International Journal of Human-Computer Interaction, vol. 21, no. 2 (2006), pp. 125-146.
Preprint version

Problem Prioritization in Usability Evaluation: From Severity Assessments toward Impact on Design

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

Abstract. Severity assessments enable prioritization of problems encountered during usability evaluations and thereby provide a device for guiding the utilization of design resources. However, designers' response to usability evaluations is also influenced by other factors, which may overshadow severity. With the purpose of enhancing the impact of severity assessments, this study combines a field study of factors that influence the impact of evaluations with an experimental study of severity assessments made during usability inspections. The results show that even in a project receptive to input from evaluations their impact was highly dependent on conducting evaluations early. This accorded with an informal method that blended elements of usability evaluation and participatory design and could be extended with user-made severity assessments. The major cost associated with the evaluations was not finding but fixing problems, emphasizing that to be effective severity assessments must be reliable, valid, and sufficiently persuasive to justify the cost of fixing problems. For the usability inspections, evaluators' ratings of problem impact and persistence were weakly correlated with the number of evaluators reporting a problem, indicating that different evaluators represent different subgroups of users or alternatively that evaluator-made severity assessments are of questionable reliability. To call designers' attention to the severe problems, the halving of the severity sum is proposed as a means of visualizing the large payoff of fixing a high-severity problem and, conversely, the modest potential of spending resources on low-severity problems.

Keywords: usability evaluation methods, problem prioritization, severity assessments, test impact

1 Introduction

Current usability evaluation methods, such as thinking-aloud studies and heuristic evaluation, focus on problem detection and treat severity assessment superficially (Cockton, Lavery, & Woolrych, 2003; Dumas, 2003). This is unfortunate for several reasons. First, practitioners often have the resources to fix only a subset of the problems identified in a usability evaluation. Thus, problem prioritization is inevitable, but if severity assessments are absent or weak they cannot guide the selection of which problems to address. Second, usability evaluation methods' superficial approach to severity assessment reflects that neither the constituent elements of severity nor their interrelations are well understood. Thus, widely held beliefs about severity may rest on uncertain evidence. One such belief holds that there is a correlation between severity and the frequency with which problems occur in usability evaluations and, hence, that the more severe problems are likely to turn up within the first few users of a thinking-aloud study (e.g., Virzi, 1992). Third, research on and comparisons of usability evaluation methods overemphasize problem detection and lack adequate means of analysing and expressing the comparatively large effect of fixing the few problems of high severity. Thus, usability evaluation methods are often assessed more on their brainstorming-like resources for discovering candidate problems than on their analysis resources for confirming or eliminating candidate problems (Cockton et al., 2003).

Severity assessments are a device for providing designers with guidance about the order in which problems should be addressed. There are, however, other factors that also enter into determining whether designers address the problems identified in a usability evaluation. Unless these competing factors are taken into account when an evaluation is planned and performed they are likely to overshadow severity assessments. Thus, for severity assessments to be effective we must know how to make them reliably and convincingly and, at the same time, understand and manage the competing factors that may otherwise determine designers' response to usability evaluations. This study analyses both these aspects of severity assessments. First, the extent to which designers are persuaded to make changes to their system in response to an evaluation concerns the fit between the evaluation and the surrounding design process. This aspect of usability evaluations, which has been referred to as persuasive power (John & Marks, 1997) and impact (Sawyer, Flanders, & Wixon, 1996), will be analysed based on a nine-month field study of the iterative design of a system for local government authorities. Second, within usability evaluations severity is typically expressed by a rating that provides an estimate of the amount

of delay, difficulty, and other inconveniences inflicted on users. This aspect of usability evaluations concerns how severity assessments are made and will be analysed based on an experimental study of the severity assessments made by eleven experienced usability professionals during a usability inspection. The purpose of this dual perspective on severity assessments is to investigate:

- factors that determine the impact of evaluations on the design process
- correlations between different aspects, or indicators, of severity
- how severity information can be made a more prominent and integrated element of evaluation and design

After outlining related work on factors that enter into determining the impact of evaluation on design (Section 2), this study briefly reviews previous work on the assessment of problem severity (Section 3). Then, the factors that determined the impact of the usability evaluations in a concrete design process are analysed based on field-study data about the interaction between the evaluations and the surrounding design process (Section 4). To complement this contextual analysis the paper, then, turns to a contrasting analysis of an experimental study in which evaluators in a usability inspection made separate assessments of different aspects of problem severity (Section 5). Taken together the two analyses point toward ways of and barriers in using severity information to ensure effective utilization of design resources (Section 6).

2 The impact of evaluation on design

Systematic evaluation at suitable points in the design process is necessary to get, assess, align, and mesh input from the different groups of people with a stake in the process (Gould, Boies, & Lewis, 1991) and is generally more realistic than attempting to formalize the design process itself (Pejtersen & Rasmussen, 1997). To fulfil this pivotal role, evaluation must draw on a variety of methods and adjust them dynamically to the specifics of the situation. This is a complex process and consequently multiple factors affect the impact an evaluation has on design. Sawyer et al. (1996) define the impact of a usability evaluation as the number of solved problems divided by the total number of problems found. Whenever a problem predictive of actual use is left unaddressed an opportunity to improve the evaluated system is missed and the effort that went into finding the problem is wasted (John & Marks, 1997). The time required to fix a problem must, however, be weighted against the benefit of fixing it and the potential of spending the resources on other outstanding tasks. Whiteside, Bennett, and Holtzblatt (1988) evaluated multiple versions of a single system and report an impact of 65% for the early, in-house evaluations and 48% for the subsequent field tests. While impact is most directly seen as the ability of an evaluation to bring about changes in the evaluated system, an evaluation may also be directed toward and have an impact on other actors in the design process, such as management and marketing (Brooks, 1994; Zirkler & Ballman, 1994). Factors such as video highlights (Dumas & Redish, 1999), wording of problem descriptions (Dumas, Molich, & Jeffries, 2004), and redesign proposals (Hornbæk & Frøkjær, 2005) have been found to influence persuasiveness and impact directly. A host of other factors influences impact in more indirect ways and must also be considered in matching usability evaluation with design conditions. These factors can be grouped into four broad categories.

Applicability. While some evaluation methods are applicable to a broad range of systems and design-process stages, others are directed toward specific stages or classes of systems. For example, usability inspections such as heuristic evaluations (Nielsen, 1994) can be performed in the early stages of the design process based on design specifications; thinking-aloud studies (Dumas & Redish, 1999) typically involve running prototypes; and beta tests (Smilowitz, Darnell, & Benson, 1994) require a functional system and are consequently confined to the late stages of the design process. Depending on the design-process stage designers may take a primary interest in different types of issues, which may or may not correspond to the types of problems an evaluation method covers the best. An important distinction is between problems concerning whether the functionality of the system in principle can do what is needed and problems concerning how well users can use that functionality. Further, some evaluation methods are directed toward specific obstacles in the evaluation situation or specific types of systems, such as mobile systems (Kjeldskov & Stage, 2004).

Validity. Gaps between the evaluation and the real world introduce a risk that what appears as a problem during evaluation will not be a problem during actual, real-world use and that some of the problems that will surface during actual use will not surface during evaluation. Such gaps may have various origins, including gaps concerning the user, task, artefact, and work context (Thomas & Kellogg, 1989). While this is a crucial issue, little is known about to what extent the problems detected during evaluations are in fact valid. In a notable study Bailey, Allan, and Raiello (1992) found that only two of the 29 problems encountered in a heuristic evaluation had an impact on the users' task completion times and subjective preference. Potentially troubling evidence has also been reported for thinking-aloud studies (Boren & Ramey, 2000; Cordes, 2001). The general recommendation for avoiding validity gaps is to conduct evaluations in the field (e.g., Whiteside et al., 1988) but this entails less control over evaluation parameters and thereby reduced reliability.

Reliability. A usability evaluation method is reliable when a rerun of an evaluation will yield essentially the same results across a range of minor variations in the test situation. The effect of many such variations have been investigated, including the sufficient number of users (e.g., Lewis, 1994) and individual versus cooperating users (Hackman & Biers, 1992). Recently, Hertzum and Jacobsen (2003) have shown that thinking-aloud studies are subject to a substantial evaluator effect in that different evaluators who analyse the same test sessions detect markedly different sets of problems. Another issue that affects usability evaluations and may vary from one evaluation to the next is the users' reaction to the evaluation setting, which typically involves trying to use a new system and performing in front of others, two circumstances that are experienced as unpleasant or stressful by many people. Whereas laboratory-like evaluations provide a controlled environment where various sources of variability can be kept to a minimum, the resulting reliability must be balanced against a reduction in validity.

Costs. Several studies provide formulas to estimate the costs of conducting usability evaluations and try to justify these costs by converting the estimated benefits of performing evaluations into cost savings (e.g., Bias & Mayhew, 2005). However, despite logical arguments to the contrary the subjective experience of many designers is that usability work lengthens projects, adds expenses, and fails to prevent that new problems show up when systems are released for actual use (Lund, 1997). As a result practitioners tend to prefer methods that are low-cost in terms of the time, expertise, and equipment required to apply them. Based on the argument that imperfect but low-cost evaluations are vastly superior to doing no evaluation work at all, Nielsen (1993) has made a case for discount evaluation methods. Recently, others have criticized these discount methods for being too poor (Cockton & Woolrych, 2002).

3 Severity assessments

Viewed from inside a usability evaluation severity assessments are the major device for influencing problem prioritization and thereby guiding the utilization of design resources. The severity of a usability problem is generally considered to be a combination of three factors (Nielsen, 1994):

- *Impact*: how much trouble will affected users experience?
- *Persistence*: how many times will a user experience the problem?
- *Frequency*: how many users will be affected by the problem?

These factors can be measured individually but they are frequently collapsed into a single severity rating. This can, for example, be done by adding or multiplying the ratings of individual factors (Lewis, 2006). Often, rating of individual factors is bypassed and evaluators simply make a single rating of severity on a scale such as: (1) not a usability problem, (2) minor usability problem, (3) major usability problem, and (4) usability disaster. In defining these categories many authors make reference to how soon the problem should be fixed and, thereby, to the design process of which the evaluation is part (e.g., "This grade [i.e., usability disasters] is for those few problems that are so serious that the user is better served by a delay in the delivery of the system", Hornbæk & Frøkjær, 2005).

Studies have consistently found weak correlations among the severity ratings of experienced evaluators (Table I). In three studies (Catani & Biers, 1998; Nielsen, 1994; Virzi, 1992) the severity ratings were purely judgement-driven; in the two other studies (Jacobsen, Hertzum, & John, 1998; Lesaigle & Biers, 2000) the ratings were, partly, data-driven in that the evaluators had seen videos of users thinking out loud while using the system. While Kendall's coefficient of concordance indicates a better-than-chance agreement among evaluators, the agreement is so low that it is risky to rely on the severity ratings of any one evaluator. Lesaigle and Biers (2000) even find that only 26 of the 78 correlations between pairs of evaluators were significant. The low agreement provides a basis for challenging severity assessments and for supplementing them with considerations about, for example, the ease of fixing problems. Indeed, the evaluators in Rodden, Green, and Kanis (1999) report that they refrained from reporting some problems because they could not think of a solution for these problems.

INSERT TABLE I ABOUT HERE

Usability evaluation methods, especially the widely used discount methods, are not meant to be perfect but their merit rests on the assumption that their performance degrades gracefully as evaluations become less costly and more informal. An important issue in this connection is whether missed problems tend to be of low severity whereas the high-severity problems are detected even when evaluations are conducted with a small number of users. Considerable debate has ensued about whether problem severity correlates with the number of users encountering a problem; that is, frequency. Table II shows that whereas some studies find highly significant,

moderate correlations between problem severity and frequency of occurrence, other studies find negligible or even negative correlations. To some extent early studies were optimistic whereas later studies call for more caution. It should be noted that except in the study by Nielsen (1994) the evaluators made data-driven severity ratings. While this may improve the robustness of the ratings it also means that the evaluators had information about frequency of occurrence when they rated severity, a potential confound. Virzi (1992) explicitly designed his third experiment to address this issue but unfortunately did not report the correlation.

INSERT TABLE II ABOUT HERE

One candidate explanation of the differences in severity-frequency correlation is that the severity ratings in several of the studies are made by only one or two evaluators and may, therefore, be less reliable because the agreement among evaluators' severity ratings is generally low (Table I). This explanation suggests that the low correlations in Table II may be the less reliable. Two other candidate explanations of the differences in severity-frequency correlation are explored below.

Heterogeneous subgroups of users. Most systems have multiple subgroups of users rather than one homogeneous user group. Caulton (2001) has pointed out that whereas some problems may be encountered with equal probability across subgroups, other problems are unique to specific subgroups. The 19 users in Law and Hvannberg (2004a) are a good example of heterogeneous subgroups as they are from different countries, speak different languages, hold different job positions, and have different levels of competence in information technology. Thus, a severe problem unique to one subgroup will be encountered by only a small part of the users, because most of the users will belong to other subgroups not affected by the problem. Assuming that some problems are unique to specific subgroups, a weak overall correlation between severity and frequency may mask that severity and frequency are more strongly correlated within individual subgroups (Caulton, 2001). In Virzi (1992, experiment 2) all 20 users were participants of an undergraduate psychology course at a private university and they were only included in the experiment if they reported little or no computer experience and no experience with electronic calendars. This suggests a homogeneous user group and, thereby, little risk that a correlation between severity and frequency is masked by subgroup effects. Indeed, Virzi (1992, experiment 2) finds a moderate severity-frequency correlation. This, partly, reconciles the differences in level of correlation among the studies in Table II.

Coverage of system features. The complexity of most systems precludes evaluations that cover the full system functionality. Instead test tasks exercise a subset of system features selected in the process of defining the purpose of the evaluation. Prior to an evaluation, designers will likely focus their attention on the features selected for evaluation, leading to improvement of the core parts of these features relative to the rest of the system. During the evaluation users will work on the test tasks and thereby also focus on the features selected for evaluation. Thus, many users will exercise the core parts of these features and any problems are likely to occur with high frequency. Conversely, exercise of peripheral parts of the features and excursions into other parts of the system will occur more randomly and by single users. The designers have, however, not devoted the same amount of attention to these parts and they are thus more likely to contain severe problems. This explanation suggests moderate correlations between severity and frequency for complex systems but also assumes a design context that is absent in the studies in Table II.

INSERT TABLE III ABOUT HERE

Two additional studies provide data about correlations between severity-related variables, see Table III. Hassenzahl (2000, experiment 1) finds no correlation between severity assessments and the time users spent recovering from problems. The correlations for the individual evaluators ranged from -0.39 to 0.46 and none of them were significant. Whereas actual problem-recovery time was not correlated with severity, evaluators' assessment of problem-recovery time was higher for the problems they assessed as the more severe (Hassenzahl, 2000, experiment 2). Overall, Hassenzahl (2000) finds a lack of correspondence between data-driven and judgement-driven estimates of severity but correspondence between some judgement-driven severity estimates.

John and Marks (1997) provide data about the correlation between the number of users that experienced a problem during a thinking-aloud study and whether or not the problem report persuaded the developer to address the problem by changing the code. The significant but weak correlation suggests that while frequency may explain r^2 = 9% of the variability in the developer's decision about whether to address a problem, this

decision is mainly based on other considerations. Apart from this one study, this author is not aware of studies correlating severity or severity-related variables with designers' decisions about whether to address a problem.

4 Field study of the impact of evaluation on design

Severity assessments are but one factor in determining designers' response to usability evaluation. The following field study, based on Hertzum (1999), investigates the factors that played a key role in determining the impact of the evaluations in an iterative design process. These competing factors may overshadow severity assessments and must therefore be understood and managed to make effective use of severity assessments.

4.1 Introduction

The field study concerns a fourteen-month project that consisted in the development of a graphical user interface for the Filing & Notification (F&N) system, which contains information about citizens for use by municipal authorities. The F&N system was developed in the 1970s as a mainframe application and is used daily by several thousand employees in the Danish municipalities. However, efficient use of its character-based user interface requires dedicated training, regular use, and an extensive printed manual. The purpose of the F&N project was to make a Windows version of the F&N system. To minimize costs and preserve the investment in the existing mainframe application it was decided to implement the new version as a graphical front end on top of the mainframe application. The F&N project group consisted of a project manager, a primary designer, a secondary designer, and an online-help writer. The primary designer was identical to the present author who was at the time employed in the organization where the study took place. The F&N project group adopted an iterative design process that evolved around five usability evaluations. This assigns evaluation a guiding role and means that this field study concerns a design process receptive to input from usability evaluations.

4.2 Method

The data collected from the project were the reports from the usability evaluations and a diary that covered the activities of the primary designer.

To indicate the extent to which the problems encountered during the usability evaluations were solved, each problem was assigned a status: (1) *Solved*, the problem was fixed. (2) *Reduced*, the problem was partly, but not fully, fixed. (3) *Unaddressed*, the problem was either deferred or rejected. The primary designer and either a usability specialist or the project manager made this assessment of problem status during the project to maintain an overview of the progress made. Subsequently these assessments have been revisited and 12% of the problems have been reassigned to indicate that they have been addressed to a lesser extent.

The activities of the primary designer, who did most of the work on the project, were tracked in a diary. The diary, which was updated successively throughout the day, covered the nine-month period from the first usability evaluation through the fifth and contained every activity with a duration of 15 minutes or more. The recordings were made on diary sheets, one for each day, and gave the starting and ending time of the activity, the project to which the activity pertained, and a terse description of the activity. To achieve this level of detail the current diary sheet was lying easily accessible on the designer's desk.

To enable investigations of when a problem was addressed and how long it took to correct it, each activity recorded in the diary was analysed and linked to the problems it addressed. Some activities were not performed in response to any of the problems encountered during the evaluations; other activities contributed to the solution of several problems. Thus, an activity could be linked to any number of problems, just as a problem could be linked to any number of activities.

4.3 The five usability evaluations

The F&N project evolved around five usability evaluations, each a major project milestone (see Table IV). The users participating in the evaluations were regular users of the mainframe version of the F&N system.

INSERT TABLE IV ABOUT HERE

Laboratory test. The first evaluation was conducted by two usability specialists in an in-house usability laboratory and involved six users who were asked to think out loud while solving eleven set tasks. Each user worked with the tasks for 1½ hours and during that period they were alone in the test room, while the usability specialists were in a control room separated from the test room by a one-way mirror. The users were frequently

asked questions such as "What would you expect to see or be able to do at this point?" The usability specialists recorded problems observed during the sessions and communicated them to the F&N project group in a report.

Workshop test. The second evaluation was also performed in-house but by the F&N project group and in a conference room rather than a usability laboratory. First, the project group gave a guided tour of the F&N system. Then the users had two two-hour sessions for testing, separated by lunch. Finally, the evaluation was concluded by a plenary discussion. Eight users participated in the evaluation and they worked two by two on a number of set tasks. Each pair of users sat at a separate table with one computer. When users discovered a problem they either called upon a designer to report it directly or made a print-out of the screen and annotated it. The designers circled among the users to observe, inquire, and receive feedback. After the test the project group produced an annotated list of the encountered problems.

Field tests. The third, fourth, and fifth evaluation were performed on site and managed by the users themselves. The users had the F&N system installed at their personal workplace and used it occasionally in the execution of their day-to-day duties. There were no set tasks to be solved during the field tests; rather the F&N system was exposed to real-life conditions in terms of tasks, workload, and technical environment. The designers contacted the users once or twice during an evaluation to inquire about their use of the system, get feedback, and motivate further testing. Problems discovered by the users were reported by telephone or on an evaluation form. The project group concluded each evaluation by compiling an annotated list of the reported problems.

4.4 Discussion

The five usability evaluations led to the detection of a total of 77 problems, of which 55 were solved or reduced while 22 were left unaddressed. Counting reduced problems as 50% solved, this gives an overall impact ratio (Sawyer et al., 1996) of 65%. Looking at the impact ratio of individual evaluations (Table IV) it is striking that only the three first evaluations had an impact while half of the unaddressed problems were encountered during field test 3. A key factor in determining whether a problem was addressed was *when* it was found. Finding a problem early profoundly increased its chances of being addressed. Conversely, the two last field tests, conducted during the last third of the project, had no impact. Field tests in which users are expected to work with a system while unsupervised can only be conducted late in a project when the system prototype is fairly stable. Therefore, the F&N study suggests that such field tests will often have a low impact. Among the reasons for a low impact of late evaluations Bass and John (2003) emphasize that many usability issues have consequences for the software architecture, which gets more and more ingrained and costly to change as projects progress. Additional reasons are discussed below.

In the beginning and middle of the F&N project much work remained and the project would be one of the designers' major concerns for some time to come. Also, many problems could be corrected at almost no extra cost when they could be addressed along with other problems concerning the same part of the design. Near the end of the project most of the designers spent the majority of their time on other projects or they were about to enter other projects, and little room was left for prolonging the F&N project even moderately. This meant that relative to the designers' other responsibilities the time required to solve a usability problem tended to appear reasonable in the beginning and middle of the project and prohibitive near the end of the project (see also, Kumar, 1990). Table V shows that during the five months from the laboratory test to field test 1 the primary designer spent an average of more than five hours a working day on the F&N project. During the remaining four months of the F&N project, other projects occupied most of the primary designer's time, and substantially fewer daily hours were left for the F&N project. A major reason for this decrease in project intensity was the duration of the field tests. During the field tests the list of outstanding tasks contained few, if any, high-priority tasks and for weeks the project was largely on hold. To deal with such project blocking the F&N designers were assigned to multiple projects simultaneously. As a consequence the prime determinants of the impact of an evaluation may be external to both the evaluation itself and other project activities and, instead, concern the activities and deadlines of the other projects on which the designer was working.

INSERT TABLE V ABOUT HERE

Table V also shows that the work involved in addressing the usability problems accounted for a substantial fraction of the total design effort. The primary designer spent 25% of his time fixing problems encountered during the five evaluations. At the time of the laboratory test several facilities were not yet developed. Thus, the problems found during this evaluation were added to an already long list of outstanding tasks. As the project progressed the list got shorter and increasingly dominated by input from the evaluations. This meant that usability issues came to occupy more of the primary designer's time. Near the end of the project the action taken in response to the evaluations was restricted to presumably severe problems and the amount of time spent

on usability issues dropped. The time spent fixing problems corroborates the impact ratios and suggests that the major cost of usability evaluation is not finding but fixing the problems. Averaged over the entire project the primary designer spent 2.8 hours fixing a problem. Field test 1 was administered by the primary designer and this task amounted to 0.4 hours for each problem found – that is, one seventh of the average time spent fixing a problem. Field tests owe their low cost to leaving almost everything to the users, and the essential task left with the persons conducting field tests is the management of users' commitment to perform a thorough evaluation. The administration of the laboratory and workshop tests was more resource demanding than the field tests.

In the laboratory test, the users digressed from the ideal way of solving the set tasks – they got into problems and recovered from these problems – but they kept pursuing the tasks and did so with little attention to their ecological validity. In the workshop test the users seemed to feel free to go beyond the set tasks to exercise the system in ways consistent with their actual tasks but not anticipated by the designers. It seems reasonable to ascribe the users' more exploratory attitude to two circumstances. First, working two by two the users were not alone when they got stuck or in doubt, and differences in their day-to-day work practices fostered discussion and divergent suggestions for solving the tasks (see also Hackman & Biers, 1992). Second, the face-to-face way of communicating with the persons conducting the evaluation, the lunch break, and the absence of detailed observation of the users' behaviour made the atmosphere more informal and the set tasks less prominent. This had a substantial effect on the types of problems detected during the two evaluations.

5 Experimental study of severity assessment in usability inspections

Whereas the preceding section investigated how factors beyond severity assessments influence the impact of evaluations on design, the following experimental study looks at the aspects of severity that enter into severity assessments. Unless severity assessments can stand up to competing factors in a persuasive way, severity assessments are wasted effort. The experimental study investigates correlations between different aspects of severity and – pertaining to the need for persuasiveness – illustrates a way of presenting severity information that emphasizes the comparatively large effect of addressing the most severe problems. The experimental data were collected by Hertzum, Jacobsen, and Molich (2002) who focused their analysis on problem detection; the severity data have not previously been analysed.

5.1 Method

5.1.1 Evaluators and task scenario

Twelve professional usability specialists participated in the study as evaluators. One evaluator was, however, removed during the data analysis because he failed to comply with the procedures of the study. On average, the remaining 11 evaluators had 7.3 years of experience with usability work and had conducted 37 usability tests (with users) and 35 usability inspections (without users).

The evaluators inspected a comprehensive e-commerce web site, www.avis.com, that enables people to rent cars all over the world. The evaluators were asked to focus their inspections on five user tasks: finding the cost of renting a car, making a reservation, getting an overview of the kinds of cars available, finding out about pick-up locations, and getting information about special deals. It should be noted that the web site has been changed since the evaluation.

5.1.2 Procedure

The evaluators individually inspected the web site and documented their inspection in a written report listing the identified usability problems. In addition to a brief description of each problem, the evaluators were asked to assess its impact and persistence. *Impact* was assessed by indicating the percentage of users who would experience: (1) no problem, (2) a minor problem, that is a brief delay, (3) a serious problem, that is a significant delay but users eventually complete their task, and (4) a disaster, that is users voice strong irritation, are unable to solve the task, or solve it incorrectly. *Persistence* was indicated by a rating on a three-point scale: (1) Users quickly learn to get around the problem. (2) Users only learn to get around the problem after encountering it several times. (3) Users never learn how to get around the problem. The evaluators were also asked to report positive aspects of the web site. As long as the evaluators complied with this format they were free to inspect the web site in any way they wanted. The evaluators spent an average of 15 hours on their individual inspection and they all performed some type of heuristic evaluation (Nielsen & Molich, 1990).

After performing their individual inspection the evaluators met for two hours in three-person groups to combine their individual inspections into group outputs. In continuation of the group discussions, all evaluators participated in a two-hour plenary discussion of their experiences from the evaluation. Finally, the evaluators individually filled in a post-evaluation questionnaire about their personal familiarity with cars and car rental.

5.1.3 Severity indicators

The data analysis involves three indicators of the severity of each reported problem: impact, persistence, and frequency. Whereas the evaluators' *persistence* ratings are used directly, some pre-processing is necessary to produce a single indicator of *impact*. On the basis of the impact percentages reported by the evaluators (e.g., no problem: 10%, minor: 20%, serious: 30%, and disaster: 40%) the impact of each problem is defined as:

$$\text{Impact} = \frac{0\times\text{NoProblem} + 1\times\text{Minor} + 2\times\text{Serious} + 3\times\text{Disaster}}{3} \quad (1)$$

Impact (in the example: $(0\times0.10 + 1\times0.20 + 2\times0.30 + 3\times0.40) / 3 = 0.67$) can be any value between 0 and 1, with higher values indicating higher impact. Hassenzahl (2000) uses a similar approach to define a severity index from categorical judgements.

To determine the *frequency* with which problems were reported the 11 evaluators' individual reports were compiled into a master list of unique problems. This was done independently by two raters. They classified 68% of the problem instances identically; disagreements were resolved through discussion and a consensus was reached. Most of the raters' disagreements were resolved by combining problem instances into fewer, more frequently reported problems. The master list contains 220 unique problems.

5.2 Results

Each evaluator assessed the impact and persistence of the problems he or she reported. The mean impact assigned by evaluators to the problems they reported ranges from 0.20 to 0.53 with an overall mean of 0.30. For persistence, the evaluators rated an average of 43% (*SD* = 21) of the problems they reported as problems users quickly learn to get around, 36% (*SD* = 14) as problems users only learn to get around after encountering them several times, and 21% (*SD* = 12) as problems users never learn to get around. Differences in impact and persistence values across evaluators may be due to the different sets of problems reported by different evaluators. The frequency with which problems were reported (i.e., the number of evaluators reporting a problem) can, however, be used to relate the impact and persistence assessments of individual evaluators to the collective outcome of their evaluations.

For each individual evaluator, the Pearson correlation was calculated between the evaluator's impact ratings of the problems he or she reported and the frequencies with which these problems were reported. Pearson correlations were used because the data satisfy parametric assumptions. The correlations between impact and persistence and between frequency and persistence were calculated by Spearman's Rho correlation, because the persistence ratings form an ordinal scale. Table VI shows the correlations between impact, persistence, and frequency. There is a moderate and for seven of the 11 evaluators highly significant correlation between impact and persistence. The correlations vary substantially from evaluator to evaluator but, on average, the correlation between impact and persistence is 0.48. Thus, $r^2 = 23\%$ of the variability in impact can be predicted from the persistence ratings, and vice versa. The correlations between impact and frequency are significant for five of the evaluators and for these evaluators the correlations are moderate. For the remaining evaluators the correlation between impact and frequency is considerably weaker, and across all evaluators frequency explains an average of only $r^2 = 8\%$ of the variability in impact. The correlation between persistence and frequency is negligible. Frequency explains an average of only $r^2 = 2\%$ of the variability in persistence, and only one of the 11 evaluators has a significant correlation between persistence and frequency.

INSERT TABLE VI ABOUT HERE

To illustrate how severity assessments may be used to guide design efforts, problem severity was defined as impact times normalized persistence. Table VII shows the sum of these severity scores for each evaluator and the number of problems that must be fixed to accomplish the first three reductions of the severity sum to half of its previous value, assuming problems are fixed in order of decreasing severity. For example, for evaluator A the seven most severe problems contain half of the severity sum, the six next problems contain a quarter of the severity sum, the next eight problems contain an eighth of the severity sum, and the 24 least severe problems contain only the last eighth of the severity sum. On average, half of the severity sum is contained in the first 22% of the problems. Though the number of problems reported by each evaluator varies from 17 to 52, the first half of the severity sum is contained in the relatively narrow range of 16-30% of the problems. Further, each of the first three recursive reductions of the severity sum to half of its previous value can be achieved by fixing an average of, roughly, the next 20% of the problems.

INSERT TABLE VII ABOUT HERE

5.3 Discussion

To examine the nature of severity, the evaluators were asked to make separate assessments of the impact and persistence of each reported problem. However, the correlation between impact and persistence suggests either that these two aspects of severity tend to co-vary or that the evaluators have difficulties distinguishing between them, at least for some problems. During the plenary discussion some of the evaluators expressed that they felt that without, for example, observing users they did not have sufficient input to assess severity at this level of detail, and several evaluators seemed to doubt whether the distinction between impact and persistence would make their reports more informative to designers. This suggests that one combined assessment of the severity of each problem may be more feasible, although it masks differences between aspects of severity.

The weak correlations of impact and persistence with frequency reiterate that more than one evaluator is needed to find most of the severe problems in a system by means of heuristic evaluation (see also, Hertzum et al., 2002). Individual evaluators appear insufficiently capable of imagining the diversity of people that may use a web site for car rental. Rather, differences in the evaluators' personal experiences with cars and car rental may have influenced their inspections. For example, in the post-evaluation questionnaire three evaluators answered that they did not have a driver's licence, and when asked about the number of car brands and types they could distinguish the evaluators' answers ranged from 'a few' (three evaluators) through 'some' (two evaluators) to 'a lot' (six evaluators). This indicates that the evaluators themselves belong to different subgroups of the user population, and their reports contain no attempts to counter this potential bias by explicitly considering different groups of users of the web site. It is, thus, likely that the severity assessments have been biased toward a subgroup of users that is too similar to the individual evaluator to be representative of the total user population, a phenomenon known as anchoring (Tversky & Kahnemann, 1974). Additional support for such anchoring is provided by Law and Hvannberg (2004b) who report that only a subset of even the severe problems detected during heuristic evaluation can be ascribed to the heuristics; the remaining problems are identified based on the evaluators' experience and intuition. Further, differences in the sets of problems detected by different evaluators may affect their severity assessments because evaluators may, to some extent, assess the severity of one problem relative to the severity of other problems. Thus, the reliability of individual evaluators' severity assessments is likely to be low and designers may, therefore, regard evaluator-made severity assessments as opinion. This equates severity assessments with designers' own specialist opinions and suggests that improved impact could be achieved if severity assessments were either derived from user data or made by users rather than evaluators.

If half of the problems detected during a usability evaluation are subsequently addressed then Sawyer et al. (1996) say that the impact of the evaluation has been 50% as their formula for the impact of an evaluation is simply the percentage of addressed problems out of the total set of problems detected during the evaluation. This amounts to presuming that all problems are equally severe and thus provides no support for problem prioritization. This study (Table VII) proposes the recursive halving of the severity sum as an aid in determining the effect of addressing problems in order of decreasing severity. Many definitions of severity partly distinguish between different levels of severity by means of different recommendations about how soon problems should be fixed. In doing so these definitions recommend that problems are fixed in order of decreasing severity. This provides for a focused use of resources, especially in iterative design, but the size of the effect of fixing the smaller number of more severe problems is easily underestimated. In this study half of the severity sum is concentrated in approximately 20% of the problems and the severity sum can be halved three times (i.e., reduced to 12.5%) by solving 60% of the problems. Whiteside et al. (1988) note that once detected the most severe problems are often trivial to fix. Thus, the work involved in fixing the 20% most severe problems may be comparable to that of fixing any other subset of 20% of the problems.

6 Concluding discussion

The severity of a usability problem is an assessment of the amount of trouble and inconvenience users will experience as a result of a specific aspect of a system. In this sense severity assessments bypass design-process considerations. Severity assessments are, however, also recommendations about the urgency of fixing problems, and in this sense they are explicitly intended to influence designers' allocation of their time and attention. Experimental usability-evaluation studies such as the one in Section 5 focus on how severity assessments are made and rarely discuss the impact such assessments have on the subsequent design process. The F&N project identifies three important factors that compete with severity assessments in determining designers' response to usability evaluations: the design-process stage, the level of formality, and cost-

effectiveness. To be effective, severity assessments must take these competing factors into account.

Design-process stages and the role of severity assessments. In the F&N project the early evaluations had substantial impact, the late field tests a considerably lower impact. This was mainly due to the time the designers had available for the F&N project in its early and middle stages and their responsibilities toward other projects in the late stages of the F&N project, and it was only to a smaller extent a consequence of properties of the evaluation methods. With respect to severity assessments this suggests that their role changes substantially in the course of a project, a finding that has remained under-recognized in previous work. This study leads to the following recommendations. In the early stages of iterative projects, the primary scope of severity assessments should be the period until the next evaluation. Fixing the severe problems before the next evaluation will substantially improve the insights that can be gained from it because severe problems consume considerable evaluation time and user resources and thereby render other, more subtle problems invisible. In the middle of a project, evaluations still bring up numerous issues and, at the same time, many facilities have yet to be designed. At this stage, the primary contribution of severity assessments should be to help ensure a consistent focus on usability in the midst of a design process encompassing manifold outstanding issues. In the late stages of a project, focus is increasingly on project completion and the primary role of severity assessments is to point out the few, high-severity problems that must necessarily be addressed. Evaluators should put their resources into providing persuasive arguments for fixing these problems, whereas problems of middle and low severity are unlikely to be addressed. Also, evaluators may inadvertently assess problem severity relative to the different sets of problems identified in different system prototypes. This provides additional reasons for distinguishing between severity assessments made at different stages of the design process.

Level of formality and user-made severity assessments. The laboratory and workshop tests both triggered many changes of the F&N system but they differed substantially in their level of formality. This may reduce the reliability (i.e., reproducibility) of the workshop test but it improves its applicability, especially during the early stages of design because the outcome of the workshop test was to a large extent to adjust what the system could do, whereas the laboratory test triggered a larger proportion of changes related to how the functionality was made available to users. The more informal approach of the workshop test is advocated by Wright and Monk (1991) and represents a seemingly viable blending of conventional usability evaluation and participatory design. A hitherto unexplored possibility is to involve users in assessing problem severity. This could be done at almost no extra cost during the concluding plenary discussion of a workshop test and would provide a structured way of collaboratively reviewing the problems reported by individual pairs of users. One aim of having users, rather than usability specialists, assess severity would be to improve the validity and persuasiveness of severity assessments. User-made severity assessments at the end of a workshop test would, however, also mean that multiple people assess the severity of the reported problems, possibly improving the reliability of the severity assessments relative to those made by individual usability specialists. The low level of agreement in the usability specialists' severity assessments during the experimental study indicates a large need for finding ways of obtaining more reliable severity assessments.

Cost-effectiveness and impact enhancement. The field tests in the F&N project were low-cost and led to the detection of multiple problems that had not been identified during the earlier evaluations. Smilowitz et al. (1994) find that field tests (they use the term beta tests) are as effective as laboratory tests and considerably less costly, but their study is restricted to whether problems are found and excludes considerations of whether they are fixed. In the F&N project a lot of the resources that went into conducting field tests were wasted. Since field tests in which users work unsupervised with a system prototype are restricted to the late stages of the design process the lower impact of such field tests may not be specific to the F&N project. A further aspect of the cost-effectiveness of usability evaluations is that their total cost includes both the cost of finding problems and of addressing them. This may have a positive and a negative consequence. First, if the total cost of evaluations is dominated by the cost of addressing the problems, as the F&N project suggests, the amount of resources that goes into the conduct of an evaluation becomes less critical. Second, the cost of fixing a problem may overshadow assessments of its severity. There appears to be a general need for extending usability evaluation methods with what could be termed means of impact enhancement. Such means are intended to increase the persuasiveness of evaluations and thereby help ensure that resources spent finding problems are not wasted, such as in the project described by Marshall, McManus, and Prail (1990) in which several usability evaluations were conducted but no action was taken to fix any of the identified problems before the system was released. As problem prioritization is inevitable, severity assessments are a powerful but seemingly underutilized element of impact enhancement. If severity is rated on, or transformed into, an interval scale this study suggests the halving of the severity sum as a means of calling designers' attention to the severe problems by visualizing the large payoff of fixing a high-severity problem and, conversely, the modest potential of spending resources on low-severity problems.

Using severity assessments to stipulate the order in which problems should be fixed presupposes reliable and valid severity assessments. Based on data about severity assessments made during usability inspections this

study yields three recommendations. First, the agreement among individual evaluators' severity assessments is so low that it is risky to use a single evaluator's severity assessments as the basis for decisions about system revisions. The extra cost of involving multiple evaluators will likely be small compared to both the increased reliability of the severity assessments and the cost of fixing the severe problems. User-made severity assessments can also be considered. Second, if subgroups are carefully handled frequency appears to provide some severity information. Thus, subgroups should be a primary consideration in the planning and analysis of evaluations, particularly if frequency data will be used to support severity assessments or substitute for having multiple evaluators assess severity. For usability evaluations with users, subgroups should be considered in the selection of users; for usability inspections, subgroups should be considered in the selection of evaluators. Third, while distinctions between different aspects of severity may be useful in establishing the severity of problems encountered by observing users, it appears that evaluators in usability inspections lack a basis for distinguishing between aspects such as impact and persistence and thus consider them to be correlated. This points toward restricting severity assessments to a single, combined rating.

Acknowledgements

Rolf Molich was the initiator and organizer of the CUE-3 study, which provides the empirical data for Section 5. Further information about the Comparative Usability Evaluation (CUE) studies is available at www.dialogdesign.dk/cue.html. Niels Jacobsen was instrumental in the planning and execution of the CUE-3 study. Special thanks are due to the evaluators who participated in the CUE-3 study and the members of the F&N project group.

References

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.

Bass, L., & John, B.E. (2003). Linking usability to software architecture patterns through general scenarios. *Journal of Systems and Software*, 66(3), 187-197.

Bias, R.G., & Mayhew, D.J. (Eds.). (2005). *Cost-justifying usability: An update for the Internet age*. San Francisco, CA: Morgan Kaufmann.

Boren, T., & Ramey, J. (2000). Thinking aloud: Reconciling theory and practice. *IEEE Transactions on Professional Communication*, 43(3), 261-278.

Brooks, P. (1994). Adding value to usability testing. In J. Nielsen & R.L. Mack (Eds.), *Usability inspection methods* (pp. 255-271). New York: Wiley.

Catani, M.B., & Biers, D.W. (1998). Usability evaluation and prototype fidelity: Users and usability professionals. In *Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting* (pp. 1331 - 1335). Santa Monica, CA: HFES.

Caulton, D.A. (2001). Relaxing the homogeneity assumption in usability testing. *Behaviour & Information Technology*, 20(1), 1-7.

Cockton, G., Lavery, D., & Woolrych, A. (2003). Inspection-based evaluations. In J. Jacko & A. Sears (Eds.), *The human-computer interaction handbook* (pp. 1118-1138). Mahwah, NJ: Erlbaum.

Cockton, G., & Woolrych, A. (2002). Sale must end: Should discount methods be cleared off HCI's shelves? *ACM Interactions*, 9(5), 13-18.

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

Dumas, J.S. (2003). User-based evaluations. In J. Jacko & A. Sears (Eds.), *The human-computer interaction handbook* (pp. 1093-1117). Mahwah, NJ: Erlbaum.

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

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

Gould, J.D., Boies, S.J., & Lewis, C. (1991). Making usable, useful, productivity-enhancing computer applications. *Communications of the ACM*, 34(1), 74-85.

Hackman, G.S., & Biers, D.W. (1992). Team usability testing: Are two heads better than one? In *Proceedings of the Human Factors Society 36th Annual Meeting* (pp. 1205-1209). Santa Monica, CA: HFS.

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

Hertzum, M. (1999). User testing in industry: A case study of laboratory, workshop, and field tests. In A. Kobsa & C. Stephanidis (Eds.), *User Interfaces for All: Proceedings of the 5th ERCIM Workshop* (pp. 59-72). Sankt Augustin, DE: GMD.

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

Hertzum, M., Jacobsen, N.E., & Molich, R. (2002). Usability inspections by groups of specialists: Perceived agreement in spite of disparate observations. In *Proceedings of the CHI 2002 Conference on Human Factors in Computing Systems: Extended Abstracts* (pp. 662-663). New York: ACM.

Hornbæk, K., & Frøkjær, E. (2005). Comparing usability problems and redesign proposals as input to practical systems development. In *Proceedings of the CHI 2005 Conference on Human Factors in Computing Systems* (pp. 391-400). New York: ACM.

Jacobsen, N.E., Hertzum, M., & John, B.E. (1998). 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.

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

Kjeldskov, J., & Stage, J. (2004). New techniques for usability evaluation of mobile systems. *International Journal of Human-Computer Studies*, 60(5&6), 599-620.

Kumar, K. (1990). Post implementation evaluation of computer-based information systems: Current practices. *Communications of the ACM*, 33(2), 203-212.

Law, E.L.-C., & Hvannberg, E.T. (2004a). 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.

Law, E.L.-C., & Hvannberg, E.T. (2004b). Analysis of strategies for improving and estimating the effectiveness of heuristic evaluation. In *NordiCHI 2004: Proceedings of the Third Nordic Conference on Human-Computer Interaction* (pp. 241-250). New York: ACM.

Lesaigle, E.M., & Biers, D.W. (2000). Effect of type of information on real time usability evaluation: Implications for remote usability testing. In *Proceedings of the XIVth Triennial Congress of the International Ergonomics Association and 44th Annual Meeting of the Human Factors and Ergonomics Society* (pp. 6-585 - 6-588). Santa Monica, CA: HFES.

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

Lewis, J.R. (2006). Usability testing. In G. Salvendy (Ed.), *Handbook of human factors and ergonomics, 3rd edition* (pp. 1275-1316). New York: Wiley.

Lund, A.M. (1997). Another approach to justifying the cost of usability. *ACM Interactions*, 4(3), 48-56.

Marshall, C., McManus, B., & Prail, A. (1990). Usability of product X – Lessons from a real project. *Behaviour & Information Technology*, 9(3), 243-253.

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

Nielsen, J. (1994). Heuristic evaluation. In J. Nielsen & R.L. Mack (Eds.), *Usability inspection methods* (pp. 25-62). New York: Wiley.

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.

Pejtersen, A.M., & Rasmussen, J. (1997). Effectiveness testing of complex systems. In G. Salvendy (Ed.), *Handbook of human factors and ergonomics, 2nd edition* (pp. 1514-1542). New York: Wiley.

Rodden, M.J., Green, W.S., & Kanis, H. (1999). Difficulties in usage of a coffeemaker predicted on the basis of design models. In *Proceedings of the Human Factors and Ergonomics Society 43rd Annual Meeting* (pp. 476-480). Santa Monica, CA: HFES.

Sawyer, P., Flanders, A., & Wixon, D. (1996). Making a difference - The impact of inspections. In *Proceedings of the CHI'96 Conference on Human Factors in Computing Systems* (pp. 376-382). New York: ACM.

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

Thomas, J.C., & Kellogg, W.A. (1989). Minimizing ecological gaps in interface design. *IEEE Software*, 6(1), 78-86.

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

Virzi, R.A. (1992). Refining the test phase of usability evaluation: How many subjects is enough? *Human Factors*, 34(4), 457-468.

Whiteside, J., Bennett, J., & Holtzblatt, K. (1988). Usability engineering: Our experience and evolution. In M. Helander (Ed.), *Handbook of human-computer interaction* (pp. 791-817). Amsterdam: Elsevier.

Woolrych, A., & Cockton, G. (2001). Why and when five test users aren't enough. In J. Vanderdonckt, A. Blandford, & A. Derycke (Eds.), *Proceedings of the IHM-HCI 2001 Conference, Vol. II* (pp. 105-108). Toulouse, FR: Cépadèus Éditions.

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

Zirkler, D., & Ballman, D.R. (1994). Usability testing in a competitive market: Lessons learned. *Behaviour & Information Technology*, 13(1&2), 191-197.

Table I. Evaluators' agreement on severity assessments. Agreement is given both as Kendall's coefficient of concordance (Kendall's W) for the group of evaluators and as the range of correlations between pairs of evaluators (for Kendall's W, significant agreements are marked with asterisks)

Reference	System	Evaluators	No. of problems	Severity scale	Kendall's W	Range of pair-wise correlations
Virzi, 1992 (experiment 3)	Voice response	6 usability and product experts	17	3 point	0.33 [a]	0.05 - 0.72
Nielsen, 1994	Integrating	11 usability specialists	40	5 point	0.31 ** [b]	4 negative, 51 positive
Jacobsen et al., 1998	Multimedia authoring	4 HCI researchers	93	2 point	0.43 *** [c]	0.10 - 0.55
Catani & Biers, 1998	Library search	5 usability professionals	99	5 point	0.06 [d]	0.26 - 0.49
Lesaigle & Biers, 2000	Library search	13 experienced usability professionals	71	5 point	0.18 [d]	-0.05 - 0.53

** $p < 0.01$, *** $p < 0.001$

[a] Mean correlation among pairs of evaluators (Kendall's W is not reported by Virzi)

[b] $\chi^2(df = 39, N = 11) = 132.3$.

[c] $\chi^2(df = 92, N = 4) = 156.7$.

[d] Significance not reported.

Table II. Correlation between assessment of problem severity and frequency of problem occurrence (significant correlations are marked with asterisks)

Reference	System	Type of evaluation	Evaluators	No. of problems	Severity scale	Severity-frequency correlation
Virzi, 1992 (experiment 2)	Electronic calendar	Thinking aloud (20 users)	3 human-factors students	40	7 point	0.46 ** [a]
Nielsen, 1994	Integrating	Thinking aloud (4 users)	11 usability specialists	40	5 point	0.46 **
Jacobsen et al., 1998	Multimedia authoring	Thinking aloud (4 users)	4 HCI researchers	93	2 point	0.46 *** [a]
Law & Hvannberg, 2004a	Educational contents sharing	Thinking aloud (19 users)	2 HCI researchers	88	3 point	0.10 [a]
Lewis, 1994	Office system	Working on set tasks (15 users)	2 usability specialists	145	4 point	0.06 [a]
Woolrych & Cockton, 2001	Powerpoint drawing editor	Thinking aloud (12 users)	1 HCI researcher	16	3 point	-0.29 [a, b]

** $p < 0.01$, *** $p < 0.001$

[a] In assessing severity the evaluators had information about frequency of occurrence

[b] Calculated on the basis of data reported in the paper

Table III. Additional correlations between severity-related variables (significant correlations are marked with asterisks)

Reference	System	Evaluators	No. of problems	Variable 1	Variable 2	Correlation
Hassenzahl, 2000	Management of funding	12 people with system experience	15	Error-recovery time (seconds)	Severity rating (3-point scale)	-0.06
John & Marks, 1997	Multimedia authoring	1 developer	54	Frequency (of 4 users)	Code change (yes, no)	0.30 * [a]

* $p < 0.05$

[a] Calculated on the basis of data reported in the paper

Table IV. Usability evaluations and their impact ratios

Evaluation	No. of users	Offset from project start	Duration	Evaluation conducted by	No. of problems	Impact ratio
Laboratory test	6	5 months	2 days	Usability specialists	38	74%
Workshop test	8	8 months	1 day	Design team	20	73%
Field test 1	8	10 months	3 weeks	Users	8	94%
Field test 2	8	12 months	2 weeks	Users	0	-
Field test 3	8	13 months	5 weeks	Users	11	0%
Total					77	65%

Note. Since the F&N system evolved from one evaluation to the next, the numbers of problems found during evaluations *cannot* be used for making direct comparisons between evaluations.

Table V. Time spent on the F&N project by the primary designer, in total and to address problems encountered during the evaluations

Period	Hours spent on project		Hours spent fixing problems	Percent of time spent fixing problems
	In total	Per working day		
Laboratory test to workshop test	279	6:29	56	20%
Workshop test to field test 1	216	4:48	70	32%
Field test 1 to field test 2	79	2:09	25	32%
Field test 2 to field test 3	4	0:40	0	0%
Field test 3 to system release	26	0:47	1	4%
Total	604	3:41	152	25%

Table VI. Correlations between impact, persistence, and frequency (significant correlations are marked with asterisks)

Evaluator	No. of problems	Correlations		
		impact-persistence [a]	impact-frequency [b]	persistence-frequency [a]
A	45	0.75 ***	0.15	0.22
C	17	0.34	0.52 *	0.44
D	52	0.59 ***	0.37 **	0.49 ***
E	19	0.60 **	0.37	0.24
F	39	0.71 ***	0.44 **	0.23
G	50	0.48 ***	0.23	0.13
H	21	0.15	0.58 **	0.01
I	25	0.65 ***	0.06	0.05
K	30	0.25	0.16	0.05
M	49	0.08	0.51 ***	-0.03
N	23	0.70 ***	-0.16	-0.12
Mean (*SD*)	33.64 (13.60)	0.48 (0.24)	0.29 (0.23)	0.15 (0.19)

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

[a] Spearman's Rho correlation

[b] Pearson correlation

Table VII. Number of problems that must be fixed to halve the remaining severity sum, assuming problems are fixed in order of decreasing severity

Evaluator	Severity sum	No. of problems	1st halving of severity sum		2nd halving of severity sum		3rd halving of severity sum		Rest of severity sum	
A	6.80	45	7	(16%)	6	(13%)	8	(18%)	24	(53%)
C	2.58	17	4	(24%)	3	(18%)	3	(18%)	7	(41%)
D	13.78	52	13	(25%)	12	(23%)	9	(17%)	18	(35%)
E	4.23	19	5	(26%)	3	(16%)	3	(16%)	8	(42%)
F	16.16	39	10	(26%)	11	(28%)	6	(15%)	12	(31%)
G	4.54	50	8	(16%)	11	(22%)	7	(14%)	24	(48%)
H	3.28	21	4	(19%)	7	(33%)	4	(19%)	6	(29%)
I	4.48	25	4	(16%)	4	(16%)	5	(20%)	12	(48%)
K	6.22	30	9	(30%)	7	(23%)	4	(13%)	10	(33%)
M	8.17	49	10	(20%)	9	(18%)	10	(20%)	20	(41%)
N	6.61	23	5	(22%)	4	(17%)	4	(17%)	10	(43%)
Mean	6.99	34	7.2	(22%)	7.0	(21%)	5.7	(17%)	13.7	(40%)

Note. *Severity sum* – the sum of problem impact times normalized persistence across all problems reported by the evaluator. *No. of problems* – number of problems reported by the evaluator. *1st halving of severity sum* – number (and percentage) of problems containing the first half of the severity sum. *2nd halving of severity sum* – number (and percentage) of problems containing half of the remaining half of the severity sum. *3rd halving of severity sum* – number (and percentage) of problems containing half of the remaining quarter of the severity sum. *Rest of severity sum* – number (and percentage) of problems containing the last eighth of the severity sum.