

This article was downloaded by: [Det Kgl Bibl Nationalbibl og Kbh Univ]

On: 15 September 2008

Access details: Access Details: [subscription number 776111357]

Publisher Taylor & Francis

Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK



Human-Computer Interaction

Publication details, including instructions for authors and subscription information:

<http://www.informaworld.com/smpp/title-content=t775653648>

A Study of the Evaluator Effect in Usability Testing

Kasper Hornbæk ^a; Erik Frøkjær ^a

^a University of Copenhagen,

Online Publication Date: 01 July 2008

To cite this Article Hornbæk, Kasper and Frøkjær, Erik(2008)'A Study of the Evaluator Effect in Usability Testing',Human-Computer Interaction,23:3,251 — 277

To link to this Article: DOI: 10.1080/07370020802278205

URL: <http://dx.doi.org/10.1080/07370020802278205>

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: <http://www.informaworld.com/terms-and-conditions-of-access.pdf>

This article may be used for research, teaching and private study purposes. Any substantial or systematic reproduction, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.



A Study of the Evaluator Effect in Usability Testing

Kasper Hornbæk and **Erik Frøkjær**
University of Copenhagen

ABSTRACT

The evaluator effect names the observation that usability evaluators in similar conditions identify substantially different sets of usability problems. Yet little is known about the factors involved in the evaluator effect. We present a study of 50 novice evaluators' usability tests and subsequent comparisons, in teams and individually, of the resulting usability problems. The same problems were analyzed independently by 10 human-computer interaction experts. The study shows an agreement between evaluators of about 40%, indicating a substantial evaluator effect. Team matching of problems following the individual matching appears to improve the agreement, and evaluators express greater satisfaction with the teams' matchings. The matchings of individuals, teams, and independent experts show evaluator effects of similar sizes; yet individuals, teams, and independent experts fundamentally disagree about which problems are similar. Previous claims in the literature about the evaluator effect are challenged by the large variability in the matching of usability problems; we identify matching as a key determinant of the evaluator effect. An alternative view of usability problems and evaluator agreement is proposed in which matching is seen as an activity that helps to

Kasper Hornbæk is a computer scientist with an interest in usability research and information visualization; he is an Associate Professor in the Department of Computer Science, University of Copenhagen. **Erik Frøkjær** is a computer scientist and Associate Professor in the Department of Computer Science, University of Copenhagen; he is interested in human-computer interaction, systems design, usability research, and descriptive psychology.

CONTENTS

- 1. INTRODUCTION**
 - 2. RELATED WORK**
 - 3. EMPIRICAL STUDY**
 - 3.1. Evaluators
 - 3.2. Application
 - 3.3. Individual Evaluation (Week 1)
 - 3.4. Individual Matching of Problems (Week 2)
 - 3.5. Team Matching of Problems (Week 3)
 - 3.6. Independent-Expert Matching of Problems
 - 4. ANALYSIS**
 - 5. RESULTS**
 - 5.1. Individual Evaluations (Week 1)
 - 5.2. Individual Matching of Problems (Week 2)
 - 5.3. Team Matching of Problems (Week 3)
 - 5.4. Independent-Expert Matching
 - 5.5. Comments on the Individual and Team Matching
 - 6. DISCUSSION**
 - 7. CONCLUSION**
-

make sense of usability problems and where the existence of a correct matching is not assumed.

1. INTRODUCTION

The phrase evaluator effect names the observation that usability evaluators report substantially different sets of usability problems when applying the same evaluation technique on the same application. We investigate possible factors in the evaluator effect and discuss implications for usability research and practical evaluation.

The evaluator effect has been found by different researchers across a variety of evaluation techniques and applications (Hertzum & Jacobsen, 2001; Hertzum, Jacobsen, & Molich, 2002; Jacobsen, Hertzum, & John, 1998a, 1998b; Vermeeren, van Kesteren, & Bekker, 2003). A review by Hertzum and Jacobsen (2001) found that the average agreement between two usability evaluators of the same system using the same technique varied between 5% and 65%. These numbers seriously challenge the assumption that usability evaluation is reliable, because it must be expected that any evaluation repeated with other evaluators identifies very different problems. Indirectly, the evaluator effect also challenges the validity of usability evaluation: If we

are really finding usability problems, how can evaluators disagree this much?

In addition to acknowledging the evaluator effect, some studies have aimed to characterize the factors involved in it (Hertzum & Jacobsen, 2001; Hertzum et al., 2002; Vermeeren et al., 2003). For example, it has been argued that vague evaluation procedures may make evaluators focus on different things during the evaluation (Hertzum & Jacobsen, 2001) and that evaluators occasionally fail to observe the evidence of a particular problem (Vermeeren et al., 2003).

Extending that work, we empirically investigate the role of matching in the evaluator effect. Matching refers to the procedure used for comparing usability problems found by different evaluators to assess whether they concern the same or different problems. Thus, matching underlies any study of the evaluator effect. In particular we look at two factors in the evaluator effect that previous studies leave unexamined. We compare the evaluator effect for a series of evaluations as assessed by the evaluators themselves and by independent-expert matching. No previous study has systematically investigated evaluators' own assessment of the extent of the evaluator effect, though their views on agreement may differ from those of independent experts. We also compare matching in teams of evaluators to individual matching. Team matching allows for discussion and clarification of usability problems, whereas variants of individual matching are used in most previous studies of the evaluator effect. We study these factors by assessing the quantitative differences of the evaluator effect, by analyzing the content of problems matched, and by analyzing participants' comments about perceived difficulties and feelings of agreement.

Describing the factors in the evaluator effect appears a useful first step in understanding how to deal with it and is a necessary prerequisite for improving usability evaluation methods so that they are less affected by the evaluator effect. So although our focus is on somewhat specific factors, we aim at raising issues of general interest to usability researchers (e.g., of matching of usability problems, a notoriously difficult problem; Lavery, Cockton, & Atkinson, 1997) and to point out how to cope with the evaluator effect in practical usability work (e.g., of doing group consolidation of usability problems).

2. RELATED WORK

The evaluator effect was named by Jacobsen et al. (1998a, 1998b). They had four participants look through four video tapes of users thinking aloud while using a multimedia authoring system. Each participant noted problems, as defined by nine criteria, such as "the user explicitly gives up" or "the user expresses surprise" (Jacobsen et al., 1998a, p. 1137). Two people then

matched the problems to form a list of unique usability problems. This list showed that only 20% of the problems were found by all evaluators and that around half of the problems were only found by one evaluator. The authors concluded that “analyzing usability test results is an activity with considerable individual variability” and that “the evaluator effect revealed in this study shows that usability tests are less reliable than previously reported” (Jacobsen et al., 1998a, p. 1339).

After the appearance of Jacobsen et al.’s work, it is clear that previously published data, though not using the phrase evaluator effect, support the conclusion that a low overlap exists between the problems found by different evaluators (e.g., Nielsen, 1992; Nielsen & Molich, 1990). Nielsen and Molich, for example, found that the problems identified by individual evaluators covered between 13% and 59% of a known set of problems. This implies a substantial evaluator effect. The evaluator effect also appears present in industrial usability evaluations. In Molich, Ede, Kaasgaard, and Karyukin’s (2004) study of usability evaluation by nine professional teams, 75% of the problems identified were found by only one team.

The results of Jacobsen et al. have been corroborated in a number of later studies (Hertzum & Jacobsen, 1999; Hertzum et al., 2002; Kessner, Wood, Dillon, & West, 2001; Vermeeren et al., 2003), using procedures for assessing the evaluator effect similar to those used in the studies by Jacobsen et al. For instance, Vermeeren et al. examined evaluators looking at videos of users’ interaction with a drawing program and a video game. Evaluators analyzed the videos using a tool for structured analysis and marking up of video segments. An evaluator effect was found for both evaluators’ detection of breakdowns and for their identification of usability problems. A review by Hertzum and Jacobsen (2001) found the evaluator effect in 11 studies of usability evaluation (average agreement between evaluators varied between 5% and 65%). In these studies evaluators had differing experience with conducting usability evaluation, used both analytical and empirical evaluation techniques, and evaluated both simple and complex systems. Moreover, the evaluator effect is seen both for detection of usability problems and for evaluators’ assessment of the severity of the problems detected. The evaluator effect has also been found both when evaluators observe the same evaluation (captured on video) and when the results of different evaluations are matched. Thus, the evaluator effect appears substantial and widely present.

A number of reasons for the evaluator effect have been put forward. Hertzum and Jacobsen (2001) suggested that the application of usability evaluation techniques is characterized by variability in the task scenarios considered by evaluators and by vagueness of evaluation procedures and of the criteria for what constitutes a usability problem. Vagueness of evaluation procedure, for example, means that the description of the evaluation method

does not make explicit what are the critical steps of the evaluation. In cognitive walkthrough, for example, it is suggested that critical knowledge about the user is not made explicit, meaning that evaluators may mistake their own problems for problems that the intended user group may experience (Hertzum & Jacobsen, 2001). Vermeeren et al. (2003) gave even more detailed descriptions of the reasons why evaluators report different problems, including differences in interpreting users' intention and mishearing utterances. These reasons all relate to the interpretation required by the evaluators, as summarized by Hertzum and Jacobsen's comment that "the principal cause for the evaluator effect is that usability evaluation is a cognitive activity which requires that the evaluators exercise judgment" (Hertzum & Jacobsen, 2001).

In our view, however, the aforementioned studies have failed to control or systematically study several factors that may impact the extent of the evaluator effect. In particular we find that too little attention has been given to factors relating to the matching of usability problems. Matching is used to compare problems from different evaluators to assess whether they concern the same or different problems. In most studies of usability evaluation, the procedures used for matching of usability problems are quite vague (Lavery et al., 1997). This appears also to hold for studies of the evaluator effect, as the studies referenced previously do not clearly report which criteria for matching they have used. Jacobsen et al. (1998a) described the matching procedure by describing how two of the authors, after splitting a few of the problem descriptions, "then examined their lists and eliminated duplicates" (p. 1337). Kessner et al. (2001) is similarly brief in writing "the evaluators independently grouped the problems into categories of problems that were essentially the same" (p. 97). Thus our focus is on the role of matching in the evaluator effect.

We focus here on two factors: (a) whether matching was done by people who either had or had not conducted the evaluation from which usability problems are being analyzed, and (b) whether matching was done individually or in groups. Both of these are motivated by the observation that whether or not problems are matched (i.e., considered similar) strongly influences the magnitude of any evaluator effect reported. Let us briefly relate these factors to previous work.

The first factor concerns who is doing the matching of usability problems that underlie the calculation of the evaluator effect. As previously mentioned, matching can be done in many different ways. Further, it proceeds on the insecure basis of very brief problem reports (Hornbæk & Frøkjær, 2005). It could be hypothesized that evaluators know more about the problems identified than is expressed in their problem reports: a lack of match among problems—and hence a high evaluator effect—could indicate just that evaluators had expressed themselves in an unclear manner, and not that they disagree.

Evidence for or against this hypothesis is precluded by the fact that studies of the evaluator effect most often let persons who have not been involved in the evaluation do the matching. Note how this factor is relevant not only for the study of the evaluator effect but for any study that compares the results of usability evaluations.

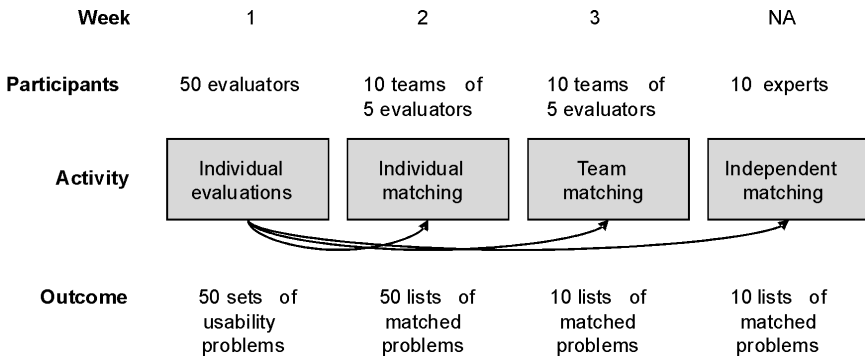
The second factor that we find lacking in previous discussions of the evaluator effect is the difference between individual and group analysis of evaluations. Hertzum et al. (2002) presented an observation pertinent to this factor. Their study found that evaluators had an overlap in reported problems of only 9%, as determined by independent matching of problems. However, when evaluators discussed their results in groups of three persons they appeared to agree: "The substantial differences in the individual reports stand in stark contrast to the perception the evaluators acquired during the group work. The evaluators left the group work with a strong, and reassuring, feeling of agreement" (Hertzum et al., 2002, p. 663). The authors drew the conclusion that the evaluators failed to notice that they disagree. An alternative view is that the evaluators were actually in agreement given an opportunity to discuss and clarify their problem reports. Perhaps evaluators agree when in contact with each other because they understand and may explain much more about the problems they have described than what is expressed in their problem descriptions. Alternatively, they may agree because of social factors (e.g., consensus-building processes and peer pressure). This factor, again, is relevant not only to the study of the evaluator effect but also to a better understanding of the benefits and drawbacks of group reporting of usability evaluations.

Next, we examine the two factors discussed above in a study that also investigates qualitative differences in the content of usability problems grouped together.

3. EMPIRICAL STUDY

The study compares assessments of the evaluator effect by evaluators versus independent experts, and by individuals versus teams.

The study was conducted over 3 weeks. In the 1st week, evaluators conducted individual think-aloud studies to find usability problems in a Web application (i.e., instances of usability problems). In the 2nd week, evaluators individually matched problems from a team of evaluators (i.e., identified types of usability problems). In the 3rd week, evaluators met with the other members of their team to match all the problems found by the team (again identifying problem types). Subsequently, matching of the problems was performed by independent experts (also identifying problem types). The procedure is summarized in Figure 1.

Figure 1. Procedure of the study.

3.1. Evaluators

Fifty computer science students (4 female) participated as evaluators in the study. Participation was part of an introductory course on human–computer interaction (HCI), which students could follow in the 3rd, 4th, and 5th year of their studies; the majority of students took the course at the 3rd year where it was their first systematic introduction to usability evaluation. At the time of their participation, evaluators had learned of models for usability, contextual design, interaction styles, visual design, and usability evaluation (e.g., inspection, think-aloud testing). Evaluators were unaware of the purpose of the study and were not explained about the evaluator effect until their participation in the study had finished.

Every evaluator was randomly assigned to a single team, each consisting of five evaluators. In Week 3 of the study the members of these teams collaborated.

Note that using students as evaluators is not optimal. However, as we expect differences in matching to be small we needed many participants; enlisting 50 usability specialists for the study was impossible. As described in Section 3.6, 10 HCI researchers also matched the problems; the Results section present the differences between the matchings of the HCI researchers and those of our students.

3.2. Application

The evaluators evaluated a Web site on science (<http://dr.dk/videnskab>), developed by the National Danish Broadcasting Company (DR), and referred to as the application in what follows. In 2004, the main site had more than 800,000 visits per week. The application was chosen because we find it

representative for many Web sites; DR also pointed it out as “in need of usability testing” because the application had not previously been thoroughly tested.

3.3. Individual Evaluation (Week 1)

In the 1st week of the study, evaluators were asked to perform a think-aloud test of the application. They were given 10 brief descriptions of tasks, developed in cooperation with DR. Evaluators received a document by Molich (2003) as an instruction in how to conduct a think-aloud study. Evaluators did the evaluation individually, each running a study with 1 participant.

To document their evaluation, evaluators describe the usability problems by a fixed format comprising five points:

- A headline, consisting of a brief characterization of the problem.
- A description that details the problem. Evaluators were asked to give as many details as possible and to ensure that their descriptions were understandable for persons who were not familiar with their think-aloud test.
- An assessment of why the problem is serious for all or some of the application’s users (e.g., because it confuses users or prevents them from finishing their tasks).
- A description of the context in which the problem arose (e.g., a particular task or part of the application).
- A description of how to solve the problem. Participants were asked to ensure that the description could be understood on its own by developers of the application, without access to the other points of the description of the problem (i.e., the other bullets in this list).

Similar problem reporting formats have been used in other research (e.g., Hornbæk & Frøkjær, 2005; Jeffries, 1994; John & Marks, 1997). Specifically, we require evaluators to describe how to solve the problem because Jeffries (1994) argued that this was a useful part of a problem report; Hornbæk and Frøkjær (2005) provided evidence that suggestions for redesigns were valued by developers.

Hertzum and Jacobsen (2001) made three recommendations on how to circumvent the evaluator effect, all of which we aimed at following with our experimental design. First, we avoid variability in task scenarios by providing tasks to the evaluators. Second, we use a fairly detailed evaluation procedure—Molich’s (2003) instruction for think-aloud studies—in an effort to reduce variability because of different understandings on part of the evaluators

on how to do a think-aloud study. Third, Molich's instruction contains hints at what may constitute a usability problem; we decided against giving more explicit criteria for what constitutes a problem (e.g., such as those used in Jacobsen et al., 1998a, or John & Marks, 1997) because evaluators were running the test individually without access to video or audio equipment. They thus had to employ a coarse definition of usability problem during or immediately after the evaluation.

3.4. Individual Matching of Problems (Week 2)

Each evaluator was assigned to a team composed of five evaluators; in Week 2 the individual matching was done on the problems identified by the team; in Week 3 the team members met and worked together (cf. Figure 1). Thus, the evaluators also matched their own problems. The rationale for having evaluators compare five sets of problems is that five appears a realistic and manageable team size; we were unable to find in the literature any firm advice on the number of evaluators needed for group matching of usability problems (though plenty of advice exists on the number of evaluators needed to conduct an evaluation).

As argued previously, the criteria used for matching problems are likely to influence the overlap between problems. We chose to use simple instructions for the participants, much depending on an intuitive understanding of what constitutes a match of usability problems. We used the following instructions for the matching:

The activity of matching usability problems has as its purpose to identify the problems that are similar and those that are different. The intention is to work up the material from several evaluators to a whole. You must therefore group those usability problems that you, given their description, consider similar.

Note that this description of what constitutes a match appears as specific as those descriptions reported in other studies of the evaluator effect (e.g., Kessner et al., 2001; Vermeeren et al., 2003). In the Discussion section, we consider the likely effects of using a more detailed instruction on how to match problems and the attempts we know of in studies of the evaluator effect of using more explicit instructions for matching.

Evaluators were given a short description of affinity diagramming (Gaffney, 1999) to support their matching of problems.

In addition to performing the matching, we required participants to record the rationale behind matching problems. As guidance to the recording we stated, "It is in this field that you should add comments or interpretations to

the instruction about grouping those usability problems that you, given their description, consider similar.” In addition, for all problems not grouped with another problem, we asked for an explanation why the particular problem was considered unique. If the participants wanted to, they could list a problem in several groups, thus splitting one problem into several.

3.5. Team Matching of Problems (Week 3)

The purpose of the 3rd week was for teams of evaluators to meet and agree on a common matching of usability problems. This was done through a discussion and matching of all the individual problems. Participants were free to choose their method for doing so but were suggested to use affinity diagramming (Gaffney, 1999). They were given the same instructions as in their individual matching regarding what would constitute a match. Note that teams matching the problems had participated in matching those individually.

Evaluators had to report the teams’ matching of problems similarly to how they reported their individual matching. After having completed the teams’ matching of problems, evaluators had to write a one-page to two-page essay about their perception of the team’s matching and of differences between the team’s matching and their individual matching of problems. Evaluators were given an open-ended question about these issues, with a few hints at what they could discuss. We chose against using a rating form for agreement because we did not want to sensitize evaluators to this particular issue.

3.6. Independent-Expert Matching of Problems

We asked 10 HCI researchers to individually match the problems from one team. These researchers were associate professors, assistant professors, or PhD students specializing in usability research; none of them had previously evaluated the application or knew the objectives of the current study. They also had not participated in any of the activities described in the preceding three sections. The rationale for including the independent experts in the study is that they have not participated in the individual evaluations: They are thus representative of the persons that typically match problems in studies of the evaluator effect.

These independent experts received the same material as was used for the individual matching of Week 2. In addition they were given a document describing the background of the evaluation and the tasks used in the evaluation, materials that the evaluators in the study had received during Week 1. The independent-expert matching was reported in the same way as the individual and the team matching.

4. ANALYSIS

Some approaches used in the analysis of the results are nonstandard in HCI research and are therefore explained next. These approaches are needed because the overlap between matchers cannot be described adequately using standard descriptive statistics such as means or percentage overlap.

The agreement between evaluators is quantified using a measure called *any-two agreement*. Hertzum and Jacobsen (2001) argued that this measure is superior to other measures of the evaluator effect because it does not require that all problems in an interface are known, nor is any-two agreement affected by the number of evaluators. Any-two agreement is defined for a set of n evaluators i, j, \dots , with problem sets P_i, P_j, \dots , as the average of

$$\frac{|P_i \cap P_j|}{|P_i \cup P_j|} \quad (1)$$

over all unique pairs of evaluators i and j , where $i \neq j$. The number of such pairs is $\frac{1}{2} n \times (n - 1)$.

We also aim at quantifying the difference between evaluators' matching of usability problems. The measure of any-two agreement works well to describe agreement between evaluators; it cannot, however, readily be used to describe the agreement between matchers. Among other reasons, an evaluator produces a set of problems; a matcher produces a set of sets of usability problems. Next we discuss three alternative measures to be used for describing the agreement of matchers. Let us first note that a *matching* is a set, say G , of *groupings*, $\{g_1, g_2, \dots, g_m\}$. Each grouping, $g \in G$, consists of one or more usability problems that have been matched together because a matcher considers them similar. Note that m may differ between evaluators because they may make a different number of groupings. However, evaluators whose groupings we compare have always received the same usability problems for their matchings.

We propose an *extension of the Jaccard measure of similarity* (Jaccard, 1912), widely used to compare expression of genes and to analyze performance in information retrieval. Jaccard similarity between two sets, P and Q , is defined similarly to the basic calculation in any-two agreement, that is, $|P \cap Q| / |P \cup Q|$. To compare two matchings, G and H , this measure cannot readily be applied because G and H are sets of sets. We therefore define a function called *bestmatch*, which given a matching $H = \{h_1, \dots, h_m\}$ and a grouping of usability problems, $g \in G$, returns the $h \in H$ with the highest Jaccard similarity to g . We

use this function on all groupings in G to sum the sizes of the union and the intersection between groupings and their best possible matches in H . We can thus calculate our extension of the Jaccard measure as:

$$\frac{\sum_{i=0}^m |g_i \cap \text{bestmatch}(g_i, H)|}{\sum_{i=0}^m |g_i \cup \text{bestmatch}(g_i, H)|} \quad (2)$$

The intuition of the formula is to consider individually every grouping in a matching, together with its best-matching grouping in the other matching chosen for comparison. For each such pair, the number of problems that overlap is related to the number of problems in total. The example in Figure 5 can support the understanding of these measures. For example, to calculate the extended Jaccard measure between the team and the independent-expert matching each grouping in the team's matching is considered in turn. The first of those is $\{a, b, c, d\}$. The bestmatch of this grouping with the independent-expert matching is $\{a, b, c, 1\}$; they have a Jaccard similarity of .75. Next we consider the grouping $\{e, f, g, h\}$; its bestmatch is $\{e, f, j\}$ with a similarity of .4. We continue doing this until the grouping $\{4\}$, the final grouping of the team's matching shown in Figure 5. From these steps we may calculate the extended Jaccard similarity between the team and the independent-expert matchings as .26.

Anderberg (1973) suggested that grouping-based measures (such as the extended Jaccard measure) and pairwise measures (similar to the any-two agreement) are quite different. We therefore also report a measure that looks at whether pairs of usability problems are grouped together or not, namely *Rand's measure of similarity* (Rand, 1971). For all pairs of problems it is considered whether they are treated similarly in the two matchings under comparison, that is, whether they—in both matchings—are either placed in the same grouping or in different groupings. The Rand measure of similarity is then simply the number of pairs treated similarly relative to the total number of pairs. Finally, we report the *specific agreement*, that is, the number of pairs of problems actually reported together in both matchings, relative to the total number of pairs.

5. RESULTS

First we present in turn the results from each of the 3 weeks in the study. Next, we present an analysis of the evaluators' comments about their matching activities. Finally, we analyze the independent-expert matching.

5.1. Individual Evaluations (Week 1)

Evaluators reported an average of 6.24 descriptions of usability problems ($SD = 2.60$), a total of 312 problems for the 50 evaluators. The number of problem descriptions ranged from 3 to 15.

Problem descriptions were on the average of 86 words long, counting all five points of the description of problems. An example of a problem is given here.

Headline: The global and the local search functions are confused.

Details: There is both a global search field for the whole of dr.dk and a local search field for Videnskab + IT [the part of the site tested]. However, both are placed in the top right corner and may be confused with each other if the user is not paying attention. The problem is particularly severe if the page has been scrolled and the upper search field has been hidden under the browser's tool bar.

Seriousness: The user can become frustrated and confused by a search that is overly broad or narrow compared to the intention of the search.

Context: Happened when the user tried to use the global search field.

Solution: Make more clear the difference between the two search fields, for example by changing the label for the local search field from "Search" (the same as for the global search field) to "Search in Videnskab + IT [the part of the site tested]."

5.2. Individual Matching of Problems (Week 2)

Figure 2 summarizes evaluators' individual matching of problems. The 50 evaluators each received an average of 31.2 problems—the average number of problems produced by a team of five evaluators—for a total of 1,560 received problems. Four percent of these were repeated in different groups of usability problems (i.e., an evaluator chose to split a problem); 1% was not included in the matching because of participants' clerical errors. In the following we analyze those problems that participants reported as separate problems, for a total of 1,600 matched problems. For the remainder of this section,

Figure 2. Evaluators' Individual Matching of Problems ($n = 50$).

	<i>M</i>	<i>SD</i>	%
No. of problems received	31.2	5.69	—
No. of problems matched	32.0	6.08	100
Problems grouped	25.9	6.31	81
Unique problems	6.14	3.86	19
No. of groups created	7.22	1.81	—

we focus on matched problems; the mentioning of received problems is made only for completeness.

On average, evaluators created 7.22 groups of problems ($SD = 1.81$). These groups contain a total of 25.86 problems, about 81% of the total number of problems matched by each evaluator. Evaluators pointed out an average of 6.14 problems ($SD = 4.69$) as unique, that is, did not group them with other problems. This amounts to 19% of the problems matched. In terms of the often-used distinction between usability problem tokens and types, the number of matched problems counts as the problem tokens; the sum of the number of groups and the number of unique problems counts as the problem types.

Figure 3 reports the any-two agreement, as calculated from each of the results of the 50 individual matchings. The average any-two agreement is 42% ($SD = 13.79\%$), meaning that on average a pair of evaluators will agree on a little less than half of the problems they have identified. This number is within the range identified by Hertzum and Jacobsen (2001). Note that 42% is the average of 50 measures of the evaluator effect, each calculated from an individual matching. The range of these measures of the evaluator effect shows agreement of 8% to 63%.

It could be expected that evaluators would treat problems that they found themselves differently from problems that other evaluators found. Of interest, a smaller percentage of the participants' own problems were considered unique ($M = 13.7\%$, $SD = 17.6$) compared to those problems found by others ($M = 19.6\%$, $SD = 13.21$), $F(1, 49) = 6.23$, $p < .05$. Thus, participants may better be able to interpret their own problems and decide for a match. This difference does not result in a higher any-two agreement when only matches with an evaluator's own problems are considered. To confirm this, we broke down the any-two agreements calculated from an individual matching made by considering separately (a) pairs of evaluators that include the person who made the matching, and (b) pairs of evaluators that do not include the person who made the matching. Figure 3 shows these as the "Own usability problems" row and the "Other evaluators' usability problems" row, respectively.

Figure 3. Average Any-Two Agreement (%) Calculated From the Individual Matching (Rows 1–3), Teams' Matching (Row 4), and Independent-Expert Matching (Last Row).

Any-Two Agreement Basis	<i>N</i>	%	<i>SD</i>
Individual matching (all problems)	50	41.7	13.79
Own usability problems	—	42.3	14.71
Other evaluators' usability problems	—	41.2	14.98
Team matching	10	46.5	13.66
Independent-expert matching	10	41.4	13.18

As may be seen from the figure, there is no significant difference between these numbers, $F(1, 49) = 0.39, p > .5$. Thus, evaluators find fewer unique problems among the problems they had identified themselves, but this is not a strong enough effect to affect the calculation of the any-two agreement.

In summary, the individual matching shows a considerable evaluator effect but also considerable variation in the magnitude of the effect.

5.3. Team Matching of Problems (Week 3)

Figure 4 summarizes the teams' matching of problems. The teams received an average of 31.2 problems. The teams found that 4% of these could be placed in more than one group, that is, an average of 32.7 problems was matched. Next we analyze the difference between individual and team matchings based on the number of problems matched.

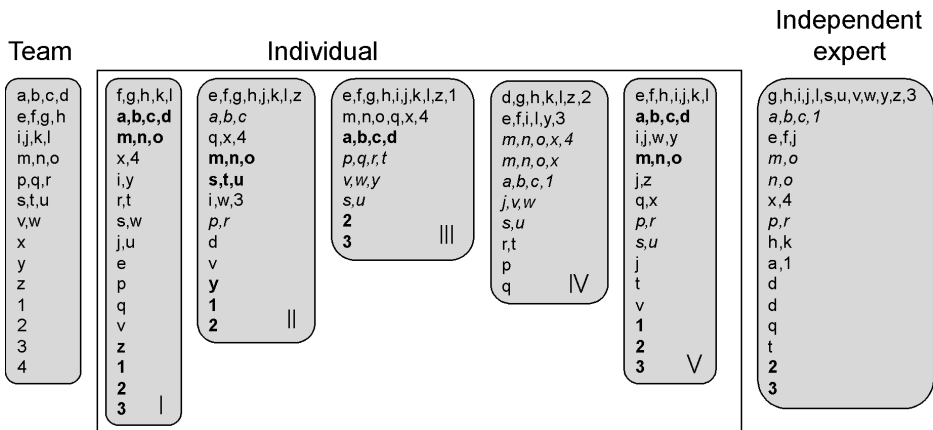
Overall, teams' matchings of problems show a higher agreement compared to individual matching. When related to Figure 2, Figure 4 shows a difference between the problems that are grouped by teams (87%) as compared to those grouped by individual matchers (81%). This difference is significant, $F(1, 49) = 9.06, p < .01$. Note that this and the following tests are made on the level of individual evaluators, which is the reason for the 49 *df* for the error term. Conversely, the number of unique problems is higher for individuals (19%) compared to teams (13%). Figure 3 shows the average any-two agreement calculated from the teams' groupings. The average any-two agreements derived from individuals' matching (42%) and from teams' matching (47%) are also significantly different, $F(1, 49) = 4.37, p < .05$. Note, however, that the sizes of these effects are small.

Thus, quantitative measures suggest a difference between the matching done by individuals and that done by teams. The aforementioned analysis, however, does not address the difference between the actual groupings made by individual evaluators and their teams. As an example of the differences in groupings consider Figure 5. This figure shows the matchings of a representative team (in terms of average any-two measures and number of problems

Figure 4. Teams of Evaluators' Matching of Problems ($n = 10$).

	<i>M</i>	<i>SD</i>	%
No. of problems received	31.2	5.69	—
No. of problems matched	32.7	6.57	100
Problems grouped	28.3	4.95	87
Unique problems	4.40	3.60	13
No. of groups created	7.50	1.78	—

Figure 5. Illustration of matching for a representative team. To the left is shown the team's matching and to the right the independent experts. In between are the five individual matchings (labeled I to V). Problems are denoted a to z and 1 to 4. Within any of the seven matchings, a set of problems appearing on a line separated by commas were considered by the matcher to be similar. The relation between sets of problems in the individual/independent-expert matchings and the closest matching set in the teams matching are visually indicated as follows: a Jaccard similarity of 1 is shown in bold, a Jaccard similarity larger than .5 but less than 1 is shown in italic. The figure and the overall overlap in the contents of the matching described in Section 3.5 indicate a lack of overlap between matchings.



that are matched). It is clear from the figure that there is little overlap between the individual matchings and those of the team.

To describe these differences in the content of the matching, Figure 6 summarizes some measures of the overlap. The average of the extended Jaccard measure of similarity between teams and individual evaluators is .52 ($SD = 13.5$). Roughly, this means that half of the problems in a grouping produced by any participant may be found together in the best-matching group of the team. The average of Rand's measure of similarity between teams and individuals is .90 ($SD = .05$). However, given the nature of the groupings this measure is not very informative: most of the similarity arises from the large number of cases where usability problems are not grouped together by the team or by individuals. The specific agreement is therefore only 6%, meaning that for any pair of problems an individual evaluator has grouped together, there is only a 6% chance that the team has likewise grouped that pair of problems together. As a more directly interpretable indicator of the difference in matchings, it is notable that among the 668 groupings that comprised the individual matchings, 30% ($SD = 15.9$) were identical to a grouping found in the team's matching. However, around half of the identical groupings consisted of only one problem.

Figure 6. Measures of Extended Jaccard Similarity, Rand's Similarity, and Specific Agreement for the Matchings Made by Individuals Versus Teams, by Individuals Versus Independent-Expert Matchers, and Among Individuals.

Comparison	Extended Jaccard		Rand		Specific Agreement	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Individual vs. Team	.52	.10	.90	.05	.06	.03
Individual vs. Independent-Expert	.44	.14	.87	.05	.04	.02
Individual vs. Individual	.49	.10	.89	.03	.04	.02

Note. All $n = 50$ (i.e., based on values for individual evaluators).

In summary, we find a small difference in agreement between individual and group matching. However, the groupings produced are very different as suggested by Figure 6.

5.4. Independent-Expert Matching

Figure 7 summarizes the results of the independent-expert matching. These are compared to the individual matching to see if having participated in an evaluation (as opposed to not having been part of the evaluation) affects matching.

The summary data in Figure 2 and Figure 7 suggest few differences. We find no difference in the number of problems that are seen as unique (independent-expert matchers 16%, individual matchers 19%), $F(1, 49) = 0.72$, $p > .4$. Again, tests are made at the level of individual evaluators. The number of groups of problems created is also comparable (independent-expert 8.3, individual 7.2). Figure 3 shows the measures of the evaluator effect that can be derived from the independent-expert matching. The average any-two agreement is about 41%, with a range between 29% and 69%. The average any-two agreement is similar to that of the individual ratings (which was 42%).

Figure 7. Independent-Expert Matching of Problems ($n = 10$).

	<i>M</i>	<i>SD</i>	%
No. of problems received	31.2	5.69	—
No. of problems matched	33.4	7.85	100
Problems grouped	27.9	5.26	84
Unique problems	5.50	3.95	16
No. of groups created	8.3	1.77	—

However, we again observe marked differences in the matching of problems at the content level (see Figure 6). The average of the extended Jaccard measure of similarity between individual matchings and the independent-expert matchings is .44. Compared to the difference among individual matchers (extended Jaccard is .49), this is significantly lower, $F(1, 49) = 11.35, p < .001$.

5.5. Comments on the Individual and Team Matching

To understand the evaluators' experience of the individual and the team matching, we analyzed the essays they wrote after Week 3 of the study. From these essays we extracted 267 comments about the matching, comments that upon reading the notes were seen as good characterizations of what an evaluator found relevant. So as to avoid unclear comments we only extracted comments that clearly described one or more aspects of matching. These 267 comments concern four themes: (a) satisfaction with the team matching compared to the individual matching, 69 comments; (b) difference or similarity between the individual and group matching, 92 comments; (c) criteria for the team matching, 81 comments; and (d) criteria for the individual matching, 25 comments. Because essays were open ended (to avoid directly prompting for information and sensitizing participants to our interests), not all participants commented on all of these themes. Next we discuss these themes.

Regarding the team matching in comparison to the individual matching, 31 evaluators (62%) commented that they find the team matching better or clearly better than their individual matching, for example, that "in all, I find the new grouping better than the one I found myself" or "in all, the team matching gave better insight, better overview and larger collections of problems." Eight evaluators listed advantages and disadvantages of both the team and individual matching, and one evaluator commented that "the team matching was not better than the individual matching."

Twenty-two evaluators specifically commented on the utility of the discussions during the team matching, suggesting that it gave a "better and more certain understanding of the usability problems," for example:

Many different viewpoints and opinions emerged during the team matching, and for several usability problems it meant that I understood the problem differently after the team meeting compared to before.

In particular, 13 evaluators pointed out the usefulness of having present during the team matching the evaluator who had conducted the think-aloud test and written a particular problem. For example, one wrote,

Each author of usability problems had an opportunity to elaborate on the description of the problem. That has played a large role in making the list of unique usability problems shorter, when it turned out that I had not understood the individual descriptions of usability problems accurately enough.

And another:

The short problem description gives an overview of the nature and extent of the problems, but is too brief to enable a full understanding of the other think-aloud participant's experiences—it is therefore not until the team work that data from other than ones own participant are done justice.

Regarding the overlap between individual and team matches, 17 evaluators wrote something clear and specific. Of these, 11 found little difference between their individual and the team's matching. One of them wrote:

With the exception that many of my unique problems have disappeared during the team matching, I find my individual matching and the team matching very similar. Most groups of problems are about the same problems, simply with a re-written common focus.

Six of the 17 evaluators discussing the overlap between matchings pointed out significant differences, for example:

One reason that we have so many common groupings and so large groups is compromises. We only had long discussions about questions where we had large disagreements.

Many evaluators made comments about how they matched problems (14 about team matching, 7 about individual matching, both excluding comments about the use of affinity diagramming). At least three matching criteria were used: (a) similarity regarding problem type, for example lack of overview in navigation; (b) similarity in terms of solution, typically seen from the programmer's viewpoint; and (c) similarity of area of the site where a problem is found, for example, a particular object in the user interface (UI). In team matching, three evaluators mentioned that their group combined criteria (a) and (b), two evaluators mentioned that the group used criteria (b), one evaluator that the group used (a), and one evaluator that the group used (c). Regarding the individual matching, three evaluators say they used matching criteria (b), two used (c), and one used (a). Combination of matching criteria was not explicitly mentioned.

Four evaluators mentioned that the team discussed the level of abstraction at which to describe problems, for example: "the significant differences in our individual grouping of topics were the choice of level of aggregation—it is

mostly about finding an appropriate level of abstraction.” Often, the team matching leads to a matching at a higher level of abstraction compared to the individual matchings. Some evaluators suggest that such changes, in addition to giving a better overview, also are a way to reach agreement.

6. DISCUSSION

The data from the study show a substantial evaluator effect, independently of whether the effect is calculated from the matchings of individuals, teams or independent experts. Surprisingly, usability problems are matched very differently among matchers. We find negligible overlap between the groupings formed, and for a given grouping of problems the best-matching grouping from another persons’ matching will only match on half of its problems. Thus it appears that matchers fundamentally disagree about which problems are similar. Regarding the comparison of individual versus team matching, it appears that when teams of evaluators meet to match problems they produce matchings with fewer unique problems. These matchings also show higher any-two agreement, and evaluators are more satisfied with them. The matching bears little resemblance, in terms of problems grouped together, to the individual matchings. Regarding the comparison of evaluators versus independent experts, we find that independent-experts are less similar to individual matchings than the average similarity among individual matchings.

The main question to be discussed is how our results relate to the research on the evaluator effect discussed in the section on related work. On the surface, our study confirms the existence of a substantial evaluator effect and thus corroborates the findings of previous studies of the evaluator effect (Hertzum & Jacobsen, 1999; Hertzum et al., 2002; Jacobsen et al., 1998a, 1998b; Kessner et al., 2001; Vermeeren et al., 2003). However, compared to these studies we are also able to say something about the variability of the evaluator effect. Our data show that evaluators’ individual matchings lead to varying measures of the evaluator effect (agreement varying between 8% and 63%), suggesting that individual variations in matching strongly influence the evaluator effect.

The foundation on which measures of the evaluator effect rest—the matching of usability problems—is highly variable. We have documented how the matchings produced by evaluators and by independent experts show little similarity in content, neither within nor between these groups. Our preferred measure for quantifying this difference, the extended Jaccard measure, suggests that the best match of grouped problems between two matchings is only identical in half of its elements. This disagreement in the content of matchings appears as large as the disagreement on the identification of usability prob-

lems. The quote by Hertzum and Jacobsen (2001) mentioned earlier (“the principal cause for the evaluator effect is that usability evaluation is a cognitive activity which requires that the evaluators exercise judgment”) is equally true for matching as for identification of usability problems. We see this as the main point of this article. If we accept the evaluator effect and are wary of its influence on usability evaluation, we should be equally wary the influence of the “matcher effect” (i.e., the difference between persons’ matchings) on usability research.

This finding has a number of implications for existing studies of the evaluator effect. Overall, it seems that these studies’ estimates of the evaluator effect are less reliable than commonly assumed. The reliability is reduced because the studies are based on matching, which we have seen to be variable. To a large extent the magnitude of the evaluator effect also depends on the exact definition of matching that is used in the individual studies of the effect. No previous study of the evaluator effect has made this clear.

It is also noteworthy that the agreement between matchers reported in previous studies of the evaluator effect (such as Hertzum et al., 2002; Jacobsen et al., 1998a; Kessner et al., 2001) is usually much higher than what we found. The agreement is typically reported in percentage of agreement between two matchers (in the studies just referenced, it is 86%, 80%, and 68%, respectively). One difference to our study is that we did not ask participants to find “unique problem tokens” (Jacobsen et al., 1998a) but merely to group similar problems. Agreement to a large extent depends on the number of unique problems found; previous studies rarely report the number of such problems. An exception is Hertzum et al. (2002), who wrote that 79% of their matched problems were unique. But if 79% of the problems are unique then an agreement of 68% is not really impressive because it may indicate that matchers agree only about the unique problems and disagree about all cases where two are more problems have been matched. When all matchers identify many unique problems, reliability expressed in terms of percentages will become much higher, and any measure of the evaluator effect will show more pronounced differences among evaluators. Again, matching pops up as the Achilles heel of the measures of the evaluator effect.

In a nutshell, how to match depends on a notion of similarity. As pointed out in the section on related work, most studies of the evaluator effect do not report in detail how they matched. In the Hertzum, Jacobsen, and John studies, however, a more specific procedure and criterion were used (M. Hertzum & N. E. Jacobsen, personal communication, November 2005):

The matching was performed by developing an experience-based coding scheme. This coding scheme handles both the intention of and the procedure for the matching, and gives a template to use when matching. The coding scheme was

iterated a few times and ended up being a few pages long before it was put to use. The procedure included a two-step coding, where problems were first sorted according to the user interface architecture of the system (where are problems located in the UI?) and second matched within each main part of the user interface. Both steps were done independently by both authors. Next it was examined which problems the two of the authors agreed/disagreed about. Disagreements were registered and discussed to reach a consensus. In cases of doubt, both in the individual and final consensus, problems were matched rather than keeping them on their own.... Our fundamental rule [for matching] is to define two problems as dissimilar if they, according to our best understanding, would lead to different corrections in the system.

Note that most of this description concerns the procedure for matching, rather than the criteria for considering problems similar. The lines describing the criteria, however, raise many questions. Does this procedure of matching inflate the magnitude of the evaluator effect because it by definition considers separately problems concerning different parts of the UI? Already Lavery et al. (1997), for instance, criticized this approach to matching. The criterion that problems are dissimilar if they lead to different corrections in the system also opens a lot of problems. For example, would descriptions that point out the same observed difficulty but mention different ways of alleviating it be considered dissimilar? Also, a question remains how to handle problems that do not mention any particular solution to a problem. Again it appears that the notion of evaluator effect rests on a particular understanding of matching which to a large extent determines the magnitude of that effect.

A possible objection could be that we have merely shown that matching is unreliable, something that was already suggested in the literature (Lavery et al., 1997) and something that more elaborate matching procedures (Andre, Hartson, Belz, & McCreary, 2001; Cockton & Lavery, 1999) may ameliorate. We do not find this objection convincing. First, such procedures have never been used in showing the evaluator effect; the most detailed procedure we know of is that by Hertzum and Jacobsen (as just described). As already pointed out, it—in virtue of its definition of similarity—seems to lead to a substantial evaluator effect. Second, the reliability of matching procedures is not well established. The evaluation by Andre et al. (2001) of the User Action Framework (UAF), intended to assist in classification of usability problems, does not reach a high overall reliability and may not improve matching. Third, our study gives some insight into how matchers relate usability problems, suggesting that matching procedures assuming only one possible classification of a usability problem guided by a standardized use of a large set of classification rules might not work. In the group matching, for example, the evaluators found it necessary and obvious to reinterpret usability problems in response to convincing arguments, and to apply and even combine alternative matching criteria. It is also the case that teams which in this study stick to

just one criterion for matching do not show less variability than groups using multiple criteria.

Another objection may be that our results are based on the work of novice evaluators; as discussed earlier we do not see that this study was practically feasible with experienced evaluators. We see no reason for expecting that matching would be less variable for more proficient evaluators, though we acknowledge that this is ultimately an empirical question. In this study, however, the matchings of the independent experts suggests few differences to the novice evaluators and matchers.

It should be noted that a number of earlier studies have hinted that matching may be quite variable, in a sense predating the general point of this paper. Connell and Hammond (1999), for instance, showed that matching was affected by the level of abstraction at which problems were aggregated. In the Supex framework (Cockton & Lavery, 1999) it is possible to match usability problems at varying levels of abstraction, thereby reaching varying levels of overlap between sets of problems (and thereby also the estimate of the evaluator effect). None of these papers, however, suggests that matching has a key role in the evaluator effect or provides empirical evidence about how matching is performed.

The difference between individual matchings and the team matchings following the individual matching is an important result. Though we have no objective measure of how good a matching is, it appears that the team process creates matchings that the evaluators are more satisfied with and that contain fewer unique problems. Note that these matchings are quite different from the matchings of the individual team members. New insights or interpretations appear to emerge during the group process; evaluators' comments confirm this speculation. Note that there is a key difference between our team matchings and the matching by a pair of researchers typically used in other studies of the evaluator effect (Hertzum et al., 2002; Jacobsen et al., 1998a; Kessner et al., 2001). In our study the teams redid the matching compared to their initial work (and reached a new result); the pair of researchers typically only discusses problems on which they disagree. In this study it should also be noted that the difference between individual and team matching is confounded with order; that is, the team matching always happens last. However, it appears illogical to try to reverse the order, so that group matching is done first because group work demands that the individuals have oriented themselves in the problems (i.e., to some extent forming an individual opinion about the problems). Thus we should think of the team matching as the series of individual and the team matchings. We note that there are large—in particular, qualitative—differences between matchings of the individuals and the teams. It is also the case that individuals treat their own problems differently from those problems of other evaluators, something that we also expect to be unrelated to the order of matchings.

We find some difference between independent-expert matching and matching performed by persons who have participated in testing. Evaluators commented that one reason for the success of the team matching is that evaluators who have observed a problem are present to discuss it. Although this difference is small, it indicates that previous studies of the evaluator effect that use matchers who have not participated in testing may underestimate the overlap between usability problems, and therefore find a lower agreement between evaluators.

Our study has some implications for practical usability work. No one would probably wish to undertake matching as part of their usability testing. Nevertheless, comparison of findings to see differences and similarities are what any analysis of results from usability evaluations is about. For practical usability evaluations, our results suggest team matching of usability problems as a solid strategy. It also seems that usability evaluators would be well advised to consider any comparison or matching of problems as something dynamic. Hertzum and Jacobsen (2001) discussed how to diminish the evaluator effect, for instance, by reducing variability in task scenarios and being specific about what constitutes a usability problem.

For the usability research, we see the main challenges for future research as the following. First, it is not clear from this study how the problem reporting format may matter in matching and consequently in establishing the evaluator effect. Some accounts suggest that problem reporting is key to matching (e.g., Lavery et al., 1997), possibly implying that more elaborate formats may help matching. We consider this an empirical question to be answered by future research. Second, matching by features in the UI is used in some of the matching approaches discussed. It could be interesting to compare empirically different ways of matching (e.g., by features in the UI, by the components suggested by Lavery et al., and by UAF) to see how they impact the evaluator effect. Third, future work should probe the difference between individual and team matchings further.

A broader and very important research question in a discussion of the evaluator effect and of the matching of usability problems is whether a substantial evaluator effect is really a problem. The studies of the evaluator effect seem to assume that a notion of the sameness of usability problem makes sense, or as Hartson, Andre, and Williges (2001) wrote, “one needs the ability to determine when two different usability problem descriptions are referring to the same underlying problem” (p. 387). This suggests that “the same underlying problem” is real, relatively static, and to be discovered in the software, not only in situated usage of the software as reflected in one particular formulation of a problem description. Given that assumption, it makes sense to seek a correct matching of problems because whether problems correspond to the same underlying problem should be clear from the description of the prob-

lems; the more able a matcher is at determining which underlying problem particular usability problems correspond to, the closer a matching will be to the correct matching of problems.

In concluding, we propose an alternative view of the evaluator effect suggesting that descriptions of usability problems are merely incomplete expressions of observations and inferences from usability evaluations, which may be useful in some practical activities such as improving an interface. This view further posits that a matching of usability problems may be understood as a way for an individual or a team to create an overview and a more clear understanding of a set of problems—associations between problems are created in the matcher's stream of thought merely by having some abstract property in common. On another occasion, possibly with another purpose, the matching could be different yet equally useful. Although we do not want to suggest that any problem may match any other problem, the range of possible matches appears far greater than commonly assumed. This view appears partly supported by our study through the variety in which individual evaluators match, and through the flexibility with which they, during the team matching, construct an entirely different view of the problems. The brief and occasionally hard-to-understand descriptions of usability problems also suggests that a view open to variability and changing interpretations, as the one presented here, is probably more fitting and descriptive than the ideal of a perfect or correct matching. The implications for research of this view are that we cannot assume a correct matching, nor that two usability problems may in general be said to be similar or different.

7. CONCLUSION

We have presented a study of the evaluator effect, that is, the finding that evaluators in similar conditions construct substantially different sets of usability problems. Previous studies of the evaluator effect, however, have not focused on the matching of similar usability problems that underlie any calculation of the evaluation effect.

This study has shown how matching is highly variable, documenting large differences in the content of problems matched. Matching in teams after the individual matching produces more coherent results and is preferred by participants. However, previous assertions about the evaluator effect are questionable because the matchings used to calculate this effect are as variable as the evaluator effect itself.

The results indicate that further research is needed concerning the evaluator effect. However, that research should not assume that there is one correct matching of problems. Practitioners are advised to use teams of evaluators to analyze usability problems.

 NOTES

Acknowledgments. We are grateful for comments on a draft of this article by Morten Hertzum, Niels Jacobsen, Effie Law, Peter Naur, and Mikael Skov. We also wish to thank the evaluators in the experiment and the persons who assisted with the independent-expert matching: Hasse Clausen, Torkil Clemmensen, Morten Hertzum, Rune Høegh, Janne Juul Jensen, Jesper Kjeldskov, Marta Lárusdóttir, Mie Nørgaard, Mikael Skov, Jan Stage, Georg Strøm, and Tobias Uldall-Espersen.

Support. This work was supported by grant #2106-04-0022 from the Danish Research Council.

Authors' Present Addresses. Kasper Hornbæk, Department of Computer Science, University of Copenhagen, Njalsgade 128, DK-2300 Copenhagen, Denmark. E-mail: kash@diku.dk. Erik Frøkjær, Department of Computer Science, University of Copenhagen, Njalsgade 128, DK-2300 Copenhagen, Denmark. E-mail: erikf@diku.dk.

HCI Editorial Record. First manuscript received January 12, 2006. Revisions received November 14, 2006 and May 15, 2007. Accepted by Judith Olson. Final manuscript received August 13, 2007.— *Editor*

REFERENCES

- Anderberg, M. A. (1973). *Cluster analysis for applications*. New York: Academic.
- 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*, 107–136.
- Cockton, G., & Lavery, D. (1999). A framework for usability problem extraction. *Proceedings of the 7th IFIP TC.13 International Conference on Human-Computer Interaction (Interact'99)*. Amsterdam: IOS Press.
- Connell, I. W., & Hammond, N. V. (1999). Comparing usability evaluation principles with heuristics: problem instances versus problem types. *Proceedings of the 7th IFIP TC.13 International Conference on Human-Computer Interaction (Interact'99)*. Amsterdam: IOS Press.
- Gaffney, G. (1999). *Affinity diagramming*. Windsor, Australia: Information & Design. Retrieved November 7, 2006, from <http://www.infodesign.com.au/usabilityresources/general/affinitydiagramming.asp>
- Hartson, H. R., Andre, T. S., & Williges, R. C. (2001). Criteria for evaluating usability evaluation methods. *International Journal of Human-Computer Interaction*, *13*, 373–410.
- Hertzum, M., & Jacobsen, N. E. (1999). The evaluator effect during first-time use of the cognitive walkthrough technique. *Proceedings of HCI International '99*. London: Erlbaum.
- Hertzum, M., & Jacobsen, N. E. (2001). The evaluator effect: A chilling fact about usability evaluation methods. *International Journal of Human-Computer Interaction*, *13*, 421–443.
- Hertzum, M., Jacobsen, N. E., & Molich, R. (2002). Usability inspections by groups of specialist: Perceived agreement in spite of disparate observations. *Extended Abstracts*

- of the ACM Conference on Human Factors in Computing Systems (CHI 2002)*. New York: ACM Press.
- Hornbæk, K., & Frøkjær, E. (2005). Comparing usability problems and redesign proposals as input to practical systems development. *Proceedings of CHI2005 Conference on Human Factors in Computing Systems*. New York: ACM Press.
- Jaccard, P. (1912). The distribution of the flora of the alpine zone. *New Phytologist*, 11, 37–50.
- Jacobsen, N., Hertzum, M., & John, B. (1998a). The evaluator effect in usability studies: Problem detection and severity judgments. *Proceedings of 42nd Annual Meeting of the Human Factors and Ergonomics Society*. Santa Monica, CA: HFES.
- Jacobsen, N. E., Hertzum, M., & John, B. E. (1998b). The evaluator effect in usability tests. *Conference Summary of CHI'98 Conference on Human Factors in Computing Systems*. New York: ACM Press.
- 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: Wiley.
- John, B. E., & Marks, S. J. (1997). Tracking the effectiveness of usability evaluation methods. *Behaviour & Information Technology*, 16, 188–202.
- Kessner, M., Wood, J., Dillon, R. F., & West, R. L. (2001). On the reliability of usability testing. *Extended Abstracts of CHI 2001 Conference on Human Factors in Computing Systems*. New York: ACM Press.
- Lavery, D., Cockton, G., & Atkinson, M. P. (1997). Comparison of evaluation methods using structured usability problem reports. *Behaviour & Information Technology*, 16, 246–266.
- Molich, R. (2003). User testing, Discount user testing. *DialogDesign*. Available from <http://www.dialogdesign.dk>
- Molich, R., Ede, M. R., Kaasgaard, K., & Karyukin, B. (2004). Comparative usability evaluation. *Behaviour & Information Technology*, 23, 65–74.
- Nielsen, J. (1992). Finding usability problems through heuristic evaluation. *Proceedings of CHI'92 Conference on Human Factors in Computing Systems*. New York: ACM Press.
- Nielsen, J., & Molich, R. (1990). Heuristic evaluation of user interfaces. *Proceedings of CHI'90 Conference on Human Factors in Computing Systems*. New York: ACM Press.
- Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66, 846–850.
- Vermeeren, A., van Kesteren, I., & Bekker, M. (2003). Managing the evaluator effect in user testing. *Proceedings of IFIP Conference on Human-Computer Interaction (Interact 2003)*. Amsterdam: IOS Press.