SINTEF

SINTEF ICT

Address: NO-7465 Trondheim, NORWAY
Location: Forskningsveien 1
Telephone: +47 22 06 73 00
Fax: +47 22 06 73 50

Enterprise No.: NO 948 007 029 MVA

SINTEF REPORT

TITLE

FLAMINKO: Usability evaluation of work-domain specific applications. Experiences and lessons learnt.

AUTHOR(S)

Asbjørn Følstad

CLIENT(S)

Norwegian research project, the VERDIKT program.

REPORT NO.	CLASSIFICATION	CLIENTS REF.	
SINTEF A11624	Open	NRC p. no. 176828/S10	
CLASS. THIS PAGE	ISBN	PROJECT NO.	NO. OF PAGES/APPENDICES
Open	978-82-14-04439-3	90B235.50	21/1
ELECTRONIC FILE CODE		PROJECT MANAGER (NAME, SIGN.)	CHECKED BY (NAME, SIGN.)
NA		Erik G. Nilsson	Jan Heim
FILE CODE	DATE	APPROVED BY (NAME, POSITION, SIGN.)	
NA	2009-06-26	Bjørn Skjellaug,	

ABSTRACT

Work-domain specific software solutions represent important challenges for usability evaluation methods; particularly related to the lack of context knowledge resources available to the usability experts involved in the evaluation. In order to meet this challenge we have investigated two methods where work-domain experts (end users with thorough experience of the context of use) are engaged as analysts in usability evaluations.

- Group-based expert evaluation is an analytical evaluation method designed to allow non-usability experts to participate as evaluators.
- Cooperative usability testing is a combined empirical and analytical method, designed to allow the participating user to reflect systematically on their experience with the test and the tested solution.

The two evaluation methods were investigated on basis of their application in two cases: (1) The redesign of a PDA application for mobile sales personnel and (2) the development of an in-vehicle solution for emergency response units. The methods' performance was investigated with regard to thoroughness, validity and impact – three well defined measures in the literature. The results from the two cases were consistent. Group-based expert walkthrough may be recommendable as a low-cost method for usability evaluation of work-domain specific solutions. However, the method missed important usability issues identified through empirical usability testing. Cooperative usability testing seemed in these cases to be a preferable alternative to regular usability testing, as the inclusion of user reflection served to increase the usability issues covered as well as positively affect the impact on the development.

The study also gave results of relevance also outside the scope of investigating the methods in question. In particular they serve to raise concern regarding the adequacy of the validity measurement as applied within the field of Human–computer interaction

KEYWORDS	ENGLISH	NORWEGIAN
GROUP 1	ICT	IKT
GROUP 2	HCI	Menneske-maskin interaksjon
SELECTED BY AUTHOR	Usability inspection	Analytisk evaluering av brukskvalitet
	Usability testing	Empirisk evaluering av brukskvalitet

TABLE OF CONTENTS

1 Introduction

Work-domain specific software solutions represent important challenges for usability evaluation methods; particularly related to the lack of context knowledge resources available to the usability experts involved in the evaluation. The context of a given solution is understood as its users, the users' tasks, and environment in which the solution will be used. Examples of work-domain-specific solutions, as seen in the FLAMINKO project, are solutions for ambulance personnel and mobile sales personnel.

The context challenge is clearly seen in usability inspections (e.g. heuristic evaluation, cognitive walkthrough), where experts aim to predict usability issues (usability problems or design suggestions) through analysis of a given solution. Typically, usability experts will need to engage in extensive context research or training in order to gain a sufficient understanding of the solutions context of use; implying costs which may be prohibitive for sufficient application of usability evaluation.

However, the context challenge is also relevant for empirical usability evaluation (e.g. usability testing), since the usability expert in charge of the evaluation will define the scope of the evaluation (through the test setup and -procedure) and thus limit which usability issues that may be predicted in the evaluation.

One promising approach to alleviate the lack of context knowledge resource available to usability experts is to involve work-domain experts as analysts. Work-domain experts are understood as end users with thorough experience of the context of use, or persons with extensive secondary knowledge of this (e.g. people that have achieved such knowledge through training, supporting, or studying end-users).

In the FLAMINKO project we have used two approaches to involve work-domain experts as analysts:

- Engaging context experts as evaluators in usability inspections
- Inviting participants in usability testing to engage in analysis of the test session and the tested solution

The experiences we have had with the methods represent new knowledge of relevance to the field of Human-Computer Interaction. The approaches also represent relevant new ways of conducting usability evaluations in the industry, in particular for developers of context-specific solutions.

In this report we will first present the two approaches to involving work-domain experts as analysts. Then we will formulate the research questions and hypotheses for the study, present the research method (including two evaluation cases), and the research results. Following this, we will give a discussion and summarize experiences and lessons learn. Finally we suggest future work.

2 Background: Users as analysts in usability evaluations

Usability evaluation methods may broadly be divided in empirical methods and usability inspections. Empirical methods typically involve observation of users as they solve predefined tasks in the software solution to be evaluated. Encountered problems or complaints from the user when solving the tasks are interpreted as usability problems. Usability inspections typically involve experts that walk through the solution or its documentation in order to predict usability problems. The outcome of usability evaluation methods is typically a list of usability issues, mainly including usability problems, but possibly also suggestions for redesign.

We have let users be involved as analysts in usability evaluations through two methods; one usability inspection method and one empirical method. Background for both of these is given below.

2.1 Work-domain experts in usability inspections: Group-based expert evaluation

Researchers and practitioners have experimented with including user representatives in usability inspections from the early nineties. Important contributions in this area include Bias' (1994) pluralistic walkthrough and Muller's (1998) participatory heuristic evaluation; both included end-users as evaluators in pluralistic evaluator teams.

Following the lead of Bias and McClard, we proposed the method Group-based expert walkthrough as a result of the UMBRA project (NRC, VERDIKT program). Group-based expert walkthrough is tailored to allow non-usability experts to participate as evaluators in usability inspections. It differs from the pluralistic walkthrough and participatory heuristic evaluation e.g. because it allows the work-domain experts to participate together with other work-domain experts only, and not in multidisciplinary teams.

General overview of the evaluation procedure: Group-based expert walkthrough is set up to predict usability issues associated with a set of task-scenarios predefined by a usability expert; typically the test leader. For each task-scenario the test leader first presents the steps required to solve the task. For each step the evaluators are required to make individual notes on what they perceive to be usability issues. Discussions are not permitted during the individual note taking. Upon completion of the individual notes for a given task, the individual issues are presented and discussed in the group consisting of all individual evaluators. The group decides on whether or not each proposed usability issue should be kept and (if kept) how it should be formulated. If agreement cannot be reached the disagreement is noted as a result. The discussions are structured by the test leader, but the input to the discussions is to come from the evaluators only. For each step of a given task scenario, issues regarded as *high-severity* by the evaluators proposing them are presented and discussed before issues regarded as *low-severity*. The procedure of the group based expert walkthrough is visualized in Figure 1. More details are given in the report from UMBRA (Følstad, 2006).

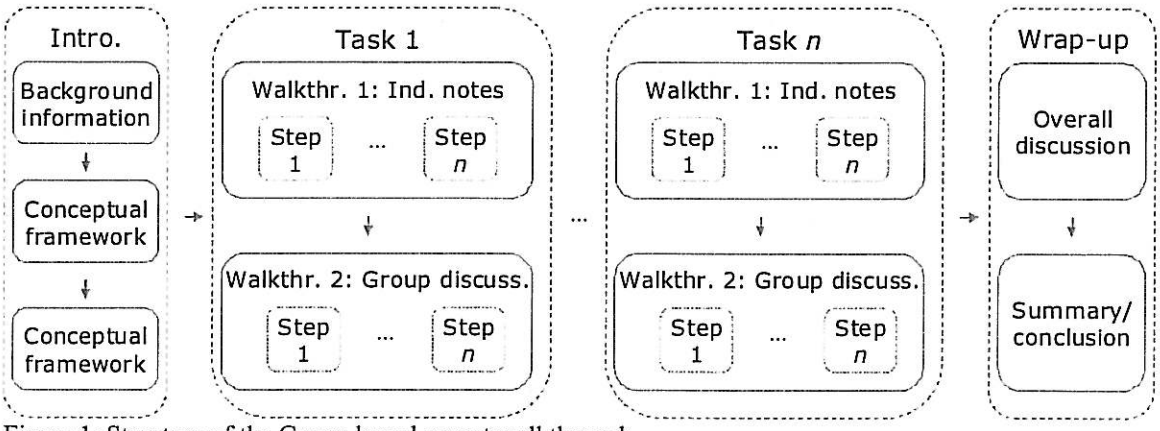


Figure 1: Structure of the Group-based expert walkthrough

Empirical experiences with the method: Work-domain experts' performance in group-based expert evaluation has been investigated in an early study targeting the impact of the evaluation results (Følstad, 2007). Impact is related to whether or not predicted usability issues were associated with changes in the software solutions. In the study, the downstream impact of context experts' evaluation results was compared to that of usability experts. In this study the work-domain experts were associated with a higher degree of impact than usability experts. However,

the usability experts predicted a higher number of usability issues. The usability experts and context experts generated highly different sets of usability issues; there was less than 20 percent overlap between the problem sets from the two evaluator categories.

2.2 User reflections in empirical evaluation: Cooperative usability testing

The main focus of usability testing have typically been observation of users, even though test protocols typically ask the participants to think aloud while doing the test tasks and also typically include short pre- and post- data collections by questionnaire or interview (Dumas & Fox, 2008; Rubin & Chisnell, 2008). The interpretation and analysis of the usability problems encountered by the test participants typically is conducted by usability experts.

One previous attempt to rethink the "observation only"-approach to usability testing was done by Frøkjær and Hornbæk (2005), in a study where they introduced an interpretation session following the interaction session of traditional usability testing. In the interpretation session, the test leader and the participant discuss what they consider the most important usability problems, supported by a video of the interaction session. This alternative set-up was termed Cooperative usability testing.

Overview of the evaluation procedure: Pursuing the idea introducing users' interpretation in usability testing, we included interpretation phases as part of the FLAMINKO usability testing protocols. Interpretation phases were included immediately following each task completion; the latter termed interaction phases. In the FLAMINKO cases, the test protocol included three to four tasks, and consequently also 3-4 interchanging interaction and interpretation phases. In the interaction phase the participant completed the task while thinking aloud, as in regular usability testing. In the interpretation phase, the test leader and the participants together went through the task as it had been done by the participant in the preceding interaction phase. In places where the participant had experienced problems or expressed concern, the test leader initiated a discussion on the participant's experience of the situation, the possible causes of the situation, and possible design changes that could remedy the predicted usability problems. In the interpretation phase, the test leader also asked the participant predefined questions related to perceived usefulness and user experience. The structure of our Cooperative usability testing, with interchanging interaction and interpretation phases, is presented in Figure 2. Since our test procedure includes several interaction and interpretation phases, instead of only one of each, this is a modification of Frøkjær and Hornbæk's Cooperative usability testing. Even so, we use this term to represent our version of this method in this report.

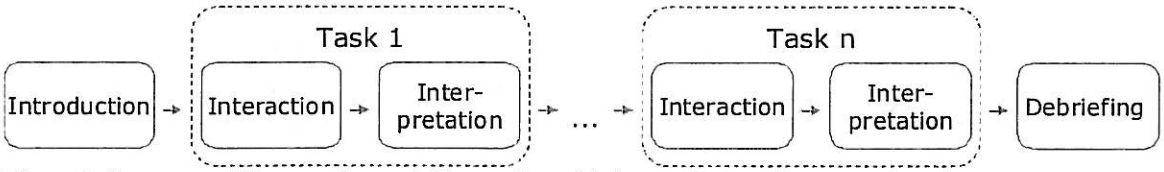


Figure 2: Structure of Cooperative usability testing with interaction and interpretation phases

3 Research questions

The main research questions targeted in our work on usability evaluation were:

1. What is the performance of group based expert evaluations with work-domain experts as evaluators? Performance was understood as:
    a. Number of predicted usability issues (compared to the interaction and interpretation phases Cooperative usability testing)
    b. Thoroughness and validity (relative to the interaction phases of Cooperative usability testing)

c. Impact (compared to the interaction and interpretation phases of Cooperative usability testing)

2. How does the inclusion of interaction phases affect the performance of usability testing? Performance was understood as:
   a. Number of predicted usability issues added in consequence of the interpretation phases
   b. Impact of usability issues of the interpretation phases (compared to the interaction phases)

3.1 What is the performance of group based expert evaluations with work-domain experts as evaluators?

In the study, we wanted to investigate the performance of Group-based expert walkthrough from different perspectives: (a) Raw counts of usability issues, (b) Thoroughness and validity and (c) impact.

Raw counts of usability issues was a much used way to compare evaluation methods in the nineties. However, more recently it has been strongly discouraged to use this alone as a performance measure (Cockton, Lavery, & Woolrych, 2003). We believe, however, that an investigation of this aspect of evaluation performance provides important background in order to understand the measures of thoroughness, validity, and impact.

Thoroughness and validity are two important approaches to the study of evaluation method performance (Hartson et al., 2001). A thorough evaluation method will predict a high proportion of all real usability problems associated with the object of evaluation. A valid evaluation method will generate a low proportion of false alarms (falsely predicted usability problems). The formulas are given below.

$$Thoroughness = \frac{\text{Number of predicted real problems}}{\text{Number of real problems that exist (including non-predicted problems)}}$$

$$Validity = \frac{\text{Number of predicted real problems}}{\text{Number of predicted problems (including false alarms)}}$$

The notion of real problems, however, is problematic. In research on the validity usability evaluation methods, these are typically defined as the set of usability issues observed in usability tests (corresponding to the interaction phases of Cooperative usability testing). However, it is conceivable that this set of usability issues does not cover all possible usability issues associated with the evaluated system. To get more insight in this regard, we also wanted to investigate (a) the nature of the false alarms found, and also (b) the proportion of usability issues generated by group-based expert walkthrough that were judged as *not tested* against the set of real problems.

Impact (Sawyer et al., 1996) is another, and possibly more controversial, approach to the study of evaluation method performance. In this paper impact is understood as the proportion of predicted usability issues (problems or suggested redesigns) that are associated with subsequent changes in the software solution.

$$Impact = \frac{\text{Number of usability issues associated with actual change}}{\text{Number of predicted usability issues}}$$

3.2 How does the inclusion of interpretation phases affect the performance of usability testing?

We wanted to investigate the results generated in the interpretation phases of Cooperative usability testing from two perspectives: (1) raw counts of usability issues and (2) comparing the impact of the usability issues generated in the interpretation and interaction phases respectively. These are the same perspectives as for our investigation of the group-based expert walkthrough, without the perspective of thoroughness and validity.

The perspective of thoroughness and validity was not applied for this part of the investigation. The reason for this was that our working definition of "real problems" was the set of problems generated in the interpretation phases, and that the results of the interaction phases and the interpretation phases were seen as not overlapping.

4 Method

Group-based expert evaluation and Cooperative usability testing were used in two separate cases, related to software solutions for ambulance personnel and mobile sales personnel respectively.

4.1 Case details

4.1.1 Case 1

The object of evaluation was a running version of a business application for PDAs. The application should support sales personnel in the field when placing customers' orders and requesting goods. The application runs on Windows mobile, and utilizes much of the standard UI elements from this operating system.

The participants in the Group-based expert evaluations and Cooperative usability testing respectively were all users of the system, and had been so for the last 0.2-1 year. Each participated in only one of the methods. All had been employed by the company where the application was used for at least two months (Group-based expert evaluations: *Median*=7.5 years, *Min*=1, *Max*=13. Cooperative usability testing: *Median*=0.9 years, *Min*=0.2, *Max*=20). Age range was 19 through 46 years (Group-based expert evaluations: *Median*=39 years, *Min*=24, *Max*=46. Cooperative usability testing: *Median*=29.5 years, *Min*=19, *Max*=45).

The Group-based expert evaluations included two sessions, each with three evaluators. The Cooperative usability testing included eight sessions with one participant in each.

4.1.2 Case 2

The object of evaluation was a UI prototype of an in-vehicle application to be used by ambulance personnel for receiving missions and messages, reporting mission status, and navigation. The prototype was to become the next version of an application that was already in use by the participants. It was made as a clickable PowerPoint presentation, and presented on an 8 inch touch screen; identical to the screen the participants had in their vehicles.

The participants in the Group-based expert evaluations and Cooperative usability testing respectively were all users of the predecessor of the prototype under evaluation, and had been so

for the last 0.2-1 year. Each participated in only one of the methods. All had been employed as ambulance personnel for at least one year (Group-based expert evaluations: *Median*=7 years, *Min*=2, *Max*=20. Cooperative usability testing: *Median*=11 years, *Min*=1, *Max*=34). Age range was 20 through 60 years (Group-based expert evaluations: *Median*=40 years, *Min*=26, *Max*=46. Cooperative usability testing: *Median*=42 years, *Min*=20, *Max*=60).

Group-based expert evaluations included two sessions, with five and four evaluators respectively. The Cooperative usability testing included nine sessions with one participant in each.

4.2 Evaluation methods

4.2.1 Usability inspection: Group-based expert walkthrough

In both cases the usability inspections were conducted according to the method group-based expert walkthrough, as described in Section 2.1. Each session lasted 2.5 hours. The sessions were video recorded.

Each of the evaluation sessions resulted in one set of usability issues from the group discussions as well as one set from each of the individual evaluators. All usability issues, both individual issues and group issues, were required to be associated with a severity rating. The severity ratings of individual issues were decided by the individual evaluator; the severity ratings of group issues were decided by the group. The following severity ratings were employed: *cosmetic* (minor obstacles or sources of irritation), *serious* (major obstacles or sources of irritation), and *critical* (insurmountable obstacles or sources of irritation).

4.2.2 Empirical usability evaluation: Cooperative usability testing

In both cases empirical usability evaluation was conducted with Cooperative usability testing, with interchanging phases of interaction and interpretation, as described in Section 2.2. Each session lasted between 20 and 60 minutes, depending on the amount of feedback provided by the participants. The sessions were video recorded. The video captured both the on-screen action and the participant dialogue with the test leader.

Following the tests, the videos were analyzed in order to establish two sets of usability issues. One "interaction set" predicted on basis of user behaviour in the interaction phases, and one "interpretation set" predicted only in the interpretation phase. The interaction phases were analysed by two individual analysts. The interpretation phases were analysed by one analyst (the test leader).

All usability issues of the "interaction set" were associated with a severity rating. Severity ratings on the level of the individual participants' usability problems were decided by the analysts. The severity rating of each general usability issue was decided by the following criteria:

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

The usability issues of the interpretation set were not associated with a severity rating. The reason for this is that we judged it to be possibly interfering with the interpretation dialogue to require severity ratings from the participant for every usability issue.

4.3 Analyses

4.3.1 Raw counts of usability issues

The results provided by the different evaluation methods were matched, in order to investigate the degree of overlap.

Matching of the results from Group-based expert walkthrough and the results form the interaction phases of the Cooperative usability testing were done by two analysts. Matching of the results from Group-based expert walkthrough and the results form the interpretation phases of the Cooperative usability testing were done by one analyst. Since any input in the interpretation phases that only repeated what was already known from the interaction phases (e.g. only stating that an UI element represented a problem when solving the task) was not included in the results from the interpretation phases, the "interaction set" and "interpretation set" of results from the Cooperative usability testing therefore only had a few instances of overlap[1] and did not require matching.

The results from the different methods were summarized as Venn diagrams to illustrate the degree of overlap.

4.3.2 The thoroughness and validity of usability inspection with work-domain experts

When investigating the thoroughness and validity of usability inspection results, the results need to be compared to a trustworthy yardstick. Within the field of HCI, the results from empirical usability testing are typically used as a yardstick against which usability inspection results may be compared.

In order to conduct the comparison, the sets of usability issues predicted in the Group-based expert walkthrough were matched with the set of usability issues from the interaction phases of the Cooperative usability testing. The matching procedure served to classify predicted usability issues in four categories:

1. Hits: Group-based expert walkthrough issues found to match Cooperative usability testing issues
2. Misses: Cooperative usability testing issues not predicted by Group-based expert walkthrough
3. False alarms: Group-based expert walkthrough issues that – given the prediction was correct - should have affected participant performance in the interaction phases of Cooperative usability testing, but did not
4. Not tested: Group-based expert walkthrough issues that could not be expected to affect participant performance in the interaction phases of Cooperative usability testing.

All usability issues classified as *false alarms*, and example issues classified as *not tested* are presented in Sections 5.3 and 5.4.

The matching was conducted by two independent analysts. Inter-analyst agreement was found to be Kappa=0.51 in Case 1 and Kappa=.49 in Case 2; corresponding to what is termed *moderate agreement* by Landis and Koch (1977) In order to calculate Cohen's Kappa expected probability was calculated on the assumption that the analyst for each item had 3 equally likely alternatives - hit, false alarm, and not tested.

[1] Across the two cases, five issues were overlapping between the two phases of Cooperative usability testing. The reason for this overlap was that these usability issues were identified in the interaction phases by some test participants and in the interpretation phases by other test participants.

Finally, all instances of false alarms were scrutinized in order to gain an improved understanding of the nature of false alarms relative to a real problem set established through the observation of users in usability testing.

4.3.3 Impact of usability issues

All usability issues resulting from Group-based expert evaluation and Cooperative usability testing group were compiled in two evaluation reports, for Cases 1 and 2 respectively.

On basis of the reports, development team representatives rated the priority of each usability issue. In Case 1 the prioritizing was done by four development team members together, in Case 2 by the project leader alone. The following priority ratings were used: *high* (change will be conducted before a given case-specific deadline); *medium* (change is relevant, but not before the given deadline); *low* (change will not be prioritized); *wrong* (the developer disagrees or the issue is not the developer's responsibility).

The different sets of usability issues were compared with regard to impact.

5 Results

5.1 Raw counts of usability issues

In Case 1 a total of 53 usability issues were predicted by the two methods. Group-based expert walkthrough predicted 38 issues. Within the Cooperative usability testing, the interpretation phases predicted 18 issues, and the interaction phases 16. Due to two instances of overlap, the total number of usability issues generated by Cooperative usability testing was 32. The reason for this overlap was that these usability issues were identified in the interaction phases by some test participants and in the interpretation phases by other test participants.

In Case 2 a total of 64 usability issues were predicted. Group-based expert walkthrough predicted 38 issues. Within the Cooperative usability testing, the interpretation phases predicted 31 issues, and the interaction phases 12. Due to two instances of overlap, the total number of usability issues generated by Cooperative usability testing was 40.

The distribution of the predicted usability issues across the methods and method phases are presented in Figure 3.

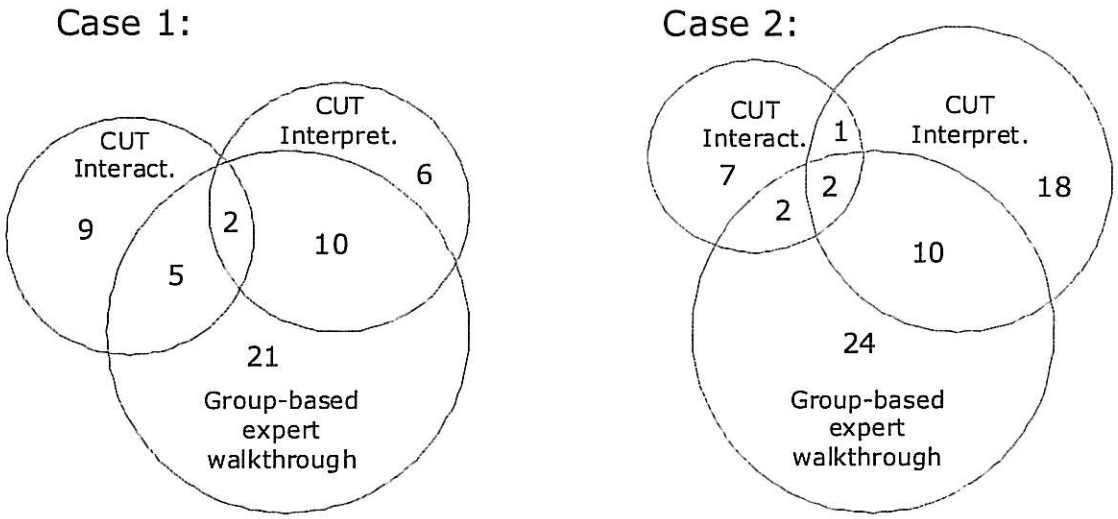


Figure 3: Distribution of predicted usability issues across the Group-based expert walkthrough and the interaction and interpretation phases of Cooperative usability testing.

We also investigated the distribution of the predicted usability issues across severity ratings (cosmetic, serious, critical). These distributions are presented below in Table 1 and 2 for case 1 and 2 respectively. It may be noted that for both methods, the proportion of cosmetic issues predicted is below 40 percent. It may also be noted that the issues overlapping between methods do not seem to be of any particular severity category – cosmetic issues may just as well be overlapped between methods as severe or critical issues.

	Group-based expert walkthrough				CUT (Interaction phases)			
	Total		Overlap		Total		Overlap	
	Freq.	%	Freq.	%	Freq.	%	Freq.	%
Critical	10	26	2	29	4	25	4	57
Serious	19	50	2	29	2	13	0	0
Cosmetic	9	24	3	43	10	63	3	43
Sum	**38**	**100**	**7**	**100**	**16**	**100**	**7**	**100**

Table 1: Distribution of Case 1 predicted usability issues (total and overlapping) across severity categories

	Group-based expert walkthrough				CUT (Interaction phases)			
	Total		Overlap		Total		Overlap	
	Freq.	%	Freq.	%	Freq.	%	Freq.	%
Critical	3	8	0	0	5	42	2	50
Serious	23	62	3	75	1	8	0	0
Cosmetic	7	19	1	25	6	50	2	50
Not rated	4	11	0	0	0	0	0	0
Sum	**37**	**100**	**4**	**100**	**12**	**100**	**4**	**100**

Table 2: Distribution of Case 2 predicted usability issues (total and overlapping) across severity categories

As described in 4.2.2, the usability issues predicted in the interpretation phases of the Cooperative usability testing were not associated with severity ratings.

5.2 Thoroughness and validity scores for Group-based expert walkthrough

Thoroughness and validity scores for Group-based expert walkthrough were calculated on basis of the classification of the predicted issues as hits, misses, false alarms, or not tested. See Section 4.3.2 for definitions of the classifications. Formulas for the scores are provided in Section 3.1.

The Group-based expert walkthrough was found to have a thoroughness score of .44 in Case 1 and .33 in Case 2. This means that in the two cases, 44 and 33 percent of the usability issues predicted in the interaction phases of the Cooperative usability testing were also predicted in the Group-based expert walkthrough. The results used to calculate the scores are presented in Table 3.

	Case 1	Case 2
Hits	7	4
Misses	9	8
Issues predicted in the CUT interaction phases	**16**	**12**

Table 3: Distribution of hits and misses in Case 1 and 2, used to calculate thoroughness scores

The same method was found to have a validity score of .70 in Case 1 and .50 in Case 2. This means that in the two cases, 70 and 50 percent of those usability issues in the Group-based expert walkthrough that were measured against the yardstick of empirical usability testing were actually observed. The results used to calculate the scores are presented in Table 4.

	Case 1	Case 2
Hits	7	4
False alarms	3	4
GBEW-issues tested against issues predicted in the CUT interaction phases	**10**	**8**

Table 4: Distribution of hits and false alarms in Case 1 and 2, used to calculate validity scores

5.3 Are all false alarms really false alarms?

The false alarms generated in both cases were scrutinized in order to help us answer the following: What does a false alarm look like? And are false alarms really false alarms, just because they did not appear during usability testing?

Due to the fairly low number of false alarms, the translated false alarms of both cases are presented and discussed below.

Case 1: False alarms

Three false alarms were identified in Case 1. These are presented in Table 5. Associated screen shots are presented in Figure 4.

1.1	The size of some screen elements (the menu choices in the lower part of the screen and the arrows in drop-down boxes) is too small and hard to hit. *(Figure 4. Explanation - small menu choices: Placing orders was the key functionality of the application. The first step in placing any new order was to click "New" ("Ny") located at the left of the bottom menu. This menu item had the standard size for Windows mobile. Explanation – small arrows: The drop-down boxes used in the application were the standard drop-down boxes of Windows mobile. These were regarded as too small for practical use in the context of the mobile salesperson, in particular if the stylus was not available.)*
1.2	The menu item "New" does not explicate what will be created. Want "New order" instead. *(Figure 4, left. Explanation: The only visual cue of the first step towards creating a new order was the menu item "New". This was thought to be insufficient for beginners or irregular users.)*
1.3	It is too easy to accidentally click "Suggestions", because this is in the same place as the item "New" on the previous page. *(Figure 4, right. Explanation: When placing internal orders to fill up the car with goods, "Suggestions" ("Forslag") provided an automatic filling in of the order based on what has previously been taken out. Most users did not want to use this functionality, and it was regarded as cumbersome to remove the auto-generated suggestion if it was requested by accident.)*

Table 5: Case 1 false alarms

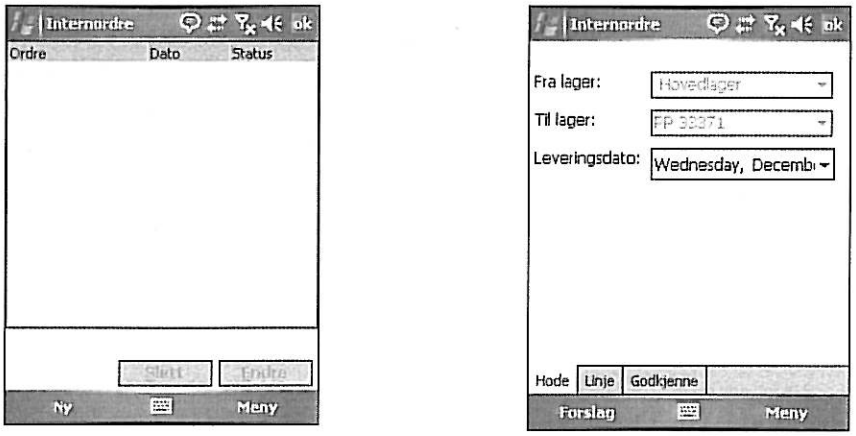


Figure 4: Screen shots associated with false alarm 1.1 (both), 1.2 (left), and 1.3 (right)

A false alarm is defined as a prediction in the usability inspection that should be expected to affect participant performance during the interaction phases of Cooperative usability testing, but did not. None of the three false alarms did appear. However, it is relevant to question whether this really justifies them being labelled false alarms.

It is quite conceivable that 1.1 and 1.2 actually may occur in the field – e.g. if the device is used while the user is busy with loading goods, driving, or talking to a customer. This is so, even though these were not registered as problems in the usability test. It may be argued that a field study would be a better validation criterion for these three than our out-of-context usability testing. However, even if we had conducted a field study and the problems had not appeared, would we then be certain that these really were false alarms? Maybe the problems appear only once in a while; often enough to be annoying to the user but too seldom to be expected to appear in a field study.

Similarly, 1.3 may be a relevant usability issue, even though we have classified it as a false alarm. Maybe this issue is only relevant for a few novice users, which were not included in the test? Or maybe, even if this issue was never to appear as an observable usability problem, neither in the field nor in the lab, we could still be justified in judging this suggestion from the work-domain experts as useful feedback in order to tailor the solution according to the users' wants.

Case 2: False alarms

Four false alarms were identified in Case 2. These are presented in Table 6. Associated screen shots are presented in Figure 5.

2.1	Difficult to notice message notification if there is an A (high priority) in mission description. *(Figure 5, left. Explanation: The UI prototype provided the users with missions and messages. Missions were related to current emergency incidents. Messages were general information. The current mission, new mission notifications, and new message notifications were visualized in the main screen of the application. The main screen contained a navigator map, top and bottom menus, and new message/mission warnings. The current mission was presented in the middle field of the top menu. New message notifications were presented as floating over the map. High priority missions were not meant to be dependent on message notifications, thus - contrary to the evaluators understanding - it was not important that new messages were noticed during high priority missions.)*
2.2	The new message notification box is too small. *(Figure 5, left. Explanation: Not only was the message notification regarded as difficult to notice in certain circumstances, but the box of the message was also considered to be too small in general)*
2.3	It is unfortunate that non-active buttons are visible *(Figure 5, middle. Explanation: In the UI prototype, menu items that referred to future functionality not to be implemented in the version under development were presented as not active (darker fill colour).)*
2.4	There might be a problem to write on the on-screen keyboard *(Figure 5, left. Explanation: The keyboard in the prototype was only used for entering search queries. It was not clarified whether this keyboard also would be used for more extensive text input.)*

Table 6: Case 2 false alarms

Figure 5: Screen shots associated with false alarm 2.1 and 2.2 (left), 2.3 (middle), and 2.4 (right)

In Case 2, three of the four false alarms may be judged to be credible, even though they were not identified as usability issues in the interaction phases of Cooperative usability testing. (From the explanation of 2.1 given in Table 6, this usability issue seems to be the consequence of the evaluators not fully understanding the purpose of the new message notification, and may thus be regarded as having low credibility.)

It is conceivable that 2.2., and 2.4 could have occur in the field. Maybe messages would be difficult to notice because the new message notification box was too small? And maybe for some users it may be difficult to write even short texts on the keyboard in the context of a fast moving ambulance? This we would not be able to find out unless we actually built and installed these parts of the system as it is presented in the visual prototype. And, as in Case 1, even if the false alarms did not appear in the field, would we then be fully satisfied that these were really false alarms?

Regarding 2.3, even if it may be less conceivable that the visibility of non-active buttons will lead to significant obstacles for the user, would we not – as in Case 1 – be justified to judge this as a useful suggestion from the work-domain experts participating as evaluators?

From this short presentation and discussion of the false alarms identified in Case 1 and 2, it should be clear that the labelling of a prediction made by a work-domain expert as a false alarm may not be as straight forward as suggested by the approach to investigating validity as is commonly used within the field of HCI. In total across the two cases, 7 out of 8 false alarms may be argued to be credible.

5.4 Usability inspection results not targeted by the validity investigation

The concepts of validity helped us investigate the quality of the usability inspection results from Group-based expert walkthrough. But, a significant part of the usability issues predicted through Group-based expert walkthrough were judged to be of a kind that could not be expected to affect participant performance during the interaction phases of Cooperative usability testing (see Section 4.3.2). This was so, even if the usability testing and Group-based expert walkthrough were conducted with similar participants on exactly the same tasks.

In the analysis of the results, two independent analysts classified all the results from Group-based expert walkthrough either as *hits*, *false alarms* or *not tested*. In this section we will present the results classified as not tested. These are results not targeted by the concepts of validity or thoroughness, since they are not associated with a trusted yardstick against which they may be compared; the nature of the usability testing tasks simply did not allow us to judge the realness of these usability issues.

The distribution of the not tested usability issues is presented in Table 7.

	Case 1		Case 2	
	Freq.	%	Freq.	%
Tested (hits and false alarms)	10	26	8	21
Not tested	28	74	30	79
Sum	**38**	**100**	**38**	**100**

Table 7: Distribution of not tested vs. tested usability issues predicted by GBEW

We see that in both cases, the vast majority of the usability issues predicted by Group-based expert walkthrough were of such a kind that they were judged by the analysts to be not tested. Examples of such items are presented in Table 8. Associated screen-shots are presented in Figure 6.

1.a	The application runs too slowly. Need a more powerful device. *(Explanation: Using several of the functions of the application had the response time of several seconds. Data transfer to a main database or a printer could take several minutes.)*
1.b	The symbol "..." is difficult to understand for novices. Should be "Search" instead. *(Figure 6, left. Explanation: In order to make a search for particular goods or customers, the user was to click a button labelled "..."after entering the search string.)*
1.c	The price presented on-screen does not include customer-specific discounts. Discounted price should also be displayed. *(Figure 6, left. Explanation. When entering a new order, price information was presented in the field labelled "Pris". Customer-specific discounted price – if relevant – was not visible.)*
2.a	It is hard to see the mission reference number with a quick glance. Need larger font. *(Figure 6, middle. Explanation: In the mission specification view, the mission reference number - presented to the right of the "A" icon – was much used and should be easily accessible, also when looking at the screen from a distance, e.g. in the back of the vehicle.)*
2.b	Information on who reported the incident is not presented. This information may be useful at times. *(Figure 6, middle: The mission specification presents all information directly relevant to the mission. The evaluators discovered that information on the person reporting the incident was missing from the mission specification.)*
2.c	It is not good if it is possible to skip a status by mistake. Need confirmation prior to skipping a status. *(Figure 6, left. During a mission, status is changed 5-6 times. These status changes are required to be reported as they happen and cannot be cancelled after they are made. It may, however, happen that the personnel forget to update status, and need to skip a status in order to report the next status change.)*

Table 8: Examples of usability issues judged as not tested in Case 1 (top) and Case 2 (bottom).

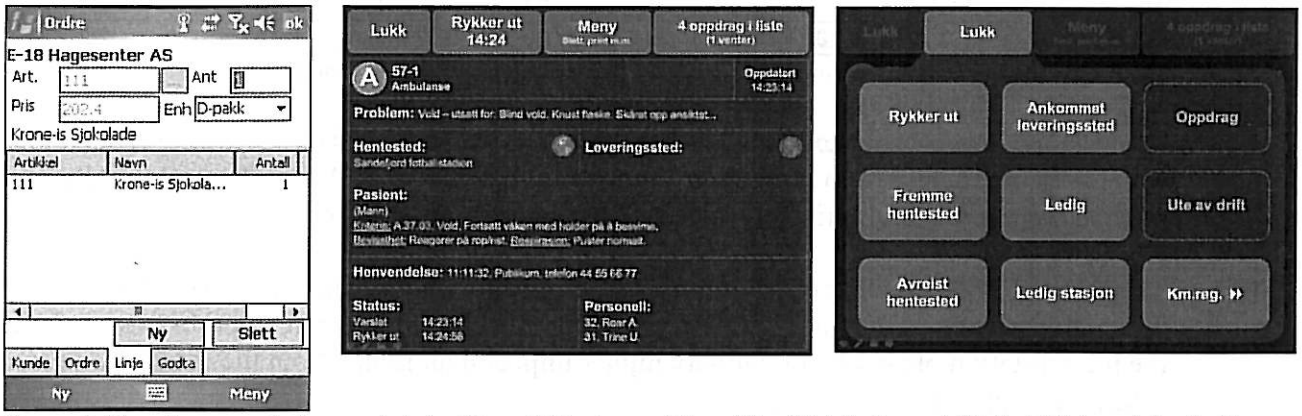


Figure 6: Case 1 screen shots associated with usability issues 1.b and 1.c (right), 2.a and 2.b (middle), and 2.c (left).

All the usability issues of Table 8 may be argued to be relevant input in the development process, even though we were not able to investigate their quality through the concept of validity.

In Case 1, the feedback that the application is perceived as too slow (1.a) is important feedback to the developers as it targets an issue highly relevant for the users' satisfaction. The symbol "..." may easily be replaced with "Search" (1.b), which also seems a more conventional label. And the left screenshot in Figure 6 may if needed, include information on discounted price (1.c) in addition to list price, provided that this information is available in the client-side database.

In Case 2, the much used and highly important mission reference number may easily be made more visible (2.a). It is also important feedback from the work-domain experts that information on reporting person is regarded as relevant information to display in the mission specification screen (2.b). Similarly, it is useful feedback that users may want a confirmation dialogue in cases of skipping status updates in the status update menu (2.c).

The two cases seem to illustrate that even usability issues from the Group-based expert walkthroughs that are not investigated for validity may contain valuable feedback to the developers. It may also be argued that we would have been hard pressed to device a usability testing procedure that would have predicted all of the issues presented in Table 8.

5.5 Impact of the usability issues

The impact of the usability issues predicted by the different methods was investigated separately for both cases. The distribution of the usability issues with respect to impact are presented in Table 9 and 10.

Case 1

Impact	Expert walkthrough		CUT total		CUT Interaction		CUT Interpretation	
High	9	25 %	6	21 %	2	13 %	4	31 %
Medium	12	33 %	13	46 %	8	53 %	5	38 %
Low	8	22 %	3	11 %	3	20 %	0	0 %
Not relevant	7	19 %	6	21 %	2	13 %	4	31 %
Sum	**36**	**100 %**	**28**	**100 %**	**15**	**100 %**	**13**	**100 %**

Table 9: Case 1 - Distribution of usability issues across method and impact categories

Case 2

Impact	Expert walkthrough		CUT total		CUT Interaction		CUT Interpretation	
High	13	42 %	16	42 %	7	58 %	8	31 %
Medium	7	23 %	6	16 %	2	17 %	5	19 %
Low	7	23 %	13	34 %	3	25 %	10	38 %
Not relevant	4	13 %	3	8 %	0	0 %	3	12 %
Sum	**31**	**100 %**	**38**	**100 %**	**12**	**100 %**	**26**	**100 %**

Table 10: Case 2 - Distribution of usability issues across method and impact categories

We see that Group-based expert walkthrough and Cooperative usability testing fared about equally well with regard to impact. In Case 1 slightly more than 1/5 of the usability items of either method were given high priority. In Case 2 the same number for either method was 2/5.

The cases provide no clear cut evidence with regard to whether usability issues from the interaction phases or the interpretation phases provided the highest impact. In Case 1, issues from the interpretation phases had relatively higher impact than issues from the interaction phases. In case 2, the opposite was the case.

It may be noted that the number of usability items receiving match scores does not match the total number of items predicted by each method. This was the cause of a clerical error where 6 of the usability issues predicted in Case 1 and 9 of the issues in Case 2 were not included in the reports delivered to the customers. Consequently no impact scores were provided for these.

In consequence the impact analysis is conducted on a dataset that is less complete than we aimed for. However, we still believe that these results provide relevant insight in the two cases.

6 Discussion

The discussion will be structured according to the two main research questions.

6.1 What is the performance of Group-based expert evaluations with work-domain experts as evaluators?

The performance of Group-based expert walkthrough was investigated from the perspectives of raw counts of usability issues, thoroughness, validity and impact.

6.1.1 Raw counts of usability issues identified by Group-based expert walkthrough.

Group-based walkthrough generated about the same number of usability issues as did Cooperative usability testing. This held true for both cases. Also, the majority of usability issues identified by Group-based expert walkthrough were not identified through Cooperative usability testing.

These findings seem to suggest that work-domain experts participating as evaluators in Group-based expert walkthrough will make a substantial contribution in addition to what is achieved through Cooperative usability testing.

In particular this is the case with respect to the interaction phases of Cooperative usability testing. Less than 15 percent of the usability issues predicted in Group-based expert walkthrough of the two cases were predicted through these interaction phases. In total, 10 critical and 38 serious usability issues predicted in the Group-based expert walkthrough were not predicted in the interaction phases. The results clearly suggest that work-domain experts participating as evaluators provide far richer results than work-domain experts participating just as users to be observed in usability testing.

6.1.2 The thoroughness of Group-based expert walkthrough

Thoroughness scores were in the area of 30-40 percent. In plain words this means that Group-based expert walkthrough failed to predict the majority of usability issues predicted through the interaction phases of Cooperative usability testing. This is somewhat lower than might be expected with usability evaluations conducted by several usability professionals (see e.g. Cockton et al., 2003); however comparisons of thoroughness scores across studies should be done only with great care as differences in the study setup may affect thoroughness.

Table 1 and 2 indicates that the severity of the usability issues do not affect which items that are predicted both in usability testing and Group-based expert walkthrough. In total, 4 critical and 8 serious usability issues predicted in the interaction phases of the Cooperative usability testing were not predicted in Group-based expert walkthrough, meaning that if we conducted only Group-based usability testing, these important issues would have been left unpredicted. This failure to predict important usability issues uncovered by observation in usability testing is an important challenge to Group-based expert walkthrough, and may suggest that the method should not be used isolated from usability testing.

6.1.3 The validity of Group-based expert walkthrough

The validity scores for the two cases were in the area of 50-70 percent. A rate of false alarms of about 50 percent seems to be too high, even though we know from the HCI literature that validity scores in the area 40-60 percent are typical in studies of usability inspection methods (Cockton et al., 2003).

However, when examining the false positives one by one it seemed fair to argue that all but one of the false positives seemed likely to represent useful feedback to the developers, even though they were not predicted on basis of observing actual users in the Cooperative usability test. This examination of the false positives lead us to conclude that the Group-based walkthrough seem to be trustworthy with respect to validity, in spite of the achieved validity scores.

Our findings related to validity have potential implications also outside the study of this method. We have presented two cases where most of the false positives made by the usability inspection method under evaluation may be argued to be credible feedback to the developers. Also the cases provide us with an example where validity as investigated with an empirical usability testing as a yardstick for comparison, do not account for the majority of the findings from the method under

evaluation; about 3/4 of the usability issues predicted through Group-based expert walkthrough were classified as "not tested" against the yardstick of empirical usability testing. Based on our experiences presented here, empirical usability testing may not be a sufficient yardstick against which other usability evaluation methods can be measured. If this is the case, this will shed new light on previous research on the validity of usability inspection methods, where results from empirical usability testing have been used as the yardstick for comparison. And the findings may serve as inspiration for future research on the validity of usability inspection methods to look elsewhere for such a yardstick.

6.1.4 The impact of Group-based expert walkthrough

The impact of the usability issues predicted in Group-based expert walkthrough may be seen from the perspective of absolute values and the perspective relative to Cooperative usability testing.

From the perspective of absolute impact, about 3/5 of the reported usability issues predicted through Group-based expert walkthrough were given high or medium priority. The impact scores may be taken to indicate that the majority of the predicted usability issues were perceived by the developers as relevant for their development process.

From the perspective of relative impact, the impact of Group-based expert walkthrough was highly similar to the impact of Cooperative usability evaluation. For both methods, in both cases, about 3/5 of the reported usability issues were given medium or high priority. This similarity should be seen in the light of the relative low overlap between the methods; the majority of the usability issues predicted through Group-based expert walkthrough were not predicted through Cooperative usability testing. Thus, the similarity in impact cannot be said to be solely a consequence of similar usability issues – rather, the similarity in impact seem to have appeared in spite of diversity in the usability issues predicted by both methods.

The presented impact of both methods is judged to be good, in particular with respect to the two cases being real development cases where a range of other constrains than user requirements apply; e.g. cost, time, and resource limitations.

6.2 How does the inclusion of interpretation phases affect the performance of usability testing?

Traditionally, usability testing is conducted as observations of users conducting tasks with the solution under evaluation. In addition to the procedure related to the task scenarios in which the user is observed, the testing procedure may include pre-test or post-test data collection of the participants' attitudes or preferences. To some degree, questions may be included to clarify incidents during the task completion, either en-route or following task completion.

In Cooperative usability testing, the participants are systematically engaged in analysis or interpretation. In our adaptation of Cooperative usability testing, an interpretation phase is included following each interaction phase, and an interaction phase consists of the completion of one task scenario.

We have investigated how these interpretation phases affect the performance of usability testing via raw counts of usability issues as well as impact. The results from the two phases have not been investigated with regard to thoroughness and validity, since there are no overlapping issues between these two phases.

6.2.1 Raw counts of usability issues identified in the interpretation phases

The distribution of predicted usability issues across methods presented in the Venn diagrams of Figure 3 shows that the interpretation phases generated at least the same number of usability issues than did the interaction phases; in Case 1 the two phases generated the same number of issues whereas in Case 2 the interpretation phase generated more than double the number of usability issues as compared to the interaction phase.

As we concluded for Group-based expert walkthrough, engaging the work-domain experts as knowledge resources and discussion partners in the interpretation phases of Cooperative usability testing seem to greatly improve the richness of the evaluation; the interpretations phases add a substantial number of usability issues to the evaluation results provided by only observation of users in usability testing.

6.2.2 The impact of usability issues identified in the interpretation phases

The analyses of impact provide interesting insight. Proportionally, the results from the interaction phase had higher impact in Case 2 and the results from the interpretation phase had higher impact in Case 1. In consequence, we do not seem to be able to conclude with regard to which method can be expected to be associated with the higher impact score in future studies.

However, in both cases the number of usability issues with medium or high impact was higher for the results of the interpretation phase than the results of the interaction phase. In consequence, if we were to take away the results associated with the interpretation phases, we would remove the majority of usability issues associated with substantial impact.

In conclusion, the inclusion of interpretation phases seems to have been valuable, both with regard to the number of usability issues generated and with regard to the total impact of the evaluation results.

7 Conclusion

7.1 The value of involving work-domain experts as analysts

In this study we have investigated the effect of involving work-domain experts as analysts in usability evaluation methods. The work-domain experts were involved as evaluators in usability inspections and reflecting participants in Cooperative usability testing.

The results from the two cases seem to sustain our initial assumption that involving work-domain experts as analysts may be useful when evaluating work-domain specific software solutions. The sheer volume of predicted usability issues increases substantially; in either of the studied cases less than 1/3 of the identified usability issues were identified through observation of users in usability testing. Also, the usability issues based on work-domain experts as analysts seemed to have similar impact on the development process compared to the usability issues based on observations in usability testing.

In conclusion, engaging work-domain experts as analysts through Cooperative usability testing was a valuable extension to traditional usability testing in the cases of this study. The addition of interpretation phases to the usability testing procedure made the output of the evaluation comparable to that of Group-based expert walkthrough both with regard to volume and impact.

With regard to work-domain experts used as evaluators in Group-based expert walkthrough, the relatively low thoroughness scores indicates that it may be wise to use Group-based expert

walkthrough in combination with usability testing, in order not to miss critical empirically predicted usability issues.

7.2 Method recommendations

On basis of the experiences from this study, we wish to make recommendations regarding the studied evaluation methods. Please note that these recommendations are made on basis of the experiences from two cases, and that the adequacy of the recommendations depends on the generality of these results.

If resources allow, the combination of Group-based expert walkthrough and Cooperative usability inspection seem to be preferable. This is in particular due to the complementary character of the results from the methods in this study.

Running Cooperative usability testing is somewhat more resource demanding than Group-based expert walkthrough, both in terms of the test sessions and analysis. However, if resources allow either a Cooperative usability evaluation or a Group-based expert walkthrough, the results of the present case seem to favour Cooperative usability testing; in particular because of the low thoroughness scores of Group-based expert walkthrough.

Running Group-based expert walkthrough is the least resource demanding method of the study. The method seems to generate valuable results. To the degree that we have been able to investigate, the validity of the method seems to be sufficient. We recommend this method as the very low-cost alternative for conducting evaluations of work-domain specific software solutions.

References

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

Dumas, J. S., Fox, J. E. (2008) Usability testing: Current practices and future directions. In: Jacko, J. A., Sears, A. (Eds.) The Human-Computer Interaction Handbook, 2. ed. Lawrence Erlbaum Associates, pp. 1171-1190.

Følstad, A. (2006). Analytisk usability-evaluering i gruppe: Sammenligning av usability-eksperter og domene-eksperter som evaluatorer. SINTEF report, STF90 A05083.

Følstad, A. (2007) Group-based Expert Walkthrough. In: Scapin, D., Law, E. (Eds.) R3UEMs: Review, Report and Refine Usability Evaluation Methods. Proceedings of the 3. COST294-MAUSE International Workshop, pp. 46-48.

Frøkjær, E., Hornbæk, K. (2005) Cooperative usability Testing: Complementing Usability Tests with User-Supported Interpretation Sessions. In proceedings of CHI 2005, ACM Press, pp. 1383-1386.

Hartson, H.R., Andre, T.S., & Williges, R.C. (2001). Criteria for evaluating usability evaluation methods. International Journal of Human-Computer Interaction 13(4), 373-410.

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. Biometrics, 33, 159-174.

Muller, M. (1998). Participatory heuristic evaluation. Interactions, 5(5), 13-18.

Muller, M. (2008) Participatory design, the third space in HCI. In J. A. Jacko, A. Sears (Eds.), The human-computer interaction handbook (2nd ed.) (pp. 1061-1082). New York: Lawrence Erlbaum Associates.

Rubin, J., Chisnell, D. (2008) Handbook of Usability Testing, Second edition: How to Plan, Design, and Conduct Effective Tests. Wiley Publishing.

Sawyer, P., Flanders, A., Wixon, D. (1996) Making a Difference - The Impact of Inspections. In Proceedings of CHI 1996, ACM Press, pp. 376–382.

Appendix 1: Cooperative usability testing, method description.

Cooperative usability testing is prepared and run as a formative usability test, as described e.g. in the usability testing handbook of Rubin and Chisnell (2008), except that a series of interpretation phases are included in the test procedure. In the following we will present what interpretation phases are, how to prepare for them, and how to run them. Also we will present how data colleted in interpretation phases may be analyzed. We assume that the reader is familiar with preparing, running and analyzing regular usability tests.

Background
Cooperative usability testing was first presented by Frøkjær and Hornbæk (2005), as a response to criticism of the adequacy of regular usability testing. We have adapted their original test procedure to include several interpretation phases instead of just one. Also, we have discontinued their reliance of video presentations in the dialogue between the test leader and the participant.

A short note on formative usability testing
Formative usability testing (hereafter only called usability testing) is conducted in order to identify usability issues related to interactive systems under development. The outcome of usability testing typically is a list of usability problems, possibly associated with redesign suggestions.

Usability testing typically involves a number of sessions with individual participants (users). Each session typically include a pre-interview/questionnaire, a set of task scenarios, and a post-interview/questionnaire. The actors of the session are the participant and the test leader (moderator). The participant should be a representative user.

Solving the task scenarios constitute the main part of the session. Task scenarios exemplify assumed core tasks for the users of the system under evaluation. The participant is typically instructed to solve a task scenario unaided by the test leader. Often the participant is instructed to "think aloud". If the participant has questions or encounters difficulties in solving the task scenario, she may ask the test leader for help, which may be given following a specified protocol. When one task scenario is completed, the test leader instructs the participant in the next task scenario. In a post-interview, the test leader may clarify with the participant incidents during task scenarios (e.g. experienced difficulties, or surprising behavior), or ask questions related to user experience, attitudes or preferences.

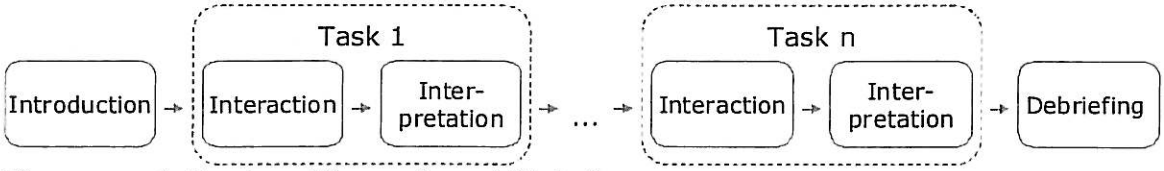
In the subsequent analysis, usability problems are defined on basis of incidents during the task scenarios, informed by data from interviews/questionnaires. Also, analysis may be done on task completion, time spent solving task, or user satisfaction.

Interpretation phases
To understand Cooperative usability testing, we need two terms: *Interaction phase* and *interpretation phase*. An interaction phase is the participant completing a single task scenario upon receiving the instructions from the test leader. An interpretation phase is the participant and test leader engaging in dialogue to analyze the incidents of the task scenario, gain understanding of the participant's perception of the solution, and discuss whether redesign suggestions may be made.

Cooperative usability testing differs from usability testing solely by the inclusion of interpretation phases. An interpretation phase is included following each interaction phase. This means that for

each task scenario the participant is confronted with, she first completes it by herself and then discusses her experiences and perceptions as she walks through the task scenario a second time.

The structure of Cooperative usability testing is visualized in the figure below.



Figure, annex 1: Structure of Cooperative usability testing

Preparing interpretation phases

Preparation of interpretation phases are done as part of the regular usability testing preparation. In the preparation, four issues needs to be considered.

1: How to walk through the preceding task interaction?

The interpretation is conducted as discussions between the test leader and the participant on basis of a walkthrough of the preceding interaction phase. The walkthrough may be conducted either by (a) walking through the entire task as it was solved by the participant in the interaction phase, or (b) moving directly to the parts of the task that represented interesting issues in the preceding task, or (c) presenting the participant for a video recording of the interaction phase. These different all have their merits. Video playback ensures that the task is presented just as it was conducted, but poses severe limitations with regard to trying out or demonstrating UI features that were not used in the interaction phase. Non-video supported walkthrough of the interaction sequence provides high levels of flexibility with regard to what can be presented and discussed, but may be problematic with regard to that the test leader needs to remember all the important details of the interaction. We recommend that a non-video supported walkthrough is chosen as an approach for relatively low complexity tasks, and that video playback is considered for relatively complex tasks.

2: How to ask for participant interpretations during task walk through?

The test leader should prepare a simple protocol for how to ask for the participant's interpretations and opinions during the walkthrough. Such a protocol may include:

- For each problem or hesitation, stop the walkthrough and ask the participant: "Please, explain to me what happened here?"
- For each case of unsuspected actions, ask the participant: "Here you chose to do X. Why did you make this choice?". Possible follow up questions include: "Do you see any other choices you could have made her in order to reach the task goal in a different manner?" and "Do you see the feature Y? What do you think this is?"
- Upon completing the discussion on a given problem/unsuspected action, ask the participant: "How could we have made this different in order to help you avoid this problem/use the solution the way we intended it to?"

The test leader may also prepare questions related to known design trade-offs associated with the task, in order engage the participant in discussions on pros and cons of design alternatives for one or more of the key UI features used in the preceding task.

1. How to ask for participant experiences after task walkthrough

The interaction phase may be concluded by asking that participant general questions on user experience and redesign suggestions. The test leader should prepare a small set of questions to ask at the end of each interpretation phase. Example questions include:

- Please describe your experience of the software as a means to solve this task?
- Would you like to use this software to solve such tasks in your everyday life? (Why/Why not?)
- Please describe the most important changes to make to the software in order to improve it for this kind of task?
- Are there any features of the software that you notice as really appreciable during this task?

2. Recording the interpretation phase

It may be beneficial to record the interpretation phase, just as the interaction phase. If you choose to conduct a non-video supported walkthrough of the preceding task, the simple solution to this will be to keep on recording during interpretation phases. If you choose to conduct a video-supported walkthrough, you will need a technical setup that allows video playback of the interaction phase, and at the same time allow recording of the interaction phase.

Running interpretation phases

The interpretation phase is run immediately following the related interaction phase.

Upon starting the interpretation phase, the test leader should clearly state the procedure and objective of the interpretation phase, and encourage the participant to comment, discuss and suggest freely during the walkthrough.

During the preceding interaction phase, the test leader or the observer will have noted participant problems, hesitations, or unsuspected actions. During the walkthrough of the interpretation phase, the test leader – possibly aided by the observer – ask the participant open-ended questions related to these observed issues. The participant may also be explicitly encouraged to state her impression of the UI during the walkthrough.

Upon completing the walkthrough the test leader asks the participant the prepared general questions on user experience and redesign suggestions.

Analyzing interpretation phases

The interpretation may be utilized for four purposes during analysis:

- Explain observed problems or surprising actions
- Redesign suggestions
- Additional utility of usability issues not observed during the interaction phases
- User experience

It is advisable to be explicit on the origin of reported usability issues. Are they observed problems or stated issues? In the case of stated issues, do several participants converge on the same issues?

When the data from the interpretation phase is used to explain observed problems, the explanation should be reported as stated by the participants. Likewise, it is necessary to separate redesign suggestions proposed by the users and redesign suggestions proposed by the involved usability personnel. Participants redesign suggestions should be commented by the usability personnel.

Additional utility issues may be an important outcome of the interpretation phases. However, such issues may be related to needed functionality, and may be involve extensive UI changes.

User experience statements should be presented, but needs to be interpreted as examples of experiences and not findings general to the user population. The reason for this is the relatively

low number of participants and sampling method used for usability testing. Interesting user experience issues may possibly be followed up in later investigations – e.g. through satisfaction surveys.