. Two of these nine problems were reported by all ten participants, the remaining seven by three to nine participants.

Comparing the data in Tables 4 and 5, we found no difference in detection rate between participants analysing moderated and unmoderated sessions. This was the case for all problems, problems rated critical or serious by at least one participant, problems experienced by at least four users, problems rated critical by at least two participants, and positive issues, Fs(1, 17) = 0.04, 1.31, 0.38, 0.58, 0.15, respectively (all ps > 0.2). However, due to considerable within-group variation, the analyses had low power (0.05-0.19). Thus, we cannot rule out that a difference in detection rate was masked by insufficient sample size. The moderated and unmoderated sessions were also similar with regard to a substantial disagreement among participants in their rating of the problems. For the moderated sessions, 13 (24%) of the 55 problems reported by more than one participant were rated as critical by one participant and minor by another. For the unmoderated sessions it was 14 (30%) of 46 problems. Also, the participants who analysed moderated and unmoderated sessions unanimously agreed on their rating of only 11 (20%) and 18 (39%), respectively, of the problems reported by more than one participant.

Figure 2 shows the average number of usability issues reported as a function of the number of participants. On average, two participants collectively reported 8.2 (moderated) and 8.3 (unmoderated) of the nine problems rated critical by at least two participants. For all problems and for positive issues, the nine (moderated) and ten (unmoderated) participants were not enough for the number of usability issues to stabilize. We fitted the data points in each data series with a logarithmic model and with the Poisson model $n(1 - (1- \lambda)^i)$, where n is the number of issues reported by the participants collectively, λ is the detection rate, and i is the number of participants (Lewis, 1994; Nielsen & Landauer, 1993). The logarithmic models (the dashed lines in Figure 2) explained 98% or more of the variation in the data for all data series, except the two data series about problems rated critical by at least two participants.

The Poisson models (the solid lines in Figure 2) for these two data series fitted very well. The other Poisson models did not fit the data as well (with standard errors of the estimate from 1.41 to 6.30). The resulting Poisson models for problems rated critical by at least two participants were:

- $9(1 - (1- 0.70)^{i})$ for moderated sessions (R^2 = 100%, standard error of estimate = 0.02).
- $9(1 - (1- 0.78)^{i})$ for unmoderated sessions (R^2 = 97%, standard error of estimate = 0.12).

The resulting logarithmic models for all problems were:

- $31.05ln(i) + 29.64$ for moderated sessions (R^2 = 99%, standard error of estimate = 2.36).
- $22.07ln(i) + 20.99$ for unmoderated sessions (R^2 = 100%, standard error of estimate = 1.18).

Figure 2 also illustrates that participants who analysed moderated sessions reported significantly more problems that were rated critical or serious at least once, than participants who analysed unmoderated sessions, $F(1, 17) = 9.33$, $p < 0.01$. We found no difference between participants analysing moderated and unmoderated sessions in number of all problems, problems experienced by at least four users, problems rated critical by at least two participants, and positive issues, Fs(1, 17) = 3.01, 3.02, 0.56, 0.42, respectively (all ps > 0.1). However, these analyses had low power (0.09-0.38) and a difference in detection rate may, therefore, be masked by insufficient sample size. Forty problems were reported by the participants who analysed moderated sessions as well as by the participants who analysed unmoderated sessions, for an overlap of 30%. The overlap increased to 80% for the problems rated critical by at least two participants.

4.2 Group Evaluations

In the group evaluation, the participants in groups of four or five merged the findings they rated as critical, serious, and bugs in their individual evaluations into a group evaluation. The findings rated as minor, ideas, and positive in the individual evaluations were excluded from the group evaluations to make the number of findings manageable to the groups. Table 6 shows the results of our analysis of the 279 findings that were included in the group evaluation.

A group member contributed an average of 36-46% of the problems that resulted from the group's evaluation. That is, the set of findings reported and rated as critical, serious, or bugs in a group member's individual evaluation differed substantially from the set of findings reported and rated as critical, serious, or bugs in the other group members' individual evaluations. This was the case for all four groups, as indicated by their similar detection rates. We found no difference in detection rate between members in the two groups that analysed moderated sessions and those in the two groups that analysed unmoderated sessions, $F(1, 17) = 0.12$, $p = 0.7$, but again low power (0.06) prevented a strong conclusion about the absence of a difference. Any-two agreements in the range 11-39% for the four groups provided further evidence of an evaluator effect. However, because the group evaluation involved only the highly rated findings, it yielded measures of the evaluator effect that reflected both differences in which problems were reported and how the reported problems were rated.

The group evaluation frequently contested the participants' individual rating of a finding as critical, serious, or a bug and led to agreement on a different group rating. Table 7 shows the group ratings of the problems. Twenty (17%) of the issues resulting from the group merging process were not considered to be problems after all, although they had been rated critical, serious, or bugs by at least one individual participant. Another 39 (34%) problems were demoted from critical, serious, or bugs to minor. Thus, the participants' own experience during the group evaluation was that many of the problems they had individually rated as critical or serious should have received a lower rating.

4.3 Comments from Participants

In the plenum discussion after the group evaluation, 13 participants expressed that in their individual evaluations, they had missed critical and serious problems that had been detected by other participants in their group. The remaining five participants felt that their individual evaluations reported all the critical and serious problems discussed in their group. The experience of an evaluator effect by the

majority of the participants was evident in several comments. For example, one participant said: "I came into this [workshop] thinking that there would be more agreement than there was. This setup and particularly the time codes, would make us agree. We didn't." Another participant expressed it like this: "I sense that there are so many judgement calls – if it is a problem, if it is serious… There are hundreds of such calls." A third participant, who also acknowledged an evaluator effect, went on by emphasising that it did not undermine the value of usability evaluation: "There is no such thing as *the* right result. We [as evaluators] are different. There is no final result but we are all providing good service."

Among the reasons for an evaluator effect, several participants mentioned lack of clarity about the goal of the evaluation. It was, for example, unclear how to balance the website's attempts to upsell against the users' dislike of preselected additional items that were probably introduced to increase sales. At least one participant wanted to maintain a pure user perspective, thereby arguing against preselected items; others argued that the goal of an evaluation should be set by the client. Another reason for the evaluator effect was disagreement about the extent to which findings had to arise from evidence in the videos. Some participants reported a problem only if the videos provided evidence for it; others felt free to report an issue they considered problematic whether or not the videos necessarily contained direct evidence. A third reason was uncertainty about the correct answer to tasks. Without local knowledge it was, for example, hard to know the correct answer to Task 7 about finding the U-Haul location nearest to an address in Fremont. One participant mentioned that such uncertainty had made it hard to know when a problem had occurred.

The group evaluation was mentioned as a way of managing the evaluator effect. One participant found that "Challenging each other in a review process is good" and expressed that the group evaluation resulted in more consistency. Five participants mentioned that in their work as usability specialists, they consistently have two evaluators analyse test sessions; most of the other participants considered this practice too expensive.

In their work as usability specialists, only about half of the participants formally rate the severity of usability findings; the others merely distinguish between a few important problems and the rest of the findings. Many participants expressed that the rating of the usability findings had been difficult and that the rating categories (Table 3) were hard to use. A major source of this difficulty was that the rating process involved multiple interrelated aspects, such as the number of users experiencing the problem, whether users were significantly delayed, whether they got frustrated, whether the problem would be easy to fix, and whether a problem could cause major difficulty for real users even if it only caused modest difficulty for a test user.

With respect to the moderation of the test sessions, participants expressed few reservations toward the unmoderated sessions. One participant, leading a large team of moderators, stated that he was impressed by the quality of the unmoderated sessions.

4.4 Assessment of Users' Task Solutions

In addition to the information they were requested to include in their individual evaluations, six participants also included information about whether the users reached the correct solution of the tasks, see Table 8. For several tasks there was disagreement about the correctness of users' task solutions, especially among the three participants who had analysed the moderated sessions. For Task 5 the participants' assessment ranged from zero to all five users solving the task correctly. This illustrates that uncertainty about task solutions may have affected participants' analyses of what problems users experienced.

5 DISCUSSION

In the following we discuss the evaluator effect and its reasons, and we make recommendations for how to manage it in conducting usability tests.

5.1 The Evaluator Effect Persists

The 19 participants in this study reported substantially different sets of usability issues for the evaluated website, even though they were experienced usability professionals and made their evaluation on the basis of a state of the art usability test. An evaluator effect was present (1) for severe problems, all problems, as well as positive issues, (2) for the detection of usability issues as well as the rating of their severity, (3) for moderated as well as unmoderated sessions, and (4) for our analysis of the participants' individual evaluations as well as the participants' own analysis in the group evaluation. We discuss these four aspects in turn.

For the full set of problems the average participant, whether analysing moderated or unmoderated sessions, reported about one third of the problems reported by the participants collectively. Less severe problems contributed more to this evaluator effect than severe problems, as evidenced by the decreasing evaluator effect for smaller subsets of more severe problems. The point at which adding another participant no longer led to the reporting of new usability issues, was approximately two participants, for the subset of problems reported as critical by at least two participants. For the two other subsets of severe problems the return on an additional participant wore off at approximately five participants. And, for all problems and for positive issues, it appeared that the point of diminishing returns had still not been reached with nine or ten participants (Figure 2). With respect to the modelling of our data, we note that Poisson models predict an earlier point of diminishing returns than logarithmic models. The slower increase of the logarithmic models fitted our data better, except for the subset of problems reported as critical by at least two participants. The group evaluation showed that the evaluator effect persists even if evaluators are restricted to report severe problems, due to differences in which problems are reported, as well as how their severity is rated. We found an evaluator effect in usability tests at least as large as in previous studies (Hertzum & Jacobsen, 2003; Hornbæk & Frøkjær, 2008; Vermeeren et al., 2003). Positive usability issues have not been investigated in previous studies of the evaluator effect. For positive issues we found a larger evaluator effect than for problems, reflecting partly that usability tests tend to focus on problems and be less systematic in their coverage of positive issues and partly that this primary focus on problems was also present in our instructions for our participants.

With respect to their severity ratings, the participants agreed about whether an issue was a problem or a positive issue but displayed disagreement in their rating of problems as critical, serious, minor, or bugs. Only 20% (moderated) and 39% (unmoderated) of the problems reported by more than one participant received a unanimous rating and 24% (moderated) and 30% (unmoderated) were rated critical by one evaluator and minor by another. Reasons for such disagreements may include that evaluators often experience difficulty imagining the causes and consequences of problems (Hassenzahl, 2000) and that correlations among the frequency, impact, and persistence of problems are often modest (Hertzum, 2006), suggesting that different severity ratings may result from assigning primary importance to different aspects of severity. During the group evaluations the participants discarded a number of findings and demoted even more to a rating of minor. This is an explicit indication of disagreement about the ratings, and it shows that a rating of critical, serious, or bug from one participant was contested and changed by the other participants about as often as it was accepted as evidence of a critical or serious problem. Law and Hvannberg (2008) also found a near even distribution between changed and retained ratings during problem merging. While their evaluators were students, our study shows a similar level of disagreement about severity ratings for experienced usability professionals. Our participants experienced the rating process as difficult, because it involved integrating multiple, uncertain, and often inconsistent aspects into a single rating. In addition to these difficulties, one may wonder if a running website can contain the number of critical and serious problems reported by the participants. We speculate that the participants tended to rate the largest usability issues on the U-Haul website as critical and thereby construed the categories of critical, serious, and minor relative to the current usability test. This speculation is consistent with a pragmatic use of the rating categories mainly as an indicator of which of the reported problems that are most important to fix. While such problem prioritisation is important, it serves a different purpose to that of severity rating. We contend that there is a need for better procedures to support severity rating.

The evaluator effect appears to be similarly large for moderated and unmoderated usability tests. The detection rates for participants analysing moderated and unmoderated sessions were within 10 percentage points of each other, and not significantly different for all problems, subsets of severe problems, and positive issues. Similarly, the any-two agreements were roughly equal. For both moderated and unmoderated sessions, there were large individual differences in the number of usability issues reported by a single participant. The average number of issues reported by a single participant was, however, similar for moderated and unmoderated sessions, with the exception that participants analysing moderated sessions reported significantly more problems that were rated as critical or serious by at least one participant. Possible reasons for this difference include, that the interaction between the moderator and the user directed the participants toward the more severe problems, or that it led to higher severity ratings by elucidating how affected users were by problems. Unmoderated usability tests appear to be a viable, and cheaper, alternative to moderated tests, as evidenced by the 80% overlap in problems rated as critical by at least two participants and by the participants' generally positive comments about the unmoderated sessions during the plenum discussion.

The participants' merging of their highly rated findings during the group evaluation showed an evaluator effect of a magnitude similar to the evaluator effect calculated on the basis of the authors' merging of the findings reported by the participants. This point is important for two reasons. First, it shows that the participants perceived the presence of an evaluator effect. This is contrary to the study by Hertzum et al. (2002) but supported by comments from the majority of the participants in the plenum discussion. Evaluators who perceive an evaluator effect are, we contend, more likely to take measures against it and, thereby, improve their practice. Second, it is methodologically important because Hornbæk and Frøkjær (2008) argued that the evaluator effect is in part an artefact of the process through which the evaluators' individual findings are merged into usability issues. The problem lists, produced in the four group evaluations, were a result of the participants' own analysis of which findings reported the same usability problems. Hence, each of the four lists provides independent evidence of the evaluator effect.

5.2 Reasons for the Evaluator Effect

This study points to five reasons for the evaluator effect. First, the detection, rating, and reporting of usability issues involve judgement in a situation characterised by uncertainty: How many users should experience difficulty or inconvenience before it constitutes a usability problem? How much difficulty or inconvenience should they experience? How often? In addition to this vagueness in problem criteria, evaluators differ in the values and previous experiences they bring to an evaluation and, therefore, attend differentially to the many aspects of a user's interaction with the system. The Rashomon effect (Heider, 1988; Roth & Mehta, 2002) suggests that in such situations, characterised by judgement and uncertainty, contradictory accounts are to be expected. This view resonates with Law and Hvannberg (2004) as well as with Hertzum and Jacobsen (2003, p. 201) who concluded that "the principal cause for the evaluator effect is that usability evaluation is a cognitive activity, which requires that the evaluators exercise judgment". This reason implies that it is futile to attempt to eliminate the evaluator effect; it must instead be managed.

Second, local or domain knowledge may be required to assess whether certain parts of a user's interaction with a system are appropriate and lead to correct conclusions. Evaluators are likely to lack such knowledge. While Chilana et al. (2010) stressed the need for evaluators to partner with domain experts when evaluating systems in complex domains, this study shows that local and domain knowledge may also be necessary in evaluations of e-commerce systems with a broad, unspecialised user group. In the absence of such knowledge, evaluators face increased uncertainty or remain unaware of issues relevant to their evaluation, leading to more judgement calls and more diversity in the usability issues reported. This reason for the evaluator effect was explicitly raised by the participants in the plenum discussion and it is implicit in the disagreements among participants about whether users solved the tasks correctly (Table 8).

Third, some evaluators may deliberately report only a subset of their usability findings. The rationale for this strategy is to maximise the usability of the test report by not flooding the client with a very

long list of usability issues but instead restricting the report to a manageable subset consisting of the more important issues. This strategy appears particularly suited for development processes that include multiple iterations of evaluation and redesign. If some evaluators apply this strategy while others report all their findings, the agreement between reports will suffer. Assuming that a manageable subset can comprise at most 30-50 usability issues and that the total number of usability issues detected by the evaluators is larger (e.g., the participants in this study collectively reported 182 usability issues) then the manageable-subset strategy alone may lead to a substantial evaluator effect. We have little evidence of the extent to which our participants applied this strategy, but it points to the possibility that incomplete reports may have desirable properties that should be remembered in discussions of the evaluator effect.

Fourth, the goal of an evaluation may be unclear or contested, thereby leading evaluators to assess the system against dissimilar images of what the system should ideally accomplish. The goal is partly something that should be specified for the individual evaluation, and partly something that relates to how evaluators understand their role. Some evaluators see themselves as advocates for the user (Boivie, Gulliksen, & Göransson, 2006) and are, thus, likely to report usability problems when a system does not serve the user's best interest but, for example, the client's economic interest in upselling. Other evaluators leave it to the client to define the goal of the evaluation and assess the system against that goal. Whenever there is a tension between the interests of different groups, as is frequently the case for e-commerce systems, lack of clarity about the goal of the evaluation may lead to an evaluator effect.

Fifth, evaluators may consider an issue problematic, yet have no evidence for it in the test sessions. Some evaluators report such problems, presumably contending that although none of the test users experienced the problem, other users will. Other evaluators refrain from reporting a problem unless they find evidence for it in the test sessions, presumably contending that in the absence of evidence, their intuition about the existence of a problem is questionable. The participants in this study consisted of both types of evaluators and this also contributed to the evaluator effect.

5.3 Recommendations for Usability Evaluation

This study has several implications for practitioners who conduct usability evaluations. We make five recommendations:

First, have more than one evaluator independently analyse test sessions, at least in important evaluations. With more than one evaluator, more problems are detected and evaluators get an opportunity to reflect on their agreements and disagreements. This recommendation echoes Hertzum and Jacobsen (2003) and appears to be entering into practice. In a survey of 155 usability practitioners, 23% of them report that their test sessions are independently analysed by at least two evaluators (Følstad, Law, & Hornbæk, 2012).

Second, consult people with local or domain knowledge to avoid uncertainty in the analysis of user actions. Local and domain knowledge may be needed to interpret whether users approach tasks appropriately, miss important information, and reach correct task solutions. The goal of a test should be clarified ahead of the test.

Third, consolidate the severity ratings of the reported usability issues in a group process. Such a process is likely to reduce the number of highly rated problems and thereby adds focus to redesign work. A group process may also support problem prioritisation, by providing the usability specialists with development people who are knowledgeable about the ease or difficulty of fixing the problems.

Fourth, consider the use of unmoderated tests. On the basis of this study, unmoderated tests appear to be a cost-effective alternative or supplement to moderated tests as the evaluator effect and the number of identified usability issues were similar for moderated and unmoderated tests.

Fifth, remember that perfect reliability is not required in order for usability testing to be worthwhile. This final recommendation is particularly relevant when multiple usability tests are conducted in an iterative process of evaluation and redesign, thereby providing additional possibilities for finding usability problems that are missed initially.

5.4 Limitations

Three limitations should be remembered in interpreting the results of this study. First, the matching process through which we identified the usability findings that reported the same usability issue was an interpretive process involving judgement. While we reached consensus about the matching of all 836 reported findings, we acknowledge that others may group the findings somewhat differently and that this may affect the magnitude of the resulting evaluator effect. The principle underlying our matching process was that two findings were instances of the same usability issue if, and only if, the same website revision would remedy both findings.

Second, to be able to assess the evaluator effect for subsets of severe problems, rather than just for the full set of all problems, we used the participants' rating of the problems to define such subsets. However, the participants expressed that the problems had been difficult to rate and they often differed in their rating of the same problems. We have partly overcome this uncertainty by including the subset for which two or more participants agreed on a rating of critical. While the detection rates and any-two agreements for the subsets of severe problems were defined on the basis of the best available evidence, we acknowledge that they are based on uncertain data about how severe the problems were.

Third, we decided against providing the participants with solutions for the seven test tasks, because we considered it part of their analysis of the test sessions to determine if users reached correct answers to the tasks. In retrospect we probably should have provided solutions along with the tasks, because it may be argued that solutions are normally devised in collaboration with the client as a part of the process of defining the tasks. The merit of not providing task solutions was that it made it evident that, in their absence, participants came to different conclusions about whether users solved the tasks correctly. Task solutions contribute to reducing the number of judgement calls made by evaluators, and may thereby help to manage the evaluator effect.

6 CONCLUSION

In this study 19 usability specialists analysed video recordings of either moderated or unmoderated test sessions from a usability test of an e-commerce website. The participants reported substantially different sets of usability issues, even though they were experienced, the test was state of the art, and the reporting procedure was consistent across participants. This evaluator effect was present for the full set of all reported problems, for subsets of severe problems, for positive issues, and for the detection of usability issues, as well as the rating of their severity. The agreement among evaluators did however increase for smaller subsets of more severe problems. Within the subset of problems reported as critical or serious by at least one participant, an average of 41% (moderated) and 45% (unmoderated) of the problems reported by a pair of participants was reported by both participants in the pair. Across all the sets of usability issues, the evaluator effect was similar for participants analysing moderated and unmoderated sessions. Finally, the evaluator effect was also present for the participants' own analysis in the group evaluation and acknowledged by the majority of the participants in the plenum discussion, suggesting that evaluators may be prepared to take steps to manage the evaluator effect.

We recommend having more than one evaluator independently analyse test sessions, at least in important evaluations. We also recommend consulting people with local and domain knowledge, consolidating severity ratings in a group process, and considering unmoderated tests.

ACKNOWLEDGEMENTS

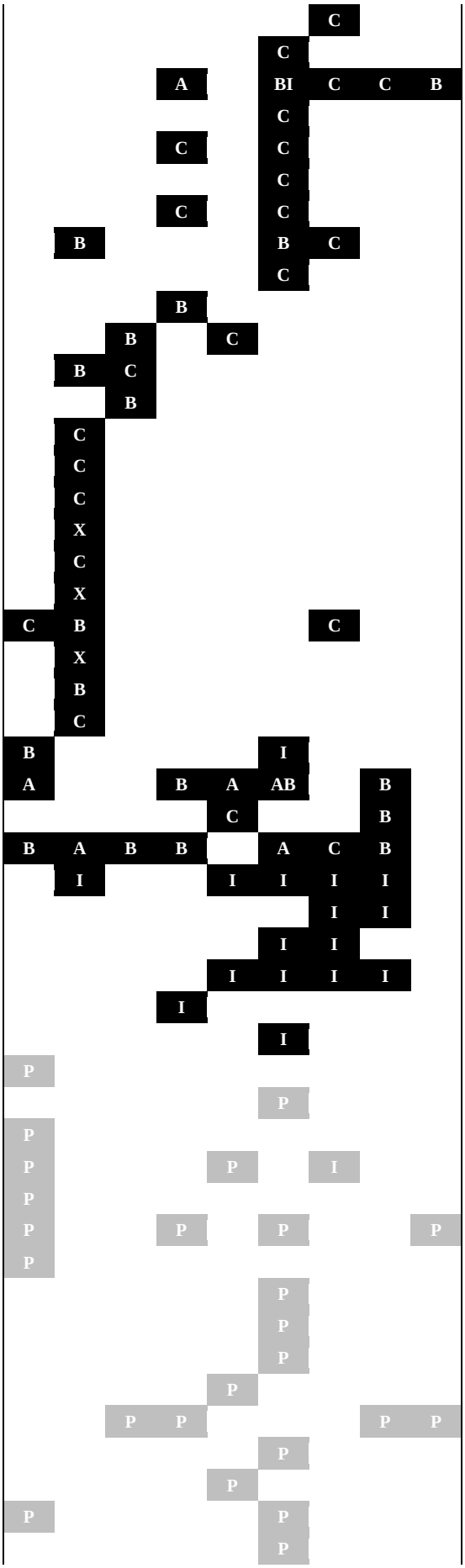
We wish to thank U-Haul for agreeing to the use of their website as the object of evaluation in this study and for their support in our design of the evaluation. We are grateful to Kate McCaffrey for conducting the moderated sessions and to Tom Tullis and Fidelity Investments, for generously allocating the resources required to run the moderated sessions. Special thanks are due to the usability specialists who participated in the study as evaluators. The 19 evaluators were Carol Barnum (Usability Center at Southern Polytechnic State University, USA), Avram Baskin (Kenexa

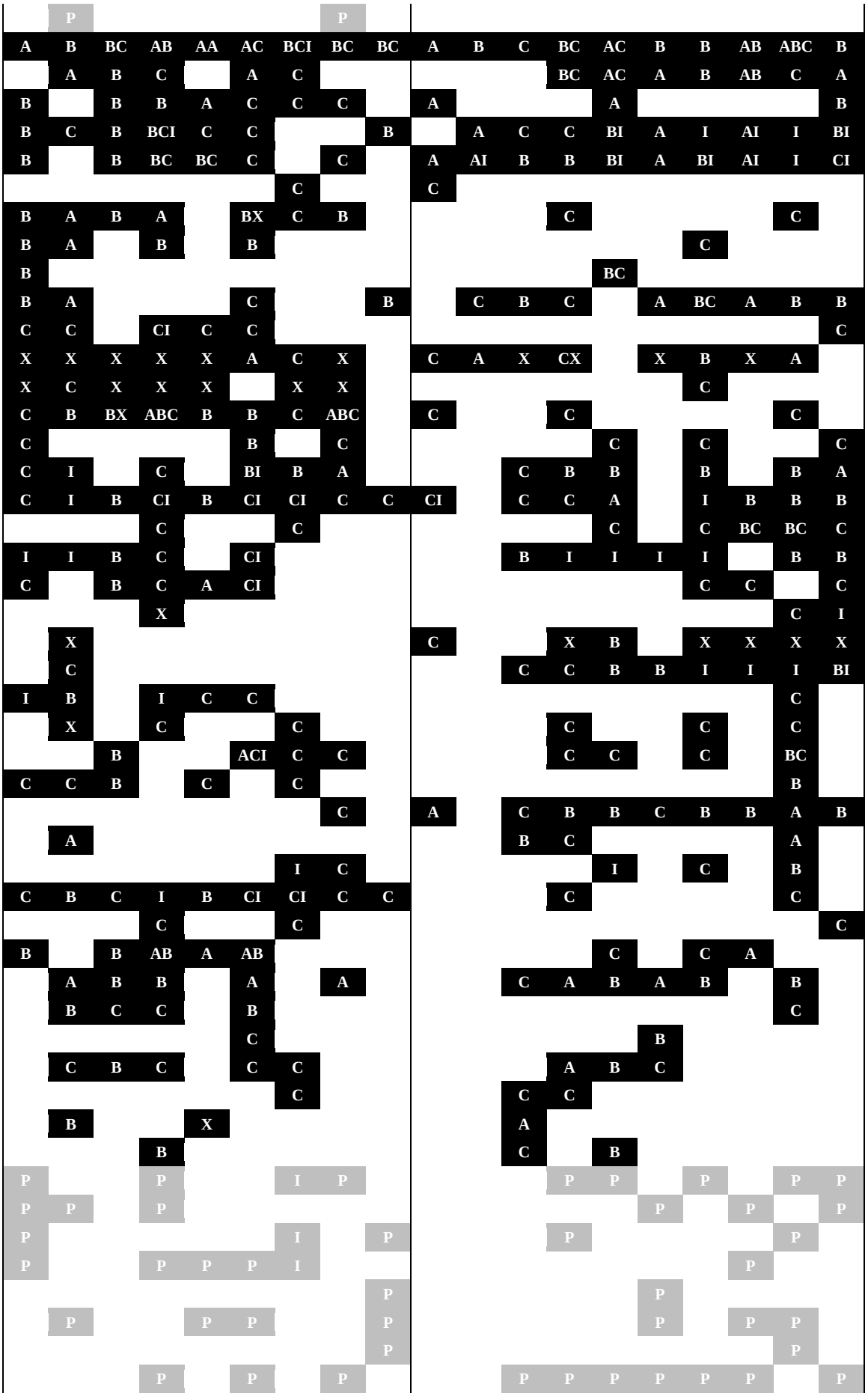
Technology, Inc., USA), Nigel Bevan (independent consultant, UK), David Bozovich (PayPal, working for U-Haul when the study was conducted, USA), Barbara Chaparro (Wichita State University, USA), Stephen Denning (UserVision, Scotland), Linna Manomaitis Ferguson (Booz Allen Hamilton, USA), Erin Friess (University of North Texas, USA), Elisa del Galdo (del Galdo Consulting Ltd., UK), Veronica Hinkle (State Farm Insurance, USA), Niels Ebbe Jacobsen (Danish Consumer Council, Denmark), Steve Krug (Advanced Common Sense, USA), Kate McCaffrey (Fidelity Investments, USA), Jen McGinn (Oracle, USA), Rolf Molich (DialogDesign, Denmark), Donna Tedesco (Fidelity Investments, USA), Brian Traynor (Mount Royal University, Canada), Tom Tullis (Fidelity Investments, USA), and Chauncey Wilson (Autodesk, USA).

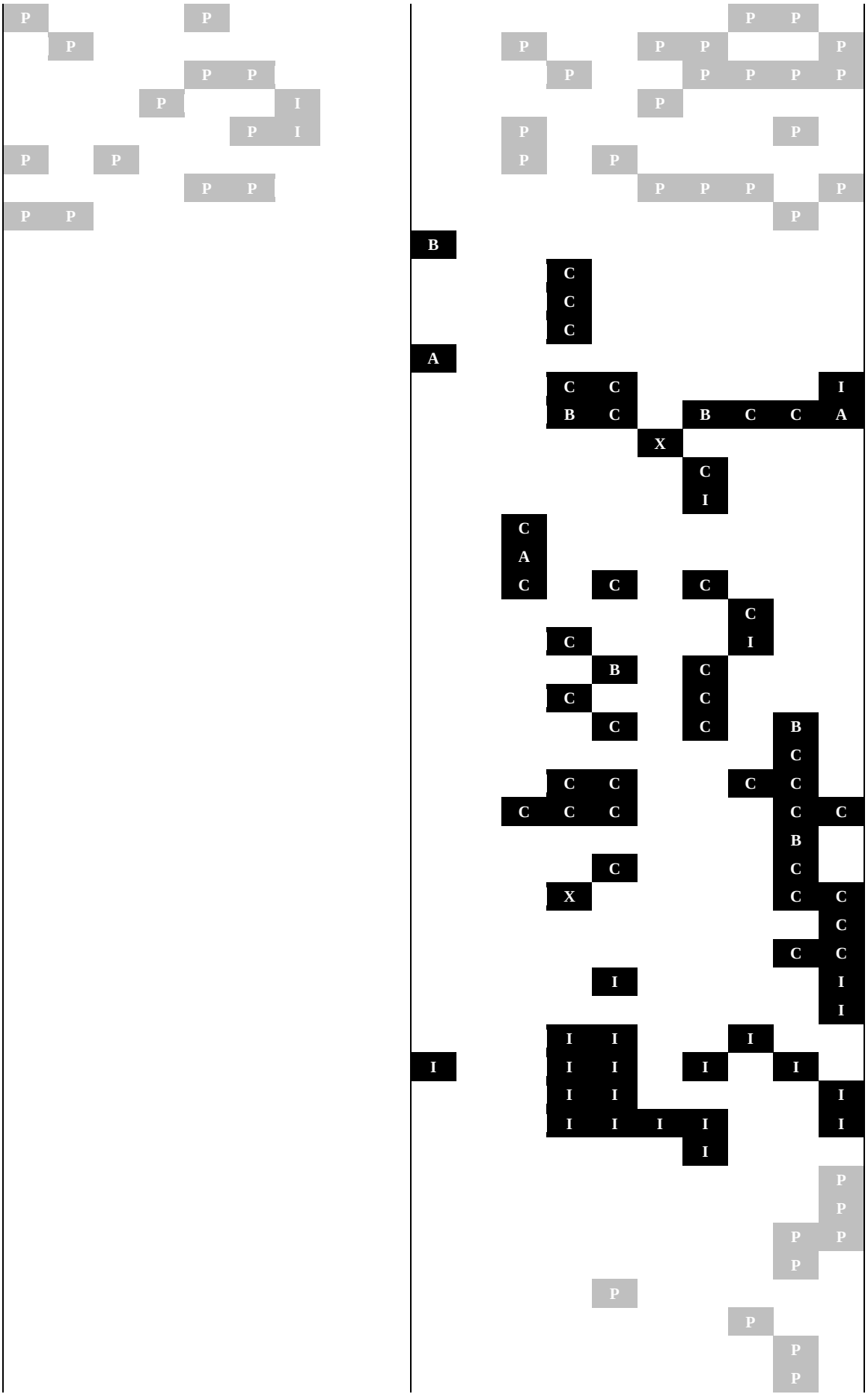
APPENDIX

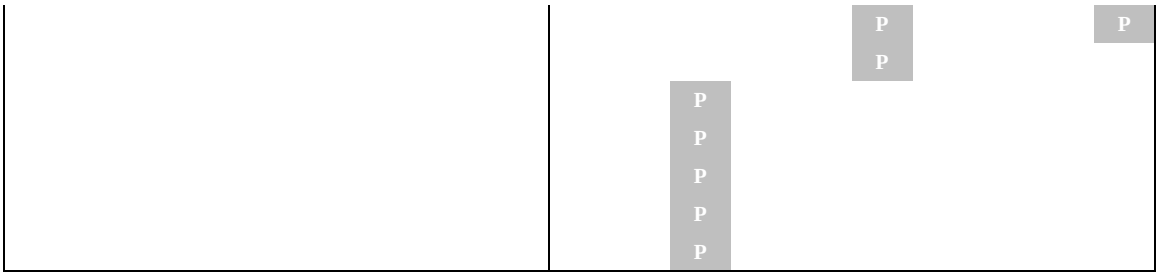
This appendix shows the usability issues reported by each participant. Each column in the matrix represents a participant and each row represents a usability issue. If a participant reported an issue, the corresponding cell in the matrix is coloured black for usability problems and grey for positive issues. The cell also gives the participant's severity rating of the issue (A – critical, B – serious, C – minor, X – bug, I – idea, P – positive issue). In some cases the same participant reported multiple usability findings with different severity ratings but these findings were merged into the same usability issue. For these cases the cell gives all the different severity ratings.

Participants analysing moderated sessions									Participants analysing unmoderated sessions									
A	B	D	H	N	O	S	W	Y	C	E	F	G	L	M	P	R	U	V
C																		
C	B		C	I	C		C											
C	I	C	C		CI		B											
C																		
C																		
C																		
C																		
I	C	C	C		CI	CI												
	A	B	ABC		C	C	CI											
	X					C	C											
C	B		C	B	C	C	C											
	I		B	C			I											
						BC												
						C												
						C												
	C					C												
						C												
					C	C												
						C												
						C												
						C												
						C												
						C												
					C													
						C												
					C	C												
						C												
						C												


C
C
A
BI C C B
C
C
C
C
C
C
B
B C
C
B
B
C
B C
B
C
C
C
X
C
X
C B
C
X
B
C
B
I
A
B A AB
B
C
B
B A B B
A C B
I
I I I I
I I
I I
I I I I
I
I
P
P
P
P
P
P
P
P
P
P
P
P
P
P
P
P
P
P
P
P
P
P
P
P
P







REFERENCES

Bailey, G. (1993). Iterative methodology and designer training in human-computer interface design. In *Proceedings of the INTERCHI'93 Conference on Human Factors in Computing Systems* (pp. 198-205). New York: ACM Press.

Bailey, R. W., Allan, R. W., & Raiello, P. (1992). Usability testing vs. heuristic evaluation: A head-to-head comparison. In *Proceedings of the Human Factors Society 36th Annual Meeting* (pp. 409-413). Santa Monica, CA: HFS.

Boivie, I., Gulliksen, J., & Göransson, B. (2006). The lonesome cowboy: A study of the usability designer role in systems development. *Interacting with Computers, 18*(4), 601-634.

Bruun, A., Gull, P., Hofmeister, L., & Stage, J. (2009). Let your users do the testing: A comparison of three remote asynchronous usability testing methods. In *Proceedings of the CHI 2009 Conference on Human Factors in Computing Systems* (pp. 1619-1628). New York: ACM Press.

Chilana, P. K., Wobbrock, J. O., & Ko, A. J. (2010). Understanding usability practices in complex domains. In *Proceedings of the CHI 2010 Conference on Human Factors in Computing Systems* (pp. 2337-2346). New York: ACM Press.

Clemmensen, T., Hertzum, M., Hornbæk, K., Shi, Q., & Yammiyavar, P. (2009). Cultural cognition in usability evaluation. *Interacting with Computers, 21*(3), 212-220.

Dumas, J., & Loring, B. (2008). *Moderating usability tests: Principles & practices for interacting*. Burlington, MA: Morgan Kaufmann.

Dumas, J. S., & Redish, J. C. (1999). *A practical guide to usability testing. Revised edition*. Exeter, UK: Intellect Books.

Frøkjær, E., & Hornbæk, K. (2008). Metaphors of human thinking for usability inspection and design. *ACM Transactions on Computer-Human Interaction, 14*(4), 20:01-20:33.

Følstad, A., Law, E. L.-C., & Hornbæk, K. (2012). Analysis in practical usability evaluation: A survey study. In *Proceedings of the CHI 2012 Conference on Human Factors in Computing Systems* (pp. 2127-2136). New York: ACM Press.

Gorlenko, L., & Englefied, P. (2006). Usability error classification: Qualitative data analysis for UX practitioners. In *CHI '06 Extended Abstracts on Human Factors in Computing Systems* (pp. 803-808). New York: ACM Press.

Gulliksen, J., Boivie, I., Persson, J., Hektor, A., & Herulf, L. (2004). Making a difference - A survey of the usability profession in Sweden. In *Proceedings of the NordiCHI 2004 Conference on Human-Computer Interaction* (pp. 207-215). New York: ACM Press.

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

Heider, K. G. (1988). The Rashomon effect: When ethnographers disagree. *American Anthropologist, 90*(1), 73-81.

Hertzum, M. (2006). Problem prioritization in usability evaluation: From severity assessments toward impact on design. *International Journal of Human-Computer Interaction, 21*(2), 125-146.

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 *CHI 2002 Extended Abstracts on Human Factors in Computing Systems* (pp. 662-663). New York: ACM Press.

Hornbæk, K., & Frøkjær, E. (2008). A study of the evaluator effect in usability testing. *Human-Computer Interaction, 23*(3), 251-277.

Jacobsen, N. E., Hertzum, M., & John, B. E. (1998a). The evaluator effect in usability studies: Problem detection and severity judgments. In *Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting* (pp. 1336-1340). Santa Monica, CA: HFES.

Jacobsen, N. E., Hertzum, M., & John, B. E. (1998b). The evaluator effect in usability tests. In *CHI'98 Conference Summary on Human Factors in Computing Systems* (pp. 255-256). New York: ACM Press.

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

Kessner, M., Wood, J., Dillon, R. F., & West, R. L. (2001). On the reliability of usability testing. In *CHI '01 Extended Abstracts on Human Factors in Computing Systems* (pp. 97-98). New York: ACM Press.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174.

Law, E. L.-C., & Hvannberg, E. T. (2004). Analysis of combinatorial user effect in international usability tests. In *Proceedings of the CHI 2004 Conference on Human Factors in Computing Systems* (pp. 9-16). New York: ACM Press.

Law, E. L.-C., & Hvannberg, E. T. (2008). Consolidating usability problems with novice evaluators. In *Proceedings of the NordiCHI 2008 Conference on Human-Computer Interaction* (pp. 495-498). New York: ACM Press.

Lazar, J., Feng, J. H., & Hochheiser, H. (2010). *Research methods in human-computer interaction*. Chichester, UK: Wiley.

Lewis, C. (1982). *Using the "thinking-aloud" method in cognitive interface design, RC 9265 (#40713)*. Yorktown Heights, NY: IBM Thomas Watson Research Center.

Lewis, J. R. (1994). Sample sizes for usability studies: Additional considerations. *Human Factors, 36*(2), 368-378.

Lewis, J. R. (2001). Introduction: Current issues in usability evaluation. *International Journal of Human-Computer Interaction, 13*(4), 343-349.

McFadden, E., Hager, D. R., Elie, C. J., & Blackwell, J. M. (2002). Remote usability evaluation: Overview and case studies. *International Journal of Human-Computer Interaction, 14*(3&4), 489-502.

Molich, R., Bevan, N., Curson, I., Butler, S., Kindlund, E., Miller, D., & Kirakowski, J. (1998). Comparative evaluation of usability tests. In *UPA'98: Proceedings of the Usability Professionals Association 1998 Conference* (pp. 189-200). Chicago, IL: UPA.

Molich, R., & Dumas, J. S. (2008). Comparative usability evaluation (CUE-4). *Behaviour & Information Technology, 27*(3), 263-281.

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

Nelson, E. T., & Stavrou, A. (2011). Advantages and disadvantages of remote asynchronous usability testing using Amazon mechanical turk. In *Proceedings of the Human Factors and Ergonomics Society 55th Annual Meeting* (pp. 1080-1084). Santa Monica, CA: HFES Press.

Nielsen, J. (1993). *Usability engineering*. Boston, MA: Academic Press.

Nielsen, J., & Landauer, T. K. (1993). A mathematical model of the finding of usability problems. In *Proceedings of the INTERCHI'93 Conference on Human Factors in Computing Systems* (pp. 206-213). New York: ACM Press.

Nielsen, J., & Molich, R. (1990). Heuristic evaluation of user interfaces. In *Proceedings of the CHI´90 Conference on Human Factors in Computing Systems* (pp. 249-256). New York: ACM Press.

O'Malley, C. E., Draper, S. W., & Riley, M. S. (1984). Constructive interaction: A method for studying human-computer interaction. In B. Shackel (Ed.), *Proceedings of the INTERACT'84 Conference on Human-Computer Interaction* (pp. 269-274). Amsterdam: Elsevier.

Roth, W. D., & Mehta, J. D. (2002). The Rashomon effect: Combining positivist and interpretivist approaches in the analysis of contested events. *Sociological Methods & Research, 31*(2), 131-173.

Rubin, J., & Chisnell, D. (2008). *Handbook of usability testing: How to plan, design, and conduct effective tests* (Second ed.). Indianapolis, IN: Wiley.

Siegel, D. A., & Dray, S. M. (2011). A professional empiricist manifesto. *ACM Interactions, 18*(4), 82-87.

Vermeeren, A. P. O. S., van Kesteren, I. E. H., & Bekker, M. M. (2003). Managing the evaluator effect in user testing. In M. Rauterberg, M. Menozzi & J. Wesson (Eds.), *Proceedings of the INTERACT '03 Conference on Human-Computer Interaction* (pp. 647-654). Amsterdam: IOS Press.

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: Wiley.

Table 1. Participants' experience as usability evaluators

	Category	Number of participants
Years having done usability evaluations	1-9	6
	10-19	4
	20-29	7
	30-39	2
Number of usability evaluations conducted	5-20	4
	30-50	4
	100-200	7
	300-	4

Table 2. Test tasks for the U-Haul moving truck rental website.

#	Task description
1	The couple needs a truck that is suitable for all the furniture and belongings in their 3 room apartment. Please find the total price the couple will have to pay for the truck. Note: They are moving on April 14th from Darlington Rd. in Pittsburgh, PA 15217 to Emerson St. in Denver, CO 80218.
2	Before you go any further, you want to check if Mike and Anna need a special driver's license to drive the truck across country. Where would you find that info?
3	They also need an indoor storage unit in Pittsburgh that can hold 10 moving boxes (18" x 18" x 16") and a large fridge. Find the per month cost of the storage.
4	You have a few questions that the U-Haul website hasn't answered. Please find the phone number for the U-Haul pickup location closest to the couple's home on Darlington Rd. in Pittsburgh, PA.
5	The couple has decided to rent the truck. Please book the truck you found the pricing for earlier. In addition, please order 20 large moving boxes, 15 small moving boxes, a utility dolly, and a dozen moving blankets. Note: Please stop when you reach the "Billing Info" page. Do *not* submit the order.
6	During the move, an unknown person scratched the truck in several places, probably with a knife. An auto body technician has estimated that the repair will cost $2,000. Since you helped the couple book the truck, they called to find out if they are liable for repair costs. And if so, how much will it cost?
7	You were impressed with U-Haul during your friends' move and you are considering U-Haul yourself. Find the nearest U-Haul pick-up/drop off to your home. Note: You live at 48105 Warm Springs Blvd., Fremont, CA 94539.

Table 3. Categories for rating usability findings.

Rating	Description
Critical problem	Causes frequent catastrophes. A catastrophe is a situation where the website "wins" over the test participant – that is, a situation where the test participant cannot solve a reasonable task or where the website annoys the test participant considerably.
Serious problem	Delays test participants in their use of the website for some minutes, but eventually allows them to continue. Causes occasional "catastrophes".
Minor problem	Causes test participants to hesitate for some seconds.
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.
Idea	A suggestion from a test participant that could lead to a significant improvement of the user experience.
Positive issue	This approach is recommendable and should be preserved.

Note: the categories are based on severity categories proposed by Nielsen (1993) and Dumas and Redish (1999).

Table 4. Individual evaluations of moderated test sessions, *N* = 9 participants

	Usability issues	Detection rate %	Any-two agreement %
All problems	101	33	31
Severe problems			
• Rated critical or serious by at least one participant	45	50	41
• Experienced by at least four users	24	53	47
• Rated critical by at least two participants	9	70	53
Positive issues	33	22	12

Table 5. Individual evaluations of unmoderated test sessions, *N* = 10 participants

	Usability issues	Detection rate %	Any-two agreement %
All problems	73	32	30
Severe problems			
• Rated critical or serious by at least one participant	32	40	45
• Experienced by at least four users	20	48	49
• Rated critical by at least two participants	9	78	69
Positive issues	31	20	20

Table 6. Group evaluation of the findings rated as critical, serious, and bugs in the participants' individual evaluations

Group	Participants	Findings from individual reports	Problems resulting from group merging process	Detection rate %	Any-two agreement %
Moderated 1	5	88	32	46	39
Moderated 2	4	80	38	36	11
Unmoderated 1	5 [a]	58	25	42	26
Unmoderated 2	5	53	20	45	31

Note: [a] The group merged five individual evaluations but only four of the five participants took part in the group evaluation.

Table 7. Group severity rating of the problems rated as critical, serious, and bugs in the participants' individual evaluations

Group	Total problems	Group severity rating			
		Not a problem	Minor	Serious	Critical
Moderated 1	32	3	14	11	4
Moderated 2	38	10	10	14	4
Unmoderated 1	25	2	10	10	3
Unmoderated 2	20	5	5	4	6

Table 8. Number of users judged by participants to be solving the tasks correctly

	Task 1	Task 2	Task 3	Task 4	Task 5	Task 6	Task 7
Moderated							
Participant B	1	5	3	2	1	5	3
Participant N	4	5	4	4	0	3	5
Participant S	5	5	4	4	5	2	5
Unmoderated							
Participant C	3	4	4	5	4	5	4
Participant P	5	5	5	5	5	5	3
Participant U	2	5	4	5	4	4	3

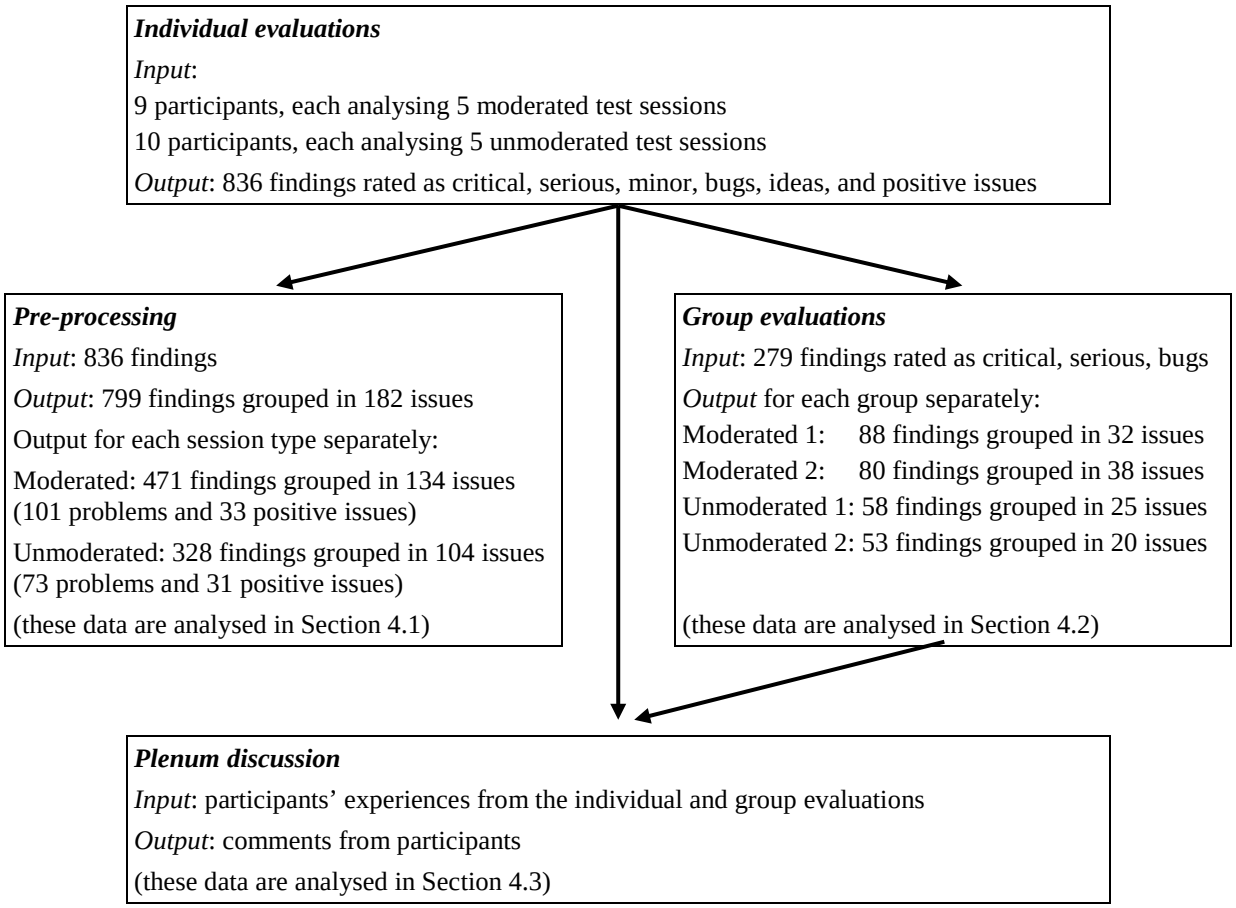


Figure 1. Overview of study

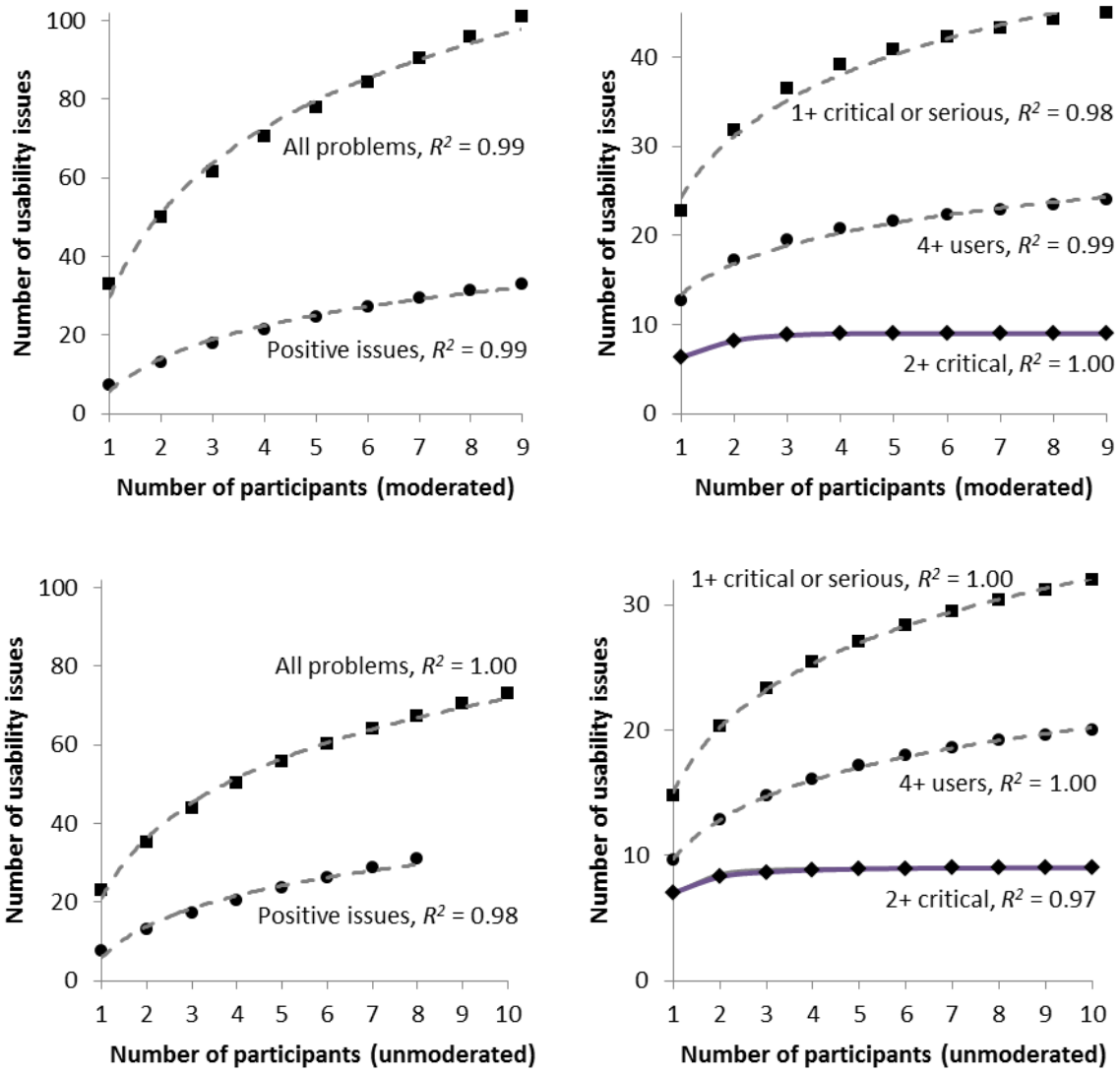


Figure 2. The average number of usability issues reported by different numbers of participants analysing moderated (top row of graphs) and unmoderated (bottom row of graphs) sessions. The graphs on the left show all problems and positive issues. The graphs on the right show three subsets of the problems: those rated critical or serious by at least one participant (1+ critical or serious), those experienced by at least four users (4+ users), and those rated critical by at least two participants (2+ critical). Solid lines show Poisson models of the data, dashed lines show logarithmic models.