The usability inspection performance of work-domain experts: An empirical study

Asbjørn Følstad a,⁎, Bente C.D. Anda b, Dag I.K. Sjøberg b

a Department of Cooperative and Trusted Systems, SINTEF, Oslo, Norway
b Department of Informatics, University of Oslo, Norway

A R T I C L E   I N F O

Article history:
Received 31 January 2008
Received in revised form 2 August 2009
Accepted 1 September 2009
Available online 6 September 2009

Keywords:
Usability inspection
Work-domain expert
Validity
Thoroughness
Group-based expert walkthrough

A B S T R A C T

It is a challenge for usability experts to perform usability inspections of interactive systems that are tailored to work-domains of which these experts have little knowledge. To counter this, usability inspections with work-domain experts have been explored, but little empirical research has been reported on these experts' performance as evaluators. The present study compared the performance of work-domain experts and usability experts with respect to validity and thoroughness. The work-domain experts were characterized by high computer experience and low system experience. The usability experts were recruited from different ICT companies. The usability inspection method applied was group-based expert walkthrough; a method particularly developed to support non-usability experts as evaluators. The criterion for performance comparison was established through user tests. Fifteen work-domain experts and 12 usability experts participated in the study. The work-domain experts generated equally valid but less thorough usability inspection results than did the usability experts. This finding implies that work-domain experts may be used as evaluators in usability inspections without compromising validity. Moreover, the usability inspection performance of nominal groups of evaluators was explored. It was found that nominal groups of work-domain experts produced results of similar quality as did nominal groups of usability experts, given that group size is disregarded. This finding may be used as basis for hypotheses in future studies on the usability inspection performance of nominal groups of work-domain experts.

© 2009 Elsevier B.V. All rights reserved.

1. Introduction

Since the late 1980s, usability inspection, or analytical evaluation, has been promoted as a cost-effective alternative to empirical evaluation (such as user tests) for identifying usability problems in interactive systems (Cockton, Lavery, & Woolrych, 2003). Today, usability inspection methods are widely used; in particular heuristic evaluation, and to some degree cognitive walkthrough (Bark, Følstad, & Gulliksen, 2005; Venturi and Troost, 2004). These methods are typically applied by usability experts (Virzi, 1997).

The usability of an interactive system depends on the context of the system (ISO, 1998). Therefore, usability evaluations need to be conducted in consideration with the context (Bødker and Madsen, 1998). However, usability inspection methods such as heuristic evaluation and cognitive walkthrough do not include sufficient context knowledge resources, which represents a “major source of inspection risk” (Cockton, Lavery, & Woolrych, 2008, p. 1179). In consequence, it is a challenge for usability experts to perform usability inspections of interactive systems that are tailored to work-domains of which these experts have little knowledge. To acquire adequate knowledge of the work-domain context, usability experts may need to receive extensive introduction or training, or conduct context research. Such activities may represent a prohibitive overhead.

Motivated by the challenge of ensuring that sufficient knowledge of the system’s context is available, alternative approaches to usability inspections have been explored. One approach has been to involve work-domain experts as evaluators in the usability inspection. Work-domain experts are “(a) potential end users with direct experience from the work-domain or (b) persons with extensive secondary knowledge of the work-domain” (Følstad, 2007b, p. 218). Usability inspection methods particularly developed to include work-domain experts as evaluators have existed since the pluralistic walkthrough of Bias (1994), with the participatory heuristic walkthrough (Muller, 1998) and group-based expert walkthrough (Følstad, 2007a) as more recent additions.

When work-domain experts are employed as evaluators in usability inspections, required knowledge of the work-domain context is ensured. However, the inspection performance will be
affected if the usability intuitions of the work-domain experts are not sufficiently strong. In particular, unsound usability judgements may decrease the validity of the usability inspection. This is potentially critical, given the risk of false positives which seem to be associated with usability inspections (e.g. Cockton et al., 2008).

Little research has been conducted on the performance of work-domain experts' as evaluators in usability inspections, with the consequence that the human–computer interaction (HCI) community has little knowledge in this regard. This is particularly problematic with respect to industrial application of usability inspections involving work-domain experts as evaluators. This article reports on a study that compared the performance of work-domain experts and usability experts as evaluators in usability inspections conducted as group-based expert walkthroughs. The main performance measures of the study were the validity and thoroughness of the usability inspection results relative to the results of empirical user testing.

2. Previous work

2.1. Work-domain experts in human–computer interaction

The importance of involving work-domain experts in the development of interactive system is highlighted by a number of HCI researchers. In particular within the participatory design tradition, the value of mutual and reciprocal meetings of users, as experts on their work-domain, and developers has been emphasized (Muller, 2008). Kensing and Munk-Madsen (1993) argued that such involvement of the users is critical in order to “build up the knowledge required for developing and using a new system” (p. 78). To explain the value of the knowledge possessed by work-domain experts, Kensing and Munk-Madsen established six categories to describe knowledge required for adequate understanding between users and developers. Two of these knowledge categories included knowledge typically possessed by the work-domain experts alone: (1) concrete experience with the users' present work and (2) knowledge about relevant structures on this work.

To understand the significance of context knowledge possessed by work-domain experts in the development projects, it is useful to consider Nielsen's (1993) three dimensions of individual differences between users: users’ varying degree of (1) computer experience, (2) system experience, and (3) work-domain knowledge (or task-domain as it is termed by Nielsen). When developing interactive systems for users with higher levels of domain knowledge, it will be more critical to involve work-domain experts as knowledge resources.

Nielsen's three dimensions also help in discriminating between different categories of work-domain experts. Work-domain experts will by necessity be high on the dimension of work-domain knowledge, and, conversely, any user with high work-domain knowledge is reckoned a work-domain expert; this characteristic is what separates work-domain experts from other users. However, work-domain experts may vary on the dimensions of (a) specific system experience and (b) general computer experience. In consequence, it may be useful to classify work-domain experts into four categories given by high/low on these two dimensions.

It should be noted that work-domain experts may vary on a range of other dimensions as well. Nielsen (1993) suggests that important differences between users may be related to age, gender, differences in spatial memory and reasoning abilities, as well as preferred learning style. Recent research by Ling and Salvendi (2009) indicates that the evaluators' cognitive style may affect usability inspection performance. On this background, a model of work-domain experts based on differences in system experience and computer experience alone may seem simplistic. Even so, the model does seem to represent a good starting point for making relevant classifications while keeping with a modest level of complexity.

In the theory of usability evaluation, over the last two decades, there has been a movement toward treating the users as work-domain experts rather than subjects just to be observed. With basis in the participatory design tradition, Ross, Ramage, and Rogers (1995) introduced PETRA (participatory evaluation through redesign and analysis) in which users are regarded as experts that may suggest design changes in interwoven design and evaluation sessions. In the field of CSCW (computer supported cooperative work), Twidale, Randall, and Bentley (1994) presented informal evaluation sessions where the participants were used as reflecting experts, in order to capitalize on their context knowledge.

Other researchers have argued that users may be utilized as work-domain experts in user testing. Boren and Ramey (2000) argued that the participating user should be “cast as an important contributor, work-domain expert, or valued customer” (p. 268), an approach that was later empirically investigated by Krahmer and Ummeln (2004). Frekjær and Hornbæk (2005) presented a new version of user testing, the cooperative usability testing, where the inclusion of an interpretation phase with discussions between the participant and the test leader enables “test users and evaluators join expertise to understand the usability problems of the application evaluated” (p. 1383). Similarly, Aborå, Sandblad, Gulliksen, and Lif (2003) presented an evaluation method where the evaluator performs observation interviews with users during their ongoing work with the evaluated system.

Different usability inspection methods have been developed in order to include work-domain experts as evaluators. In Bias (1994) pluralistic walkthrough users, developers, and usability experts participate as interacting groups of evaluators. In the field of participatory design, the participatory heuristic evaluation was developed (Muller and McClard, 1995; Muller, 1998). This usability inspection method includes heuristics that target the systems’ fit to the user needs and work environment, and allows usability experts and work-domain experts to cooperate in usability inspections.

2.2. The inspection performance of work-domain experts

Little is known about the usability inspection performance of work-domain experts. Muller and McClard (1995) presented a validation study of the participatory heuristic evaluation, but they targeted only the effect of the added usability heuristics without singling out the relative contribution of the different evaluators. However, some knowledge exists on the performance of inexperienced evaluators; this may be relevant because work-domain experts are typically inexperienced evaluators. Nielsen (1992) found that, on average, it may take between 10 and 15 inexperienced evaluators to identify the same proportion of usability problems that can be identified by three usability specialists.3 Similarly, Desurvire, Kondziela, and Atwood (1992) found that a group of non-experts identified less than one third of the total number of usability problems than did a group of usability experts. These studies might indicate that both a lower number of problem predictions and lower thoroughness can be expected in the performance of work-domain experts than in that of the usability experts, even though the recent work of Howarth, Smith-Jackson, and Hartson (2009) indicates that analysis support may improve the performance of novice usability evaluators.

3. Nielsen (1992) also included a third category, double experts, who are experts in both usability and the application domain. Nielsen’s application domain (e.g. the domain of voice response telephone menus) should not be confused with work-domain as used in this article. Nielsen’s term refers to a particular category of applications and not on a particular work-context.
However, work-domain experts are more than just inexperienced evaluators. They possess knowledge that the usability experts typically do not. Therefore, studies on inexperienced evaluators might not completely reveal the performance of work-domain experts. Evidence in this regard is provided by Desurvire, Lawrence, and Atwood (1991) in a study comparing usability experts, non-experts, and users on their predictions of task completion and error-free task completion. Users and usability experts were found to perform equally well, whereas non-experts were found to have poorer performance. The conditions did, however, differ markedly between the different evaluator categories, potentially threatening the validity of this study’s conclusions.

Possibly, one reason for the lack of knowledge on the performance of work-domain experts in usability inspections is that previous usability inspection methods developed to support work-domain experts as evaluators prescribe the inspections to be conducted in multidisciplinary groups, making it difficult to single out the relative contribution of the work-domain experts. Recently, the group-based expert walkthrough has been proposed as a usability inspection method to support usability inspections with evaluator groups that consist of work-domain experts led by a usability professional (Falstad, 2007a). In this method, the test leader walks the evaluators through the system’s user interface according to predefined task scenarios. Each task scenario is walked through twice; the first time with individual note-taking, the second with plenary presentation of notes and discussion among the evaluators. The evaluation output is a set of predicted usability problems, design suggestions, and notifications of good design solutions. Group-based expert walkthrough is particularly advantageous in a study comparing the usability inspection results from work-domain experts.

The approach to the study of usability evaluation performance has advanced within the field over the last two decades. Early studies relied on comparisons on the basis of simple problem counts, which is an approach that is strongly discouraged by Cockton et al. (2003). In later years, and till date, studies on usability evaluation performance have typically included measures of validity and thoroughness. Other measures in use include reliability, effectiveness, and cost-effectiveness (Hartson, Andre, & Williges, 2001), as well as downstream utility measures (John and Marks, 1997).

Cockton et al. (2003) proposed two notions to understand the measures of validity and thoroughness: usability problems that actually exist in a given interactive system (henceforth, real problems) and usability problems that are predicted on the basis of a usability evaluation (henceforth, problem predictions). A set of problem predictions will probably include both predictions that identify real problems (henceforth, correct predictions) and predictions that constitute false positives, and will probably not include all real problems. Validity provides control regarding false positives. Thoroughness indicates the degree to which all real problems are identified.

$$\text{Validity} = \frac{\text{Number of correct predictions}}{\text{Number of problem predictions}}$$

$$\text{Thoroughness} = \frac{\text{Number of correct predictions}}{\text{Number of real problems}}$$

In recent studies on usability inspection performance, user test results have been used to identify real problems (Law and Hvannberg, 2004; Chattratichart and Brodie, 2004; Hvannberg, Law, & Lárşdóttir, 2007). However, concern has been voiced regarding user tests as such a “gold standard” (Hornbæk and Stage, 2004) and whether or not it is possible to develop sufficiently comprehensive sets of real problems through user tests (Lindgaard, 2006). As argued by Cockton et al. (2008), it is not possible to determine with confidence all real problems associated with an interactive system. Also, usability inspection may predict usability problems that may not be realistically discovered in a usability test, though they may be relevant usability problems to bring forward as the result of a usability evaluation. As a consequence, the study of validity and thoroughness seem to have two important limitations: (1) Classifying a problem as a false positive only means that it was not associated with observable problems in the user tests (Hartson et al., 2001). (2) Not all problems predicted in a usability inspection can be expected to appear in a user test and should therefore be left out in the analysis (Lindgaard, 2006).

2.4. Usability inspection performance on validity and thoroughness

We are not aware of any studies that investigate the validity and thoroughness of work-domain experts as evaluators in usability inspections. However, given that the present study compares the performance of work-domain experts and usability experts, it is relevant to provide background on previous studies on the performance of usability experts.

A summary of the literature on usability inspection performance is provided by Cockton et al. (2003). It appears that it is difficult to compare the results of existing studies, because of variations in the usability inspection procedures. The study designs also vary greatly. Some studies are conducted with many but inexperienced evaluators, typically students (Hvannberg et al., 2007; Law and Hvannberg, 2004; Cockton and Woolrych, 2001). Others include few but experienced evaluators (Desurvire et al., 1992; Lewis, Polson, Wharton, & Rieman, 1990).

Cognitive walkthrough performance has been investigated in a number of studies, typically small-scale. A study by Desurvire et al. (1992) suggested a validity value slightly more than 40% and a thoroughness value of 28%. (The suggested validity value was calculated on the basis of the information that the experts predicted seven problems that were also identified in user tests and nine problems that were not identified in user tests.) A study by Lewis et al. (1990) was interpreted by Cockton et al. (2003) to indicate a thoroughness value of approximately 30% (validity not reported). Other studies reported by Cockton et al. indicate validity values that range from 11% to 58% and thoroughness values from 10% to 50%.

Recent studies on heuristic evaluation performance indicate validity values from 28% to 45% and thoroughness values from 53% to 76% (Hvannberg et al., 2007; Law and Hvannberg, 2004; Chattratichart and Brodie, 2004; Cockton and Woolrych, 2001).
The number of evaluators used in the studies ranged from 10 to 99. All the evaluators were inexperienced.

Comparisons of thoroughness and validity values across studies should be made with great care, because these are affected strongly by the number of evaluators. Nielsen (1992) found that a single usability specialist who participated in a particular heuristic evaluation identified, on an average, less than 50% of the total number of problems identified, whereas the combined results of four evaluators included, on an average, more than 80% of all problems identified. Similarly, Sears’ (1997) reported that thoroughness values for cognitive walkthrough increased on average from 41% to 61% as the number of evaluators increased from 2 to 5.

Cockton and Woolrych (2002) argued that although increasing the number of evaluators in a usability inspection tends to improve thoroughness, it may simultaneously decrease validity; when more evaluators are available, they will make more correct predictions but also run the risk of identifying more false positives. In a study presented by these authors, the positive effect on thoroughness by increasing the number of evaluators was canceled out by the negative effect on validity when the nominal group size was increased above seven evaluators. Using several evaluators may thus increase inspection performance up to a certain point. In line with this, it is recommended to use several evaluators in usability inspections (Nielsen, 1992). Inspection performance therefore needs to be investigated also for multiple evaluators.

3. Research question

The main research question of the present study was:

How do work-domain experts perform as evaluators in usability inspections, given that the applied usability inspection method supports such experts as evaluators?

The over-all research question is formulated so as to include all work-domain experts. However, because there are different categories of work-domain expertise, this study is limited to focus only on work-domain experts with high computer experience and low system knowledge. This category of work-domain experts may be assumed to have fairly high levels of usability intuition given their high computer experience, and may thus serve as a benchmark for later studies on the performance of other work-domain expert categories.

Two hypotheses were formulated on the basis of the main research question:

Hypothesis 1. Usability inspection results generated by work-domain experts will hold the same validity as comparable results generated by usability experts, given that the applied usability inspection method supports work-domain experts as evaluators.

Little research is available to support Hypothesis 1; we are not aware of any studies that compare the validity of work-domain experts and usability experts. However, several studies (e.g. the studies presented in Section 2.4) have found that usability experts perform rather poorly with respect to validity. Also, the study of Desurvire et al. (1991) indicates similar levels of prediction performance among users and usability experts for a given set of task scenarios. Because this seems to be the most solid evidence to date, it is expected that the two evaluator categories perform equally on validity.

Hypothesis 2. Usability inspection results generated by work-domain experts will be less thorough than comparable results generated by usability experts.

There is also a lack of evidence to support Hypothesis 2. No existing research seems to compare work-domain experts and usability experts on thoroughness. However, there exists a body of research comparing the inspection performance of usability experts and non-experts, which indicates that usability experts identify significantly more usability problems than do non-experts. Although it is problematic to equate work-domain experts and non-experts, this seems to be the most relevant present evidence. The hypothesis is therefore formulated in favor of the usability experts. It is expected that the usability experts make more problem predictions than do the work-domain experts, and when comparing the problem predictions to the user test results, it is expected that the usability experts make more correct predictions.

In addition to the formulated hypotheses, the performance of nominal groups of either work-domain experts or usability experts is explored. Such exploration is useful, because usability inspections are often conducted with multiple evaluators.

4. Research method

The research design was set up as a controlled comparison of the evaluator performance of work-domain experts and usability experts. The study was confined to work-domain experts that had high computer experience and low system experience. Both evaluator categories participated in group-based expert walkthroughs under exactly the same conditions. This usability inspection method was chosen because it (1) has been developed to support non-usability experts as evaluators and (2) enables data collection at the level of individual evaluators. The dependent variables of the study were validity and thoroughness. The basis for comparing the performance was user test results. The research design may be classified as a controlled experiment according to the operational definition of Sjøberg, Hannay et al. (2005): “A randomized experiment or a quasi-experiment in which individuals or teams (the experimental units) conduct one or more software engineering tasks for the sake of comparing different populations, processes, methods, techniques, languages, or tools (the treatments)” (p. 735).

4.1. The objects of evaluation

The objects of evaluation were four web-based applications where researchers can register, maintain, and retrieve information and documentation related to empirical research studies. The respective applications were developed by four independent commercial ICT (information and communication technology) companies on the basis of the same requirements specification. The functional requirements were described in detail to ensure functional equivalence between the applications. The motivation for developing more than one application was to study the effects of different development processes on project and product quality (Anda, Benestad, & Hove, 2005; Anda, 2007; Anda, Sjøberg, & Mockus, 2009).

Each of the four applications was developed to be used as part of the web site of the research organization Simula Research Laboratories. The purpose of each application was to increase openness in research, by making information and documentation from research studies available through the Internet. Internal end users could use each application to enter key information about their research studies (e.g. study type, duration, number of respondents, and related publications) and upload documentation (e.g. procedure descriptions, forms, and reports). External and internal users could use each application to access the registered information and documentation. Both entry of, and access to, the information and documentation were conducted through the web user interfaces of each application. The main screen for registering data and documentation in one of the four applications is presented in Fig. 1.
At the time of evaluation, the applications were in pilot use and integrated in the research organization’s website. Since the four applications were functionally equivalent, the end users were randomly assigned to one application in the pilot period. No changes were made to the applications between the time of the user tests and the expert walkthroughs.

The end users of the four applications were researchers in the field of ICT. The full range of functionality was to be used by PhD students and scientific employees at the research organization. The retrieval functionality was also to be used by external end users. The evaluated functionality, which was related to entering and maintaining information and documentation, was meant for internal end users only.

Neither the work-domain experts nor the usability experts had previous knowledge of the applications, making both of them low on system knowledge. The basic functionality of the applications was understood easily by all the participants regardless of the evaluator background; at the same time, context knowledge of the work-domain experts was assumed to provide them with an advantage in the evaluation situation. This set up controls for possible confounds between differences in system knowledge (which was equal between both evaluator categories) and work-domain knowledge (which differed between the two evaluator categories).

4.2. Conditions and participants

Twenty-nine evaluators participated in the usability inspections, but the results from two of them were discarded from the analyses. One usability expert was found to have too little work experience (<1 year). One work-domain expert reported to have taken a course in HCI.

The usability inspection results of 15 work-domain experts and 12 usability experts were analyzed. The work-domain experts were all PhD students of ICT at the University of Oslo, recruited to match the sample of user test participants (described in Section 4.3.3). Their mean age was 29 years (SD = 3.5). Their mean work experience as PhD students was 2 years (SD = 1.2). They had no previous experience with the evaluated applications. None had taken courses in HCI.

The 12 usability experts were recruited from seven ICT companies. All worked as interaction designers, information architects, or usability consultants. Their mean age was 33 years (SD = 5.9). All of them had more than 1 year of work experience with HCI; (M = 5, SD = 3.8). None had any previous experience with the evaluated applications.

4.3. Usability inspection material and procedure

Four usability inspection sessions were conducted. Each session included the test leader and 6–8 evaluators, and lasted for 2.5 h, of which 30 min were used for introductory purposes. Each evaluator participated in one session only. All the four applications were evaluated in each session. To control for confounding variables, each session included both work-domain experts and usability experts.

By organizing the sessions in this way, the two evaluator categories received exactly the same information. Each session was planned to include four work-domain experts and four usability experts. However, three usability experts were not able to attend their session.

4.3.1. Usability inspection material

The introductory material provided descriptions of the work-domain, the applications, the general objectives of a usability evaluation, and the usability inspection procedure. Sets of three task scenarios were provided for each of the four applications. Each task scenario consisted of predefined steps toward task completion. Each set of task scenarios included:

(a) Login; 3–5 steps (depending on the application).
(b) Register a new research study and related information and documentation; 15–22 steps.
(c) Open an existing study to edit its related information and documentation; 3–6 steps.

to structure the reporting of the usability problems, all the evaluators were given a form structured according to the steps of the task scenarios. Each step of each scenario was assigned a reference number, and spaces were allocated for writing down usability problem predictions and classifications regarding severity and frequency.

Severity classifications for problem predictions were defined as follows:

- **Cosmetic**: Typical users will experience minor delays or obstacles when performing the task.
- **Serious**: Typical users will experience major delays or obstacles when performing the task.
- **Critical**: Typical users will not be able to complete the task (or this part of the task) without assistance.

Problem frequency was defined as the proportion of actual users that were estimated to encounter the predicted...
problem. The problem frequency categories were 25%, 50%, 75% and 100%.

4.3.2. Usability inspection procedure

After the introduction, walkthroughs were conducted serially for all four applications; about 30 min were spent on each application. This evaluation time was found to be sufficient to walk through the three predefined task scenarios for each application. The application user interfaces were presented by means of a video projector.

The usability inspection followed an adapted version of the group-based expert walkthrough. For each step of a task scenario, the test leader asked the evaluators which action they believed would bring them closer to completing the task scenario. After allowing the evaluators a few seconds of silent reflection, the test leader presented the correct action and asked the evaluators to write down usability problem predictions and associated classifications of severity and frequency. If relevant, the test leader also presented error messages, empty search result messages, and different possible actions leading to the same result. An example of different actions leading to the same result is to initiate a search by either typing Enter or clicking the Search button. At any time in the walkthrough, the evaluators were allowed to ask for clarification or to have steps repeated. Following the completion of a task scenario, the test leader asked the evaluators to look over their notes and write down any over-all problem predictions, if relevant.

In contrast to the original group-based expert walkthrough (Følstad, 2007a), the usability inspection procedure did not allow communication between the evaluators to (a) facilitate data collection on the level of individual evaluators and (b) eliminate the exchange of usability judgments between the two categories of evaluator. However, to ensure that all the evaluators had understood their task, they were permitted to make plenary presentations of a small number of predictions associated with the first application. No changes in the evaluator notes were to be made on the basis of these presentations.

4.3.3. The basis for comparison: user test results

Before the usability inspections, the basis for comparing the two categories of evaluator was established: a master set of real user problems derived from user tests. The user tests were conducted with 18 participants who used all the four applications. All the participants were PhD students (17) or Scientific Programmer (1) at the research organization Simula Research Laboratories. All but one had worked there for 1 year or more (M = 2.1, SD = 1.2). Their mean age was 29 years (SD = 3.9). None had tried the evaluated applications, even though five had superficial knowledge of them. The participants tested the applications in one of four different orders.

The tasks for the user tests were developed on the basis of the specification documents for the application and an initial usability inspection.

All the test sessions were recorded and analyzed using MORÆ. All the analysts, the first author of this article and an independent analyst, conducted the analysis independently of each other. Neither of the analysts were the same person as the test leader. An event was registered as a usability problem if it caused (a) an incomplete task outcome or (b) an obstruction or delay in the task completion.

Upon completion of their individual analyzes, both the evaluators independently classified their observed problem instances according to a commonly agreed set of problem descriptions. The classification was conducted in two iterations to update the common set of problem descriptions. The usability problem instances from the individual analyzes were then merged on the basis of the timestamps and classifications of the observed problem instances. This merged set included a total of 200 observed problem instances; 148 of these were observed by both of the analysts, 39 by only Analyst 1, and 13 by only Analyst 2. In addition, 28 problem instances identified by Analyst 1 and seven problem instances identified by Analyst 2 were discarded during the problem merging, because the two evaluators upon discussion did not find the observed event to be a usability problem instance. All observations made by one analyst only were discussed between the analysts during the problem merging. If necessary, as it indeed was for the vast majority of the cases, the recordings from the user test sessions were reviewed as part of this discussion.

The merged set of real problems included 53 usability problems across the four applications. All the problems were in the level of problem tokens, not types; that is, similar problems in different contexts of the application were counted as different problems. This approach is in line with the previous studies on the validity and thoroughness of usability inspections (e.g. Hvannberg et al., 2007; Chattratichart and Brodie, 2004; Lindgaard, 2006). The severity of the merged problems was classified according to the following categories, balancing the frequency and severity of the observed problem instances:

- Cosmetic: Either 1–50% of the participants experienced minor (cosmetic) obstructions/delays, or 1–20% experienced obstructions/delays of which some were major (serious or critical).
- Serious: 1–50% Experienced obstructions/delays in total, and for some of these, but not more than 20%, these were major.
- Critical: Either >20% experienced major obstructions/delays, or >50% experienced obstructions/delays in total.

Of the 53 observed problems, 41 were classified as cosmetic, 7 serious, and 5 critical. It should be noted that the observed problems were related only to three task scenarios for each application. The evaluated applications may have been associated with a higher number of usability problems, but these were not relevant to this study, because both the user tests and the usability inspection targeted the same three task scenarios for each of the applications.

4.4. Data collection and analyzes

4.4.1. Data collection and prediction merging

During the usability inspections, data on predicted usability problems were collected using the structured problem reporting forms that were filled by the individual evaluators. Personal data and background on the evaluators were collected by means of questionnaires.

The problem recording format included space for reporting problem predictions associated with each of the steps of the walkthrough task scenarios. In addition, the form included space for reporting problem predictions that were not directly associated with any of the steps in the walkthrough procedure. Problem predictions reported in this space were typically either high level or only remotely associated with the evaluation, and were therefore excluded from the subsequent analysis (this amounted to 8% of the total number of problem predictions).

All problem predictions were analyzed by two independent analysts. The analysis was blind; that is, none of the analysts were aware of neither which of the evaluators produced the different problem predictions nor which of the evaluator categories were associated with each of the problem predictions.

The first step of the analysis was to match the problem predictions (from the usability inspection) with the real problems (from
the user tests). In this matching, the problems were classified as follows:

- **Correct predictions**: A usability inspection prediction that corresponded to a given real usability problem.
- **Missed problems**: Real usability problems that did not correspond to any usability inspection prediction.
- **False positives**: Usability inspection predictions that, given they were correct, should have affected user test performance.
- **Not tested**: Problem predictions that could not be expected to affect user test performance.
- **Not a problem**: Evaluator feedback that was not problem predictions, for instance positive feedback.
- **Uncertain**: The analyst was not sure how to interpret the usability inspection prediction.

On individual completion of the matching, any disagreements were resolved through an asynchronous process between the analysts. First, Analyst 2 suggested a conclusion to the disagreement. Subsequently Analyst 1 went through the suggestions. Ninety-two percent of the suggested conclusions of Analyst 2 were retained. Six percent of the remaining suggestions were changed by Analyst 1 back to the original matching result of Analyst 2, and 2% were changed back to the original matching result of Analyst 1. Since, at this point in the resolving of disagreements, Analyst 1 only made a small proportion of changes back to his original matching result, the matching process was concluded.

The matching and subsequent resolving of disagreements returned a merged set of problem predictions where all the correct predictions were matched with a given real problem. However, the analysis of validity and thoroughness for multiple evaluators also required that all the problem predictions that were classified as false positives were matched with a master set of false positives. This matching was conducted as follows: a master set of false positives was established by Analyst 1 on the basis of the iterations of categorization and trial coding on the set of false positive instances. This master set of false positives was then used independently by two independent analysts in a final coding of the false positive instances, allowing the investigation of inter-rater reliability.

4.4.2. Validity and thoroughness at the level of individual evaluators

Validity and thoroughness values were calculated for each evaluator according to the formulae presented in Section 2.1. Hypothesis 1 predicted no group differences regarding validity. Since the non-existence of group differences are difficult to investigate on the basis of statistical tests alone, the validity values across the two evaluator categories were also investigated through exploratory analyses. Hypothesis 2 predicted group differences regarding thoroughness. The thoroughness values were therefore investigated only by statistical tests. Also, statistical tests were used to compare work-domain experts and usability experts with respect to the number of problem predictions that they made.

4.4.3. Validity and thoroughness for nominal groups of evaluators

To explore the effect of combining the results from multiple evaluators, validity and thoroughness values for nominal groups of evaluators were estimated based on the prediction probability for each individual problem prediction. Prediction probability is understood as the probability that a prediction is made by at least one evaluator in a nominal group of evaluators, and was calculated for homogenous evaluator groups of fixed sizes, that is, for nominal groups of either work-domain experts or usability experts, group sizes from 2 to 10.

The prediction probabilities were calculated on the basis of the proportion of evaluators in our data set that either did or did not make a given problem prediction. The prediction probability for a given problem prediction is equivalent to 1 minus the probability that none of the evaluators in the group make the prediction. If \( n \) is the total number of evaluators of a given category, \( k \) is the total number of evaluators of that category that made the prediction, and \( m \) is the nominal group size, the prediction probability was calculated by the following formula:

\[
\text{Prediction probability} = 1 - \frac{n - k}{n} \frac{n - k - 1}{n - 1} \cdots \frac{n - k - m + 1}{n - m + 1}
\]

(The product of \( m \) fractions)

For work-domain experts, \( n = 15 \). For usability experts, \( n = 12 \). The validity and thoroughness values for each nominal group size were calculated as follows:

\[
\text{Validity} = \frac{ef}{(ef) + (hf)}
\]

\( e = \text{Mean prediction probability for real problems} \),
\( f = \text{Number of real problems} \),
\( h = \text{Number of false positives} \),
\( j = \text{Mean of false positives} \)

**Thoroughness** = Mean prediction probability for real problems

These validity and thoroughness values thus serve as estimations of the average validity and thoroughness values for all possible combinations of evaluators in fixed size homogenous nominal groups.

4.4.4. Controlling for session effects

To provide identical conditions for the two evaluator categories, four usability inspection sessions including both work-domain experts and usability experts were conducted. However, such simultaneous presentation may possibly have the unwanted effect that evaluators in the same session influence each other. Such within-session influence is considered to be unlikely, given that the evaluators were not allowed to communicate during the evaluation. However, analyzes to control this needed to be included.

If such effects were present, they would probably manifest themselves by between-session variation; that is, the results from the evaluators of one session would differ significantly from the results of the evaluators of other sessions. Group differences were therefore investigated with respect to validity values, thoroughness values, and number of problem predictions. Validity values were investigated for all evaluators. Thoroughness values and number of problem predictions were investigated for homogenous evaluators because the two evaluator categories were assumed to differ on these two measures. In addition to these analyzes, the risk of within-session influence was also considered with respect to the proportion of evaluators within each group that made the different predictions.

5. Results

5.1. Overview of the usability inspection results

The 27 evaluators produced 1335 problem prediction instances across the four applications. Each application received between 22% and 29% of the problem predictions from either of the evaluator categories. The work-domain experts produced 562 problem prediction instances (\( M = 37, SD = 11 \)); the usability experts produced 773 (\( M = 64, SD = 19 \)). This difference in problem prediction instances between the evaluator categories was found to be statistically significant in an independent-samples t-test (\( t(25) = -4.54, p \ (one-tailed) < 0.01 \)).
Problem severity ratings were provided by the evaluators for 89% of the prediction instances (95% for the work-domain experts and 85% for the usability experts). Only 24% of the prediction instances were rated as serious or critical by the evaluators (22% for the work-domain experts, 26% for the usability experts). Frequency classifications were provided by the evaluators for less than 60% of the problem predictions, and are therefore not included in what follows.

In addition to the above 1335 problem prediction instances, 337 problem prediction instances were discarded from the analysis because they were classified by the evaluators as not directly belonging to any of the walkthrough steps (135) or classified by the analysts as Uncertain (57) or Not a problem (66). In addition, 79 problem predictions were discarded because they were found to be doubles of correct predictions and false positives already identified by the same evaluator. Of the discarded prediction instances, 108 were produced by the work-domain experts and 229 by the usability experts.

In the matching of the 1335 problem predictions with the master set of real problems, 456 were classified as correct predictions, 394 as false positives, and 485 as not tested. Problem predictions classified as not tested were not included in the analyzes of validity and thoroughness. Of the problem predictions classified as not tested, 197 ($M = 13, SD = 8$) were produced by the work-domain experts and 288 ($M = 24, SD = 12$) by the usability experts.

The merged set of correct predictions included 41 problem predictions; that is, 77% of the 53 real problems identified in the user tests were predicted by one or more of the evaluators. Twelve of the real problems (23%) were not predicted by any evaluator. The merged set of false positives included 129 false positives, predicted by one or more evaluator. The reason for the larger proportion of false positives was the relatively low number of evaluators that predicted each false positive ($Mdn = 2$) as opposed to the number of those that made each correct prediction ($Mdn = 8.5$). Details of the distribution across applications and evaluator categories are presented in Table 1.

The matching was conducted by two analysts. As a measure of inter-rater agreement for the matching of the full set of problem predictions with the master set of real problems, Cohen’s Kappa was calculated for the set of problem predictions classified by both analysts as correct predictions, false positives, or not tested (1349 of the problem prediction instances). Because the classification of correct predictions required the analysts to choose between several real problems, the number of alternatives for classifying each problem prediction instance was greater than these three basic classification categories. For simplicity, it was assumed that the evaluators could choose between two relevant real problems for each correct prediction, and that the number of alternative classifications for each problem prediction was therefore four. The resulting Kappa value was 0.53, indicating moderate agreement (Landis and Koch, 1977).

Regarding the matching of the subset of the 409 instances of false positives with the master set of 129 false positives, the analysts agreed in 86% (354) of the instances.

5.2. Validity and thoroughness of individual experts’ usability inspection results

5.2.1. Validity

No differences in validity values were expected between the evaluator categories. The distributions of validity values for the different evaluated applications and problem severity levels were explored through box-plots. The box-plot that visualizes the overall validity values for each of the two evaluator categories is shown in Fig. 2. Note that the interval of the validity values for the work-domain experts completely overlaps that of the usability experts.

To indicate the similarity in the distributions of the validity values for the two evaluator categories, the result of an independent-samples t-test comparing the two categories on validity values for all problem predictions is reported. No significant difference was found between the overall validity value for work-domain experts ($M = 0.6, SD = 0.1$) and usability experts ($M = 0.5, SD = 0.1$). The t-test values were: $t(25) = 0.59, p$ (two-tailed) $= 0.56, r = 0.12$. The effect size $r$ was calculated according to Field (2005, p 296).

To provide additional detail, independent-samples t-tests and associated effect sizes are reported for each prediction severity level (Table 2) and evaluated application (Table 3). No significant differences were found, even though the higher validity score of the work-domain experts for one of the evaluated applications approached significance. The work-domain experts had the larger validity value in all of the tests associated with effect sizes exceeding $r = 0.10$.

5.2.2. Thoroughness

Differences in thoroughness values were investigated with Mann–Whitney U tests. This distribution free test was used because Shapiro–Wilk tests indicated non-normal distributions for four of the combinations of severity level and evaluator category. The difference was significant for the overall thoroughness values for the work-domain experts ($Mdn = 0.25$) and usability experts ($Mdn = 0.40$). The Mann–Whitney U test values were: $U = 8.00, p$ (one-tailed) $< 0.01, r = 0.77$. The effect size $r$ was calculated according to Field (2005, p 532). Significant differences between the evaluator categories were also found for each of the problem

![Fig. 2](image-url)
severity levels (Table 4) and for each of the evaluated applications (Table 5).

5.3. Validity and thoroughness for nominal groups of evaluator

Validity and thoroughness values were calculated for nominal groups of, respectively, work-domain experts and usability experts. The calculations were based on the prediction probabilities for the each of the 53 real problems and 129 false positives. The results were plotted with nominal group size on the x-axis, and validity and thoroughness values on the y-axis (see Fig. 3). For both the evaluator categories, when the number of evaluators increases, the thoroughness values may reach well above 0.6, however, at the cost of the validity values, which become below 0.4. The figure also shows that a larger number of work-domain experts are required to reach the same thoroughness levels as for usability experts.

5.4. Controlling for within-session influence between the evaluators

Influence between the evaluators participating in the same usability inspection session was, if existing, assumed to be manifested as between-session differences. These were investigated using ANOVA. No significant session differences were found neither for validity ($F(26) = 0.64, p$ (two-tailed) = 0.61), thoroughness (Work-domain experts: $F(14) = 0.75, p$ (two-tailed) = 0.55; Usability experts: $F(11) = 0.98, p$ (two-tailed) = 0.45), nor the total number of problem prediction instances (Work-domain experts: $F(14) = 0.79, p$ (two-tailed) = 0.53; Usability experts: $F(11) = 0.30, p$ (two-tailed) = 0.83). Session differences on thoroughness and the total number of problem prediction instances were investigated separately for the work-domain experts and usability experts due to significant group differences.

Moreover, the low proportion of problem predictions made by more than two participants within each group seems to indicate no substantial within-group influence between the evaluators. In all four groups more than 40% of the problem predictions were made by only one evaluator, and more that 60% of the problem predictions were made by only one or two evaluators. Less than 4% of the predictions made in any of the groups were made by all evaluators.

6. Discussion

This section discusses the results as follows: First, some considerations regarding the work-domain experts are presented. Second, the validity and thoroughness of the results generated by

### Table 2
Mean validity values for individual evaluators for each prediction severity level, with independent-samples $t$-test statistics and effect sizes. Mean numbers of problem predictions are shown in parentheses.

<table>
<thead>
<tr>
<th>Evaluators’ severity classifications</th>
<th>Work-domain experts</th>
<th>Usability experts</th>
<th>$t$-test</th>
<th>Effect size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>Cosmetic</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(21)</td>
<td>0.5</td>
<td>0.1</td>
<td>0.5</td>
<td>0.1</td>
</tr>
<tr>
<td>Serious</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(7)</td>
<td>0.7</td>
<td>0.3</td>
<td>0.7</td>
<td>0.2</td>
</tr>
<tr>
<td>Critical</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1)</td>
<td>1.0</td>
<td>0.0</td>
<td>0.8</td>
<td>0.4</td>
</tr>
<tr>
<td>Not classified</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(3)</td>
<td>0.7</td>
<td>0.5</td>
<td>0.6</td>
<td>0.3</td>
</tr>
</tbody>
</table>

### Table 3
Mean validity values for individual evaluators for each of the evaluated applications, with independent-samples $t$-test statistics and effect sizes. Mean numbers of problem predictions are shown in parentheses.

<table>
<thead>
<tr>
<th>Evaluated applications</th>
<th>Work-domain experts</th>
<th>Usability experts</th>
<th>$t$-test</th>
<th>Effect size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>Appl. A (8)</td>
<td>0.4</td>
<td>0.2</td>
<td>0.4</td>
<td>0.1</td>
</tr>
<tr>
<td>Appl. B (7)</td>
<td>0.6</td>
<td>0.2</td>
<td>0.6</td>
<td>0.1</td>
</tr>
<tr>
<td>Appl. C (8)</td>
<td>0.6</td>
<td>0.2</td>
<td>0.6</td>
<td>0.1</td>
</tr>
<tr>
<td>Appl. D (8)</td>
<td>0.7</td>
<td>0.2</td>
<td>0.7</td>
<td>0.1</td>
</tr>
</tbody>
</table>

### Table 4
Thoroughness values for individual evaluators for each problem severity level, with Mann–Whitney $U$ test statistics and effect size estimates. Real problem counts are presented in parentheses.

<table>
<thead>
<tr>
<th>Severity of real problems</th>
<th>Work-domain experts ($n = 15$)</th>
<th>Usability experts ($n = 12$)</th>
<th>Mann–Whitney $U$</th>
<th>Effect size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>p25</td>
<td>p50</td>
<td>p75</td>
<td>Mean</td>
</tr>
<tr>
<td>Cosmetic</td>
<td>(41)</td>
<td>0.17</td>
<td>0.19</td>
<td>0.24</td>
</tr>
<tr>
<td>Serious</td>
<td>(7)</td>
<td>0.14</td>
<td>0.28</td>
<td>0.28</td>
</tr>
<tr>
<td>Critical</td>
<td>(5)</td>
<td>0.40</td>
<td>0.60</td>
<td>0.80</td>
</tr>
</tbody>
</table>

### Table 5
Thoroughness values for individual evaluators for each of the evaluated applications, with Mann–Whitney $U$ test statistics and effect size estimates. Real problem counts are presented in parentheses.

<table>
<thead>
<tr>
<th>Evaluated applications</th>
<th>Work-domain experts ($n = 15$)</th>
<th>Usability experts ($n = 12$)</th>
<th>Mann–Whitney $U$</th>
<th>Effect size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>p25</td>
<td>p50</td>
<td>p75</td>
<td>Mean</td>
</tr>
<tr>
<td>Appl. A (8)</td>
<td>0.25</td>
<td>0.25</td>
<td>0.38</td>
<td>0.41</td>
</tr>
<tr>
<td>Appl. B (10)</td>
<td>0.20</td>
<td>0.30</td>
<td>0.40</td>
<td>0.30</td>
</tr>
<tr>
<td>Appl. C (17)</td>
<td>0.12</td>
<td>0.18</td>
<td>0.29</td>
<td>0.29</td>
</tr>
<tr>
<td>Appl. D (18)</td>
<td>0.17</td>
<td>0.28</td>
<td>0.33</td>
<td>0.33</td>
</tr>
</tbody>
</table>
individual evaluators of the two evaluator categories are discussed. Third, the findings related to combining the results from different evaluators in nominal groups are conferred. Last, the validity of the study itself is examined.

6.1. On the work-domain experts

The evaluated applications were work-domain specific: they were developed particularly for researchers within the field of ICT. The evaluators categorized as work-domain experts were all PhD students in this field, and had all been so for at least 1 year, indicating acceptable work-domain expertise. Their research interest also ensures high computer experience, and their lack of experience with the evaluated applications ensures low system experience. In consequence, the present study may be held to provide knowledge regarding the usability inspection performance of work-domain experts that have high computer experience and low system experience; cf. the work-domain expert categories described in Section 2.1.

6.2. Differences between evaluator categories regarding validity and thoroughness

6.2.1. No difference in validity?

As predicted in Hypothesis 1, the data indicated no significant differences in validity between the two evaluator categories. The validity distributions of the two evaluator categories overlapped greatly (Fig. 2), and the similarities in distribution characteristics for the two groups also extended to the subsets of problem predictions for each of the severity levels (Table 2) and each of the evaluated applications (Table 3). Only for one of the evaluated applications did the difference between the evaluator categories approach statistical significance, and in this case the validity values of the work-domain experts were better than those of the usability experts.

In particular, the two evaluator categories performed similarly for all levels of problem severity, though the absolute values of validity changed across the problem severity levels; when problem severity increased so did the validity values for both evaluator categories. This indicates that the similarity in validity was a stable characteristic of the analyzed data, even though there might well have been interaction effects between predicted problem severity level and validity values.

The study results support Hypothesis 1. Given that the usability inspection method supports work-domain experts as evaluators, these experts seem to be just as effective as usability experts when it comes to avoiding the inclusion of false positives in problem prediction sets.

A final note on the validity values for individual evaluators: these values ranged between 0.4 and 0.7 across the evaluated applications, with an over-all mean of slightly above 0.5. On the basis of informal inspection, these values are fairly high compared with the studies of cognitive walkthrough and heuristic evaluation reviewed by Cockton et al. (2003). This should not be taken to mean that group-based expert walkthrough may be expected to return more valid results than does cognitive walkthrough or heuristic evaluation, because the present study did not include controlled comparisons between these usability inspection methods. However, the validity values found in the present study may serve as an inspiration for future comparative studies.

6.2.2. Difference in thoroughness

The individual usability experts made, on an average, both more problem predictions and more correct predictions than did the individual work-domain experts. As predicted in Hypothesis 2, independent-samples t-tests indicated a significant difference in thoroughness between the work-domain experts and usability experts.

This difference in thoroughness held true both for each of the problem severity levels (Table 4) and each of the evaluated applications (Table 5). It was observed that both evaluator categories seemed to improve their thoroughness values when the severity of real problems increased. This is a comforting observation, because it indicates that expert evaluators are more likely to discover problems of high severity than those of low severity.

The study results support Hypothesis 2. Individual work-domain experts seem to be outperformed by individual usability experts when it comes to identifying as many as possible of the real problems associated with a given application.

6.3. Validity and thoroughness in nominal groups

Given that the existing literature indicates that validity and thoroughness values change with the number of evaluators involved, the effect of combining the results of nominal groups in the present sample was explored. Only homogenous nominal groups were investigated, that is, groups consisting of evaluators from the same evaluator category only.

No hypotheses were associated with this exploration. However, because of the similarity in the validity values for the individual work-domain experts and usability experts, it was considered that similar usability inspection performance results could be achieved by nominal groups of either of the evaluator categories, provided the group size was disregarded. The graph presented in Fig. 3 indicates some support for this. Both for the work-domain experts and usability experts, the combination of evaluators in nominal groups improved thoroughness and reduced validity. Moreover, the validity and thoroughness values of m work-domain experts approach the corresponding values of m usability experts.

In the present study, an increase in the number of experts from one to two yielded an increase in thoroughness of 0.12 for work-domain experts and 0.14 for usability experts, and a decrease in validity of 0.05 and 0.06, respectively. Subsequent increases in the number of evaluators were associated with diminishing returns for both evaluator categories. On the basis of Fig. 3, one may speculate that the combined results of 4–6 work-domain experts, and similarly 2–3 usability experts, may provide a good balance of the costs and benefits associated with involving more evaluators, even though this clearly depends on the actual costs of more evaluators and the importance of increased thoroughness.
The design of this study does not enable conclusions neither with respect to whether 2m work-domain experts generate equivalent results as m usability experts, nor with respect to the number of evaluators required to reach an optimal balance of costs and benefits. However, the study may serve as an inspiration for these kinds of hypotheses in future research.

Følstad (2007b) concluded that usability inspection results generated by work-domain experts may have a greater impact on subsequent development than the results of usability experts. These findings, combined with the results of the present study, seem to justify recommending a more extensive use of work-domain experts as evaluators in usability inspections.

6.4. Generality of the results across usability inspection methods

In the present study, the usability method used was group-based expert walkthrough. This method was chosen in order to investigate the evaluator performance of work-domain experts at the level of individual evaluators. Since the study did not include other usability inspection methods, no claims can be made regarding the generality of the results across usability inspection methods. However, one may speculate that similar results may be achieved in other usability inspection methods that include support for work-domain experts as evaluators; in particular Bias' (1994) pluralistic walkthrough, because it shares many of the characteristics of the group-based walkthrough (e.g. walkthrough structure, and individual note-taking).

On the other hand, one may speculate that the validity and thoroughness obtained by the work-domain experts in the present study would not have been achieved if they had used usability inspection methods not intended to support work-domain experts; e.g. heuristic evaluation and cognitive walkthrough. The main reason is that these usability inspection methods, however well suited for usability experts, lack process support for non-usability experts.

In order to address the above speculations, future research should compare the performance of work-domain experts in different usability inspection methods. Comparisons should include (1) usability inspection methods for which claims have been made in the literature that they support work-domain experts as evaluators (e.g. pluralistic walkthrough, participatory heuristic evaluation, and group-based expert walkthrough) and (2) usability inspection methods for which no such claims have been made (e.g. heuristic evaluation and cognitive walkthrough).

6.5. Validity issues of the present study

Early empirical investigations of usability evaluation performance were severely criticized because of their lack of scientific rigor. The validity of the present study is therefore discussed according to the structure of Gray and Salzman (1998).

6.5.1. Statistical conclusion validity

Given the challenge of recruiting evaluators to participate in the study, the sample sizes had to be somewhat smaller than would have been desirable with respect to the statistical tests of differences between the evaluator categories. For an independent-samples t-test with group sample sizes of 12 and 15 to have a statistical power of \(1 - \beta = 0.8\) given a statistical significance level \(\alpha = 0.05\), the effect size needs to be \(r = 0.44\). This less power-efficient Mann–Whitney U test will, according to Clark-Carter (1997, p. 224), have a comparable statistical power to an independent-samples t-test with a 5% reduction in sample size, and therefore requires an effect size of \(r = 0.45\) to have the same statistical power as an independent-samples t-test. This means that only large effect sizes could have been expected to appear as statistically significant with the present sample size. However, this should not be regarded as particularly threatening to the statistical validity of the findings of the present study. Regarding Hypothesis 1: A number of t-tests were associated with an effect size \(r > 0.10\). For all of these the work-domain experts had higher mean validity values than did the usability experts. In consequence, the data in no way supports the speculation that usability experts would be associated with higher validity values than that of work-domain experts if only the samples were large enough, which would presumably be the favoured alternative hypothesis of the HCI community. Regarding Hypothesis 2: The effect size associated with the Mann–Whitney U test for all real problems was estimated to be \(r = -0.77\); cf. the effect size of \(r = 0.45\) required for a statistical power of 0.8 for a Mann–Whitney U test. This clearly indicates that the statistical power of the tests was sufficient for the purpose.

6.5.2. Internal validity

The problem prediction merging was conducted by two independent analysts. This strengthens the study’s internal validity, and may also be regarded as necessary given the evaluator effect documented within the field of HCI (Hertzum and Jacobsen, 2003).

The study also included other important controls against analysts’ bias. The analysts were blind to which problem predictions that belonged to which evaluators. The analysis was simplified by the highly structured report formats used by the evaluators, which forced problem prediction instances to be embedded in explicates contexts and to refer unambiguously to particular application properties. In addition, the analysis was conducted to be as transparent as possible, to allow detailed reconstructions. For example, following the advice of Cockton et al. (2008), the analyses were held at the level of individual evaluators.

A potential threat to internal validity, however, is the possible effects of having both work-domain experts and usability experts present in the same usability inspection sessions. The evaluators were required not to communicate with each other. However, it is possible that evaluators influenced each other by their nonverbal behavior or questions addressed to the test leader. To control for such effects, between-session differences in validity, thoroughness, and number of problem predictions were analyzed. No significant differences were detected. Also, the fact that more than 60% of the problem predictions in a usability inspection session were made by only one or two analysts seems to reduce the likelihood of within-session influence. The possibility that there was such influence cannot be eliminated completely, but the control measures indicate that such effects were hardly substantial.

6.5.3. Construct validity

A potential threat to the construct validity of the study is the utilization of user test results as a master set of real user problems. As discussed in Section 2.3, such a master set is, at best, only an approximation of the full set of real user problems. This is reflected in the proportion of predicted problems classified as not tested, that is, not expected to affect user test performance. In consequence, the validity and thoroughness values presented in this study do not include all predictions made by the evaluators: only those predictions that may be compared with user test results are taken into account.

The authors believe that as long as one agrees with the current tradition of evaluating usability inspection performance on the basis of user test results, the master set of real problems used in the present study may be regarded as a sufficiently valid basis for evaluation; the master set was based on user test results from...

---

4 Power analyses for independent-samples t-test was conducted with G’Power 3.0.8 (Faul, Erdfelder, Lang, & Buchner, 2007). The effect size \(r = 0.44\) corresponds to \(d = 0.99\) (Cohen, 1988).
18 participants, a number that is high enough to comply with Niel-
sen’s (2006) recommendations for quantitative evaluations. How-
ever, future research should include complementary measures of
evaluator performance that do not depend on all real problems
being known. Such measures may, for example, be related to the
developer’s judgement or prioritizing of the problem predictions.

Another issue of relevance for construct validity is the actual
work-domain experts used in the study. As PhD students, they
were all in their early years as researchers. Possibly, work-domain
expertise would have been even higher if we had used highly expe-
rienced researchers as evaluators. We do not, however, regard this
issue particularly problematic with respect to the conclusions of
the study. If evaluators of even higher work-domain expertise
had been used, the quality of the evaluation output should, if any-
thing, increase; something that would not alter the main conclu-
sions of the study.

6.5.4. External validity

The present study is associated with limitations related to
external validity. As discussed above, only one usability inspection
method was used. Moreover, the study involved objects of evalua-
tion and experts of one work-domain only. These limitations do
not reduce the value of the study as an early investigation of the
usability inspection performance of work-domain experts. How-
ever, the limitations make it difficult to decide whether or not
the study’s results depend on this particular kind of work-domain
experts or applications. Further research involving a broader range
of work-domains is needed to resolve this issue.

It is, however, comforting that the main hypotheses of the study
are supported both across all the evaluated applications and for
each of the applications in isolation, although it should be noticed
that the four applications are functionally equivalent.

Another potential threat to the study’s external validity is that
the samples of participating evaluators were not randomly se-
lected from a pool of the two expert category populations. It is
not easy to define the population of the usability experts (Bark
et al., 2005). In addition, potential evaluators had busy work sched-
ules, in particular the usability experts. The lack of random selec-
tion of participants makes it difficult to rule out the possibility
that unknown sample characteristics compromised external valid-
ity. However, the sample of the present study seems to comply
with the authors’ intuitions regarding the population characteris-
tics age, background, and work experience.

6.5.5. Conclusion validity

The conclusion validity of the present study depends on
whether or not the claims made may be justified on the basis of
the presented background and results. The study has clear research
problem and hypotheses, variables based in an established tradi-
tion of studying usability evaluation performance, and a simple re-
search design constructed as a controlled experiment. These
characteristics would have made it fairly easy to identify breaches
in conclusion validity. The authors are therefore confident in the
conclusions of the study.

7. Conclusion and future work

The results of the study indicate that work-domain experts may
be used as evaluators in usability inspection without compromis-
ing prediction performance. Hence, work-domain experts may rep-
resent a valuable alternative when knowledge of the context is
required in the usability inspections. However, note that inspec-
tions with work-domain experts seem to require a higher number
of evaluators to reach a level of thoroughness comparable with
that of evaluations with usability experts.

Existing studies of the usability inspection performance of
work-domain experts are associated with important validity is-
sues. Future studies should, therefore, be designed to investigate
this:

• across several work-domains and with work-domain experts
  of different levels of computer and system experience (improving
  external validity),
• with other usability inspection methods (improving external
  validity),
• and additional measures of usability inspection performance
  complementing the measures of validity and thoroughness
  (improving construct validity).

Fairly high validity values were found in the present study com-
pared with studies reported in the literature. It may be the case
that the process structure provided by the applied usability inspec-
tion method, group-based expert walkthrough, represents a way
to increase the usability inspection validity for usability experts.
It would be interesting to investigate whether group-based expert
walkthrough has a more positive effect on the validity of the
usability inspection results generated by usability experts than
has heuristic evaluation or cognitive walkthrough.

Moreover, it would be interesting to see future studies that per-
mit extended investigations of the changes in performance, and
associated costs and benefits, when increasing the number of
work-domain experts participating as evaluators in usability
inspections. Possibly, this could be investigated both on the level
of nominal groups and the level of interacting groups. Since the
optimal number of evaluators may be affected by the work-do-
main, such research should include evaluations within different
work-domains.

Last, it is hoped that the results of the present study, along with
previous research on the importance of work-domain experts in
HCI work, will encourage the HCI research community to continue
to investigate the performance of work-domain experts in usability
inspections, as well as to continue to develop and investigate
usability inspection methods that support work-domain experts
as evaluators.

Acknowledgements

The work presented in this paper has been supported by the
Norwegian Research Council through the projects Efficiency
through user involvement (FIFOS programme) and RECORD (VER-
DIKT programme). We wish to thank Jan Håvard Skjetne (SINTEF)
for his help in the data collection, and Jan Heim (SINTEF), Pål Ulle-
berg (University of Oslo), and Line C. Gjerde (University of Oslo) for
their help in the analyzes.

References

engineering: a study of four companies that developed the same system. IEEE
measures and expert assessments. In: Canfora, G., Talvildari, L. (Eds.),
Proceedings of the 23rd International Conference on Software Maintenance,
based on use case points. In: Jeffery, R. (Ed.), Proceedings of the Fourth
Åborg, C., Sandblad, B., Gulliksen, J., Lif, M., 2003. Integrating work environment
considerations into usability evaluation methods – the ADA approach.
from an exploratory study among Nordic HCI practitioners. In: McEwan, T.,
Gulliksen, J., Benyon, D. (Eds.), People and Computers XIX – The Bigger Picture: