What Do Usability Evaluators Do in Practice? An Explorative Study of Think-Aloud Testing

Mie Nørgaard
Department of Computer Science
University of Copenhagen
mien@diku.dk

Kasper Hornbæk
Department of Computer Science
University of Copenhagen
kash@diku.dk

ABSTRACT
Think-aloud testing is a widely employed usability evaluation method, yet its use in practice is rarely studied. We report an explorative study of 14 think-aloud sessions, the audio recordings of which were examined in detail. The study shows that immediate analysis of observations made in the think-aloud sessions is done only sporadically, if at all. When testing, evaluators seem to seek confirmation of problems that they are already aware of. During testing, evaluators often ask users about their expectations and about hypothetical situations, rather than about experienced problems. In addition, evaluators learn much about the usability of the tested system but little about its utility. The study shows how practical realities rarely discussed in the literature on usability evaluation influence sessions. We discuss implications for usability researchers and professionals, including techniques for fast-paced analysis and tools for capturing observations during sessions.

Author Keywords
Usability evaluation, think aloud testing, industrial software development, user-centered design

ACM Classification Keywords
H.5.2 Information Interfaces and Presentation (e.g., HCI): User Interfaces—Evaluation/Methodology; D.2.2 Software Engineering: Design Tools and Techniques

INTRODUCTION
Methods for usability evaluation are one of the successes of human-computer interaction: they are widely used and in many cases improve the usability of the software to which they are applied. According to recent surveys [14,35], think-aloud testing (TA) is widely used and valued by usability evaluators. Numerous studies have been made of usability evaluation methods in general, and of TA testing in particular [17,21,23,25,29]; for recent reviews see [7,9]. In our view, however, these studies are biased in two respects. First, most studies do not take place in a practical software development context, but in a laboratory-style set-up with non-expert participants. While such studies give insight into benefits and drawbacks of particular evaluation methods, they miss how practical realities of software development shape the use of evaluation methods [37]. Second, studies of usability evaluation tend to focus on coarse measures of outcomes such as the number of problems identified; they rarely describe the process of evaluation in detail. One exception is diary studies of usability evaluation, such as [24], which have provided valuable input on how evaluation methods are used. In a 2004 keynote, John called for more studies of the process of using HCI methods [22], seemingly dissatisfied with the current literature.

Addressing the two biases above, this paper reports an explorative study of how TA testing is practiced. We do so by observing the setting up, carrying out, and handling of results from TA sessions in professional consultancies or software development organizations. Inspired by grounded theory and verbal protocol analysis, we analyze and summarize data with two expected benefits. For usability researchers, we intend the paper to deliver insights into some issues of practical usability work. For usability professionals, we identify some of the problems and tradeoffs they face, hoping that this may assist the planning and conducting of future TA tests.

RELATED WORK
The question of how TA testing is done in practice is related to studies (a) describing experiences from real-life usability evaluation or (b) presenting detailed information on the process of usability evaluation. Below we review this research and discuss the extent to which it helps understand the practice of TA testing.

One group of studies describes *real-life usability evaluation.* Some of these studies systematically collect data through observation and interviews of usability specialists and other stakeholders in software development projects, see for example [3,19,36]. These studies focus on factors that facilitate or impede usability evaluations and the impact of their results. They have identified several strategic concerns in real-life usability evaluation, such as the need for users to be involved throughout the design

DIS 2006, June 26–28, 2006, University Park, Pennsylvania, USA.

process to facilitate useful contributions [36] or that the organization of usability work, to some extent, shape usability results [19]. They do not, however, in detail discuss how evaluations are undertaken.

Other studies have focused more on tactical issues of usability evaluation, see for example [10,15,27,31,34]. These issues include how to make the results of usability evaluations such as TA testing impact software development [15,31] and how to deliver feedback that is useful to developers [10,17]. As an example, Molich et al. [27] discussed how the usability reports produced by nine teams of mostly professional evaluators differ in content. They found great variation in selection of tasks for usability tests and in reporting of results. Studies of tactical issues of usability evaluation rarely describe the process but focus mainly on the outcome of usability evaluation.

Equally interesting are studies where professionals report how practical circumstances have forced them to adapt and develop the evaluation procedures they use, see for example [1,32,38]. Spencer [32], for example, described how the evaluation technique cognitive walkthrough was modified to better fit the realities of the software development organization in which he worked. Those realities include time pressure and a defensive attitude among participants in the walkthrough. Spencer reported that the modified technique worked better in his organization. Such studies provide interesting observations on factors influencing practical usability work, such as the influence of a particular kind of product on the decisions about which evaluation method to use [38]. Yet, they lack the methodological rigor of the studies mentioned above and may not provide general lessons for usability research.

Another group of studies has focused on the *process of usability evaluation.* Mostly, the academic literature on usability evaluation has been concerned with the outcome of evaluation in the form of problem lists or suggestions for redesigns. A few studies, however, have reported diary studies of usability evaluation [18,20,24]. In those studies, evaluators typically keep a diary in which they make notes on their planning, conducting and reporting of an evaluation. John and Packer [24] showed how participants in a diary study made severity judgments based on personal judgment rather than on the usability evaluation technique used. Hornbæk and Frøkjær [18] argued that the evaluation process observed in their diary study was complex, with participants identifying usability problems not just while conducting the actual evaluation, but also during planning and reporting of the evaluation. The studies referenced above, however, look only at non-expert evaluators outside an industrial software development context. These studies, and studies where the evaluator fill out forms during evaluation [8], present the most detailed data on evaluation currently available. We know of no studies that have systematically observed and analyzed usability evaluation, for example using video. Overall, it appears that studies looking at real-life usability evaluation place little focus on describing the *process* of usability evaluation; studies of the evaluation process look at somewhat artificial evaluation settings with diaries as the data-collection method with the finest granularity.

The paper by Boren and Ramey [4] is a notable exception to these shortcomings. Boren and Ramey observed TA sessions in two companies, and related their observations to what some consider the theoretical basis of TA testing, the work of Ericsson and Simon [12]. The analysis by Boren and Ramey showed discrepancies between the observed TA testing and the recommendations of Ericsson and Simon. While the work of Boren and Ramey has given unique insights to usability research, it is limited in that they reported mainly discrepancies to Ericsson and Simon's prescriptions (in particular about prompting the user), and not more general issues confronting a usability specialist conducting an evaluation.

Attempting to broaden the focus of Boren and Ramey's paper we next present an explorative study concerning how usability evaluations are conducted in practice.

EXPLORING THE USE OF THINK ALOUD PROTOCOL

The question guiding the study is: *what do usability evaluators do in practice?* To get a better understanding of this we observed 14 TA test sessions in seven companies. We chose to focus on TA testing because it is widely used and because observing analytic usability evaluation, such as heuristic evaluation, presents methodological difficulties (e.g., concerning introspection) that we wanted to avoid. Our data comprise mainly audio recordings of the setting up, running and analysis of the TA sessions. Our intention is *not* to reprehend the practice of usability testing. Rather, we aim to explore what usability evaluators do so as to (a) sensitize usability research to industrial practice and (b) help evaluators understand better the strengths and weaknesses of what they do.

Companies Participating in the Study

Seven companies agreed to participate in the study by letting us observe how they conduct TA tests. The companies were recruited among Danish enterprises that either offer usability evaluation as consultancy or integrate usability evaluation in their systems development. Table 1 provides a summary of the companies; their names replaced by the letters A through G.

Our sample comprises three companies that provide usability evaluations solely to customers outside of the company and work with information technology as part of their core business (companies B, D, F). Two of the companies in the sample (companies A, C) perform usability evaluation both in-house and to customers outside of the company. These two companies have information technology and systems development as their core business. Finally, two of the companies solely perform usability evaluation in-house (companies E, G); while both companies have a strong presence online, their core business is in the service sector. The companies vary in size from 2 to 8500. They had varying levels of experience with usability evaluation; some of the evaluators we observed had only worked with usability for one year, while one had been conducting usability evaluations for eight years. Four companies evaluated running prototypes (companies A, C, E, F), two companies evaluated deployed applications (companies B, D), and company G evaluated paper prototypes. All tests observed were formative tests in that they were usability evaluations with users seeking to investigate issues such as concept, tools and navigation.

Company	*Total no. of employees (no. of employees working with usability)*	*Test sessions observed*	*Evaluators present during tests*	*Evaluators' experience in years*	*Customer of test results*
A	810 (6)	2	2	1 - 6	Intern
B	2 (2)	1	2	1 - 8	Extern
C	165 (3)	1	2	2.5 - 6	Intern
D	7 (7)	1	2	1 - 6	Extern
E	8500 (7)	4	2	4 - 6	Intern
F	16 (3)	3	2	1.5 - 6	Extern
G	3464 (8)	2	1	6.5	Intern

Table 1. The companies participating in the study and the test sessions observed within each company.

Data Collection

Methodologically we were inspired by grounded theory which dictates that researchers should not initiate an investigation on the basis of a list of hypotheses [30]. Our data collection was thus broad and open-ended. We tried to participate in as many of the activities surrounding the usability evaluations as possible, wanting to probe how the TA protocol is put into practice. Data was collected over a period of three months and the focus of attention developed during this time, as suggested by [33] and [30].

The core of our data is the observations, field notes, and audio recordings from 14 TA sessions, that is, the period of time from the arrival of the test participant until that participant leaves. These sessions were distributed among the companies as shown in Table 1; the number of sessions we could observe was largely dictated by practical circumstances. In all sessions, except those of company G, two evaluators from the company were present. On average, an evaluation consisted of a series of six sessions, of which we typically participated in two. The sessions we participated in were placed both at the beginning, middle and end of the series. In one session, the recording made from an observation room was of such poor quality that it allowed only sporadic transcription of the interaction between user and evaluator.

When possible, discussions, analysis, and informal conversations among usability evaluators before and after the test sessions were also observed and recorded. Sometimes customers (i.e., the persons who commissioned the test) were also present and took part in these discussions (e.g., company B). In two cases we recorded when usability evaluators delivered test results to the customers (companies C and G). In two cases we collected reports, summaries or notes that documented the tests (companies F, G). In two cases (companies A and C) we additionally conducted semi-structured interviews with the persons responsible for the usability work in the company.

The data collection described above resulted in, among other material, 24 hours and 54 minutes of audio recordings. Below we focus on the test sessions and the discussions immediately following tests – we only mention material from feedback sessions, usability reports, and the semi-structured interviews, when it corroborates findings from the core data.

Data Analysis

Analysis was conducted in three phases. First we segmented the recordings applying descriptive keywords to each segment. Second we re-evaluated segments and keywords in order to adjust keywords or apply new ones. Third we analyzed and tried to form a coherent interpretation of segments that shared keywords. We explain this procedure more thoroughly below, and briefly relate it to grounded theory [30] and Chi's proposal for how to analyze verbal protocols [6].

Segmenting and open coding of the recordings

The audio recordings were initially divided into 641 segments. One segment could concern a usability evaluator analyzing the test results, or explaining how to ensure scientifically valid test results. A segment could last from a few seconds to several minutes. We chose to do only a partial transcription of the recordings, but listened repeatedly to the segments during our analysis.

In order to code the segments, keywords were attached to each segment allowing us to analyze and group segments. Thirty-five keywords were generated as the study proceeded. Some segments regarded more than one interesting topic and hence got more keywords attached to it. This process is similar to open coding in grounded

theory [30] or to Chi's [6] phase of developing or choosing a coding scheme or formalism.

Re-evaluating and crosschecking the coding

In order to ensure that a segment contained evidence for a specific keyword, the coding was carried out in two iterations, one by each of the authors. Disagreements or questions about the attachment of a keyword to a segment were discussed before attaching an existing or creating a new keyword. This is similar to Chi's phase of operationalizing evidence in the protocols [6] and, in part, to axial coding in grounded theory [30].

Synthesizing and interpreting the data

Groups of segments, which shared the same keyword, were analyzed to identify the most interesting areas and thus reduce the size of data. For interesting areas, we looked for the observations that were most surprising to us, or seemed to contrast the literature on usability research and textbook recommendations on how to do a usability evaluation. Such areas were selected for further analysis and interpretation. This phase is similar to Chi's phases of seeking patterns in the mapped formalism [6] or selective coding in grounded theory [30].

RESULTS

The following section describes our results organized in six areas. Table 2 summarizes these areas and the main findings within each of them. The areas concern (1) analysis of the results from a session, (2) confirmation of known issues, (3) practical realities, (4) questions asked during a test, (5) laboratory-style scientific standards, and (6) uncovering usability problems or utility concerns. Below we present each area in turn. For findings we give the number of sessions in which they were observed. We use sessions rather than segments as an indication of frequency, because the number of segments is strongly influenced by the nature of a session, especially how much the evaluator and the user talks, how much they jump between topics, etc.

Analysis of Results from a Test Session

The first area concerns how usability evaluators analyze test sessions. By analysis we mean the task of understanding and agreeing upon important observations from a session. Analysis also includes attempts to understand the causes of those observations, interpret user behavior and find design solutions to observed problems.

None of the sessions included attempts to carry out a

Area of attention	*Main finding*	*N*	*Example of observations and quotes*
Analysis of results from a test session	Analysis is unstructured	9	Scattered fragments of analysis; no systematic approach used
	Analysis is incomplete	9	Does not identify causes or solutions; restricts discussion to user traits
	Analysis as a summary with the user	3	"Let's sum up"; selecting a few problems for further questioning; listing key findings
Confirmation of known issues as a test' focus	Looking for known issues	8	"Now, I am just looking for ammunition"; develops ideas of problems before testing; tasks and questions designed to point out known issues
	Practitioners have foreseen problems	5	"We have a gut feeling", "I told you so"
Practical realities influencing tests	Technical problems	8	System breaking down; long response times in test environment; installation or security messages interrupt workflow
	Unfinished prototypes	6	Parts of prototype missing or inaccessible; "a log-in name should not be WaddleFish"; texts and pictures are wrong or out of date
Questions asked during a test	Problems are explained, not experienced	13	"Do you think you would go back to the front page"; "did you notice this column"; "what do you expect to see"
	Leading questions	13	Questions address certain parts of GUI or system; evaluator hints the solution; "Can you do this in another way?"
	Unnecessary or obvious questions	10	"You did figure out to press the print button?"; asking user to locate information that clearly appears on the present screen; asking if user would like relevant information
Trying to meet laboratory-style scientific standards	Evaluators want similar conditions for users under test	5	"We have to make sure all users get the same questions"
	Rigid or artificial procedures	3	Laboratory style procedures; Danish evaluators speaking English to a Danish user; measuring subjective satisfaction overly systematic
Uncovering usability problems or utility concerns	User points to utility or lack thereof	10	"I would not do it like this"; user chooses to solve task without help of system
	Evaluator probes utility concerns	7	Asking about normal workflow; asking whether a task is realistic; "What would you typically do?"

Table 2. Overview of results. N refers to the number of sessions in which a finding was made (out of 14 sessions in total)

structured analysis of the results immediately after the session, for example by systematically agreeing on and then analyzing, say, the ten most prominent observations of user difficulties. However, as we have not in this study covered every step from test design to final report, we are not able to say if analysis took place later.

One evaluator did carry out a semi-structured analysis in the last minutes of three sessions though, focusing on summarizing key findings while the user was present:

> F1: "Let's sum up: The front page [should] maybe emphasize what they have in mind […] and the logo [gesturing where a logo should be]….and eventually [we should] list these sections. And the picture behind [we should] make it a bit more interesting. The editorial ends down here [points]…"

Three other evaluators (in a total of five sessions) also tried to sum up a few problematic topics and return to those topics for further questioning before ending the session. However, we did not encounter any systematic attempt to cover the most important observations directly after a session.

After a session had finished, the most common activity was that usability evaluators, and in four sessions also customers, discussed the session. We observed how they presented overall impressions intertwined with a general discussion about the system, social talk, observations, ideas for re-design, and occasionally analysis of the problems. To illustrate, an 11 minutes long discussion of a session was shaped as follows:

> Impressions of user attitude, discussing problems with prototype (3 min); Identification of one problem, analysis, summary of observations from session (2 min); Talk about old ideas, identify two problems, analysis (2 min); Discussion of recommendations and re-design (30 sec); Customer calls–and gets a short general summary (1.5 min); Summary of findings combined with general talk (2.5 min).

After this discussion, one evaluator went on to write a summary of findings to the customer. In other sessions the evaluators would just have a short conversation about general impressions before leaving the room, and thus ending the attempt to carry out an immediate analysis.

In nine sessions we saw examples of incomplete analysis. By incomplete analysis we mean remarks or observations that, if they were intended to assist in uncovering usability problems and solutions to such problems, needed to be elaborated and discussed. In seven sessions, for instance, evaluators would quickly characterize a user as being for example confused or insecure, but fail to follow up on this characterization or even identify what made the user become confused or insecure.

Confirmation of Known Issues as a Test Focus

The evaluators made comments before, during and after sessions, which let us to believe that they held more or less strong ideas about usability problems of the particular system being tested, even before commencing on the test. These ideas appear to shape the design of tasks and the questions raised during a test session. While such ideas are natural and may be important hypotheses, they sometime appear to focus the test on a particular topic or hypothesis. This delicate balance seem difficult to master.

After a session one evaluator stated, for example, that the test should provide proof for the conclusions in a usability report, which she had already begun writing:

> C1: "I think we agree on many of the issues"
> C2: "Yes – I have already written the chapter, I just need the ammunition".

A total of four evaluators stated that they had a more or less clear idea of the usability problems before commencing a test. In an interview another evaluator said that usability tests in some cases merely serve to confirm the evaluators' assumptions:

> A: "When we design a test we practically always have a gut feeling where it will fail […] in a way it is just an 'I told you so'-kind of thing, but it is nice to be able to document it".

The quotes suggest that usability evaluators see a need to support expert opinion with something more concrete when presenting customers with advice on usability. This may lead to tests that in part serve only to confirm.

In addition to these expressed opinions, it also appears that the actual activities of a test are sometimes chosen to confirm, or at least explore, areas known to be problematic. Questions and tasks within a test, for example, would be chosen to explore well-known issues. This led to test situations where evaluators literally waited for the user to point to the problem area. A1 explained to us how a certain task that required the entry of percentages most likely would cause problems. During the test, the user did actually spot the problem, and the response from the evaluator suggested almost a relief that the user did so:

> U: [Typing]
> A1: "So you just added minus 10 on both lines?"
> U: "…And then I got 20%.....WHAT?"
> A1: "Yes" [laughs out conformingly]

In another session, in response to a user severely criticizing a particular functionality, the usability evaluator broke out in laughter and said "this is really good", suggesting to us, that this issue was already anticipated as being problematic. In this way, 8 of the 14 sessions had examples of evaluators directly or indirectly expressing that they were confirmed in their preconceived opinions about usability problems.

It is hard to say whether a test focused at confirmation influences how evaluators interpret the observations they make during a test. An evaluator from company A noted after a session; "we really wanted to test this because we are confident it will fail, he [the participant] managed it, but I am sure others will not". The quote suggests that the expectation to find the problem in future tests could overshadow the possible interesting observation that at least one user successfully used a particular part of the interface. We return to discuss the balance between known issues and new findings in the discussion.

Practical Realities Influencing Tests

The study revealed numerous practical problems that usability evaluators experience when testing. In 12 sessions we observed examples of such problems or practical realities. These include system failures, users not showing up for a session, disturbing surroundings, and technical problems with recording devices. Despite such problems the evaluators managed to carry out all of the sessions.

Data show that the practical realities surrounding a test are produced by many factors, some out of the evaluators' control. In eight sessions, for example, we observed severe technical problems interfering with the session. As an example one session had a technical problem approximately every five minutes, each resulting in a break in workflow.

In two sessions problems arose because the customer had failed to provide the required number of test participants, thus forcing the evaluators to quickly find a solution in order to carry through the test within the scheduled time:

> F1: "The next user is one of my old friends […]"
> F2: "[…] they are not the first ones we choose, but if the customer fail to recruit [when they have agreed to do so] then we take whomever we can get."

Six sessions had problems with unfinished prototypes or last-minute changes to the prototype. One evaluator noted:

> D1: "Some things will, if not done properly, affect the users' perception rather dramatically…A log-in name should not be "WaddleFish", it's such a developer-kind-of-thing to make up funny log-in names like that"

Unfinished prototypes or prototypes recently changed are two reasons that evaluators often were confused or in doubt about the functionality of the prototype. In seven sessions evaluators stated that they were not familiar with aspects of the prototype's functionality:

> G1: "Now…let us see…[searches in prototype paper sheets]…these are brand new, so I have not looked at them before"

In sum, severe practical problems in some sessions lead to a continual interruption of the participants' attempts to complete their tasks. In this study, the practical realities influencing tests are much more frequent and severe than one would expect from textbooks or research papers on usability evaluation.

Questions Asked During a Test

The study showed variations in the types of questions asked by the evaluators. We analyzed these to understand which kinds of information usability evaluators are interested in, and to discuss later the validity of the information gained by different kinds of questions.

A large number of questions were reminders to keep talking like "Hmmm" and "Yes?". These kinds of questions were omnipresent and should be uncontroversial. Equally unsurprising is the many questions that simply try to elicit what the user is currently doing, or what problems the user is facing, for example "What is happening?", "What are you looking for?", or "What is the problem?". Many of these questions concerned the users' *experienced problems* in solving concrete tasks.

We encountered evaluators asking questions that differed dramatically from how Ericsson and Simon [12], and in part also Boren and Ramey [4], suggest to interact with test participants. Some questions concerned, for example, *nonexistent parts* of the system, such as asking how the user would use a mouse to interact with a paper prototype or what the user would feel about having to create a user profile in order to be able to use the system.

Other questions appeared *speculative* or hypothetical. One evaluator asked, for example, "Do you think you would go back to the front page at some point?" and "Let us say that something here [in a list of articles] would interest you…" (both F1), asking the user to continue on this assumption.

Some questions urge users to look back in time and remember their thoughts, that is, *retrospective questions*. For example "Did you notice this column [when you were here before]?" (F1), or "Do you remember if you got what you expected from the web shop?" (E1).

Questions about the user's *expectation of the system* were also frequent, for example: "What would you expect to see?" or "How many would you expect to find?" (both from company G). Questions about the expectations of the system were often asked in the beginning of the session, for example:

> D1: "Then you enter this page, and my first question is: Try looking at the page and try not to click on anything but just tell me what is happening on this page, what can you do, how do you like it and give me all of you general impressions. You may go into detail and if you point at something you are encouraged to do so with the mouse so that the secretary can see what is going on"

Another type of question apparently aims to elicit information about the *users' feelings*, typically by asking directly about what the user liked, trusted or were

interested in. E1 asked, for example "So…you feel more secure now…or?", and F1 probed "Is there anything where you think: 'Wow! I would like to click on that'…or?".

In 13 sessions we observed one or more questions of the five kinds described above. In contrast to the experienced problems discussed earlier they did not concern problems experienced as part of solving a task, but rather imagined, indirectly experienced or *expected problems*. This intensive probing for such problems surprised us.

Thirteen sessions showed another kind of question, best characterized as leading questions. One evaluator, for instance, asked a question aiming at a certain issue of interest and the user would without much trouble solve the task or answer the question as anticipated:

> [The user has pressed play to see an episode of a series of video clips in a media player:]
> G1: "What would happen when this episode was over?"
> User: "The series would end"
> G1: "It was just a short version of the series or what?"
> U: "[…]I have pressed to see the whole series...Ah! I have pressed to see the whole series [...] something [other episodes] could come afterwards […]"

Trying to Meet Laboratory-Style Scientific Standards

The evaluators made several remarks suggesting that they find validity to be of great importance when testing. The concepts of validity upon which evaluators rely seem primarily to be those of scientific experiments, such as keeping the same procedure throughout a test, using representative subjects, and using elaborate questionnaires to get information on users' satisfaction. Note that we here mainly describe the evaluators' beliefs; in the discussion we will look closer at the relation of these views to those presented by the literature.

Evaluators from three companies (representing five sessions) emphasized that one should not change a test design between the sessions of a test. Changing a test design could include making changes to questions, tasks, prototype and choice of language. One evaluator, for example, stated the importance of maintaining the same tasks and phrasings of questions throughout a TA test even though it was evident after a few test sessions that the users misunderstood some of the tasks.

> C1: "I think it is really annoying that we already now can see problems, which we cannot correct as we go along…but we have to make sure that all users get the same questions"

In three sessions we observed how the fact that evaluators were trying to adhere to laboratory-style validity resulted in rigid and artificial procedures. For instance, we observed a session where Danish evaluators asked questions in English to a Danish user. The aim was to make test conditions similar among Scandinavian participants. In another session, evaluators tried to collect data about the system through a series of questions (e.g., "I will be more effective with the system") that users should rate on a one-to-seven scale. These questions are similar to instruments for measuring subjective satisfaction typically used in laboratory-style experiments. While such scales certainly have their uses, in this case they seemed to contradict what had happened during the session minutes before. This observation was supported by the evaluator:

> A1: "When users rate statements […] we take the results with a kilo of salt. This guy – it is a pretty good score right but […] in the beginning he was right-clicking all over the place and he mentioned that he did not like the buttons disappearing…"

Thus, the questions were seemingly included to adhere to some perception of how scientific user testing should be conducted. In this case, the answers were apparently not used, but had they been, it might have lead to a de-emphasis of the user difficulties just observed.

In sum, the attempt to adhere to scientific standards in some cases lead to rigid or artificial procedures that appeared unnecessary given the influence of practical realities and the rather informal analysis of test results mentioned earlier.

Uncovering Usability Problems or Utility Concerns

All sessions in the study would naturally include segments where usability problems were identified, including problems with scrolling, positioning of information, how links should be emphasized, how the user was prompted for information several times, etc. Other segments concern the utility of the system, for example which tasks the system should support or whether tasks from the test were unrealistic with regard to how the user usually uses the system or would want to use the system [28]. We observed utility concerns being discussed in 10 sessions.

In seven sessions we observed how the evaluator asked more or less specific questions concerning the utility of the system. Consider the following example:

> F1: "Lets look at the article again...What would you typically do?"
> User: "I would pass it on...if it was fun and interesting…"
> F1: "Like printing it?"
> U: "No just by word of mouth…"
> F1: "Word of mouth. Ok…"
> U: "…unless it was really good - then I would forward it electronically..."
> F1: "Would you ever print articles?"
> U: "No....I actually save them […]"
> F1: "So…do you copy the text and paste it into a Word document?"
> U: "Yes, I could do that"

Ten sessions had examples of users who were pointing to utility problems like the following from company C:

U: "[reads question loud:] 'Where would I look for an employee?'.... I would use a phonebook [which is not a part of the system]"

Some users specifically pointed to areas of the system, which they found failed to support their workflow, for example from company E: "This is just to tell you that I would not do it like this".

In 13 sessions we observed how problems relating to usability seemed to be favored over problems relating to the utility of the system. A remark from a user about not wanting to solve a task in the way suggested by the system did for example not result in an attempt to investigate that utility problem further; nor did it get reported to the customer during the feedback session we observed. This study suggests that utility problems are much less frequently examined than usability problems. Given the little attention problems regarding utility got in the sessions we observed, we do not expect them to be treated more thoroughly in discussions that we did not attend.

DISCUSSION

To sum up, this study shows that careful and systematic analysis of usability problems rarely take place immediately after the sessions in which they occur. Evaluators do not always, either, ensure that they agree on even the most important observations from a test. In addition, many tests appear to search also–and sometimes mainly–for confirmation of issues known beforehand or observed in other tests. Most of the sessions we observed were affected by practical realities such as incomplete prototypes and evaluators' limited experience with the system being tested. The questions raised by the evaluators during the test varied, but some questions appeared hypothetical and probed only users' expectations and not the problems they actually experienced. Some evaluators seemed to regard TA testing as a scientific laboratory-style method resulting in rigid and artificial procedures when conducting the test. Finally, seemingly important observations about the utility of the system being evaluated were made during sessions. These were infrequent, however, compared to results and discussions concerning usability issues.

Most surprising to us is the lack of systematic analysis while the results of a test are still fresh in mind. As we have not covered every step from test design to final report in this study, we are unable to rule out whether analysis were done at a later stage. Still, the fact that evaluators rarely check whether they agree on the most important observations from a session adds to the picture of analysis as being a weak part of the evaluation process. Work on the evaluator effect [16] show that evaluators observing the same test find substantially different usability problems, making collecting and discussing different views of the main observations important. Summaries of the main observations by the evaluator while the test participant was present worked well—similarly to the idea of cooperative usability testing [13]. However, using this or similar techniques to agree on observations from a test does not in itself reveal usability problems, the causes of those problems, or possible remedies for them.

Perhaps the lack of systematic analysis is understandable, given the scarce advice about analysis of usability tests we receive from textbooks and introductions about how to do a TA study. Molich [26], for example, used 2 pages of his 33-page instruction on how to do TA testing to discuss analysis. Dumas and Redish [11] used around 31 pages of their 404-page textbook on analysis. It appears desirable that usability research develops and validates techniques supporting fast-paced analysis. Usability evaluators would be well advised to more systematically relate and discuss their observations when they are fresh in mind. Evaluators might take up using post-it notes for capturing observations during a session, and analyze these immediately after the session. They might find it rewarding to prioritize these post-its, possibly together with the user, to develop a common understanding, and discuss problems and feasible solutions.

The extent to which usability practitioners already before testing had a clear idea of the usability problems to be found was surprising. Interestingly, recommendations are made in the literature (e.g., [11], p. 160) about looking for known problems. Some views of the psychology of confirmation suggest that as a result of this, evaluators are very likely to confirm what they are looking for, perhaps failing to make other equally important observations. If the answer is not known with confidence prior to testing, we agree with the practice of exploring these explicit questions in the test. However, if usability issues are already known with such confidence that the practitioner is only "looking for ammunition", why test at all? Finding the balance between on the one hand testing specific areas of concern and on the other hand exploring the system in a more open manner seems to be an important but difficult challenge to evaluators.

The practical realities surrounding the tests we observed are far from the expectations about the test situation presented in textbooks such as [11]. Techniques and tools that are usable under such less-than-ideal circumstances are needed, for example to enable the analysis of observations in the usually short time available between sessions. Evaluators should for their part consider preparing material to be used on the fly in case of system failure.

Given the work of Boren and Ramey [4], we had expected open and varied questions. Quite surprisingly we saw hypothetical questions, abstract questions, leading questions, and plain impossible-to-answer questions: in short, questions that did not aim at understanding problems experienced by the user, but rather at encouraging users to predict possible problems. On the one hand this suggests that evaluators may be looking for information about

feelings and perceptions, which cannot be gained from a traditional TA testing. On the other hand we feel obliged to point out that some of the questions we encountered could never produce useful answers.

Questions about "first impressions", "what would you expect to be there [e.g., on the next page]", or "what do you feel about this" may imply that evaluators need researchers to provide more valid and systematic ways of probing for, say, participants' feelings of trust. Evaluators are advised to pay closer attention to the way they phrase their questions.

Questions probing for information about utility also seem to warrant further investigation. Molich [26] suggested to ask test participants about their impressions of the tasks *after* a TA session. In two sessions we observed how useful discussions about the users' real-life tasks developed from such a question being asked during a test session. However when the same type of question appeared at the end of a session as advised by Molich, it became more general and received also a general answer. We suggest for researchers to provide further techniques for initiating discussions about utility *during* tests, which would help address the concern that usability testing might "tune a user interface at the tail end of design, to clean up any rough edges or unnecessary difficulty in understanding or interacting with the interface" [2, p.373], instead of concern the user' tasks or needs. In order to understand and discuss how to improve the utility of a system evaluators may find it helpful to question the system's utility and ask users how they usually go about solving a specific task.

The study suggests a belief amongst some evaluators that usability testing is science, and therefore must meet the same criteria as science. Iivari [19] recently reported an explorative study in which similar attitudes were present among some usability professionals, "staid researchers" in Iivari's terms. The insistence on, for example, not changing tasks or procedure during a test appears rigid and counter-productive. We encourage evaluators to change set-up or make alterations to the prototype in the middle of a test if they believe it will help them answer important questions about the use of the system. Since TA testing is not a classical laboratory-style scientific testing method evaluators may feel they need to support the formative test results with summative measures. This need for bolstering a usability claim is discussed by [5] who points at highlights videos as one way of providing such evidence. Researchers are encouraged to search for other, less expensive methods, for backing up usability results.

Acknowledging the work of Boren and Ramey [4] this study aims at providing a needed description of how usability evaluation is conducted in practice. Two limitations are worth mentioning. First, we have only collected data in seven companies. Obviously, there are great variations in how usability work is conducted in those companies, which we have not touched upon. A goal for future work should be to collect more coarse-grained data, which would capture the process of usability evaluation in a greater number of companies. Second, we have mainly focused on test sessions. Thus, we did not explore the relation between test sessions and the feedback given to customers; nor did we collect any material on the planning of tests.

CONCLUSION

We have presented an explorative study of how usability professionals conduct think-aloud tests. It suggests that think-aloud tests might not get sufficiently analyzed. We see a tendency that evaluators end up focusing too much on already known problems, and that the questions they ask during a test seem to concern problems that the user expects, rather than problems actually experienced during the test. The tests were to some extent shaped by practical realities and by some evaluators' adherence to a strict, laboratory-style procedure. Finally evaluators seem to prioritize problems regarding usability over problems regarding utility, when they conduct think-aloud tests.

We encourage further work on methods for fast-paced analysis. Methods and procedures for investigating the utility and probing for users' perception of a system may also be of value for evaluators. Practitioners are advised to more systematically capture and discuss observations from a test. Questions about the practical relevance of the system evaluated could be one way to address utility issues. Investigating problems that are experienced rather than expected may also improve think-aloud tests.

REFERENCES

1. Arnowitz, J., Gray, D., Dorsch, N., Heidelberg, M., & Arent, M. The Stakeholder Forest: Designing an Expense Application for the Enterprise, *Proc. CHI 2005*, ACM Press (2005), 941-956.
2. Beyer, H. & Holtzblatt, K. *Contextual Design*, Morgan Kaufman Publishers, San Francisco, 1998.
3. Boivie, I., Åborg, C., Persson, J., & Löfberg, M. Why Usability for Lost or Usability in in-House Software Development, *Interacting with Computers*, *15* (2003), 623-639.
4. Boren, M. T. & Ramey, J. Thinking Aloud: Reconciling Theory and Practice, *IEEE Transactions on Professional Communication*, *43*, 3 (2000), 261-277.
5. Carter, L. & Yeats, D. The Role of Highlights Video in Usability Testing: Rhetorical and Generic Expectations, *Technical Communications*, *52*, 2 (2005), 1-7.
6. Chi, M. T. H. Quantifying Qualitative Analyses of Verbal Data: A Practical Guide, *The Journal of the Learning Sciences*, *6*, 3 (1997), 271-315.
7. Cockton G., Lavery, D., & Woolrych, A., Inspection-Based Evaluations, in Jacko, J. A. & Sears, A. *The Human-Computer Interaction Handbook,* Lawrence Erlbaum Associates, 2003, 1118-1138.

8. Cockton, G., Woolrych, A., Hall, L., & Hidemarch, M. Changing Analysts' Tunes: The Surprising Impact of a New Instrument for Usability Inspection Method Assessment, *Proc. HCI 2003*, Springer Verlag (2003), 145-162.
9. Dumas J., User-Based Evaluations, in Jacko, J. A. & Sears, A. *The Human-Computer Interaction Handbook,* Lawrence Erlbaum Associates, 2003, 1093-1117.
10. Dumas, J., Molich, R., & Jefferies, R. Describing Usability Problems: Are We Sending the Right Message?, *interactions*, 4 (2004), 24-29.
11. Dumas, J. & Redish, J. *A Practical Guide to Usability Testing*, Intellect, 1999.
12. Ericsson, K. A. & Simon, H. *Protocol Analysis: Verbal Reports As Data, Revised Edition*, MIT Press, Cambridge, MA, 1993.
13. Frøkjær, E. & Hornbæk, K. Cooperative Usability Testing: Complementing Usability Tests With User-Supported Interpretation Sessions, *Extended Abstracts of ACM Conference on Human Factors in Computing Systems* (2005), 1383-1386.
14. Gulliksen, J., Boivie, I., Persson, J., Hektor, A., & Herulf, L. Making a Difference - a Survey of the Usability Profession in Sweden, *Proc. Nordichi 2004*, ACM Press (2004), 207-215.
15. Hertzum, M. User Testing in Industry: A Case Study of Laboratory, Workshop, and Field Tests, *Proc. ERCIM Workshop on User Interfaces for All*, (1999), 59-72.
16. Hertzum, M. & Jacobsen, N. E. The Evaluator Effect: A Chilling Fact About Usability Evaluation Methods, *International Journal of Human-Computer Interaction*, *13* (2001), 421-443.
17. Hornbæk, K. & Frøkjær, E. Comparing Usability Problems and Redesign Proposals As Input to Practical Systems Development, *Proc. CHI'2005*, ACM Press (2005), 391-400.
18. Hornbæk, K. & Frøkjær, E. Two Psychology-Based Usability Inspection Techniques Studied in a Diary Experiment , *Proc. Nordichi 2004*, ACM Press (2004), 3-12.
19. Iivari, N. Usability Specialists - 'a Mommy Mob', 'Realistic Humanists' or 'Staid Researchers'? An Analysis of Usability Work in Software Product Development, *Proc. Interact 2005*, Edizioni Guiseppe Laterza, (2005), 418-430.
20. Jacobsen, N. E. & John, B. E. Two Case Studies in Using Cognitive Walkthroughs for Interface Evaluation, *CMU-CS-00-132* (2000).
21. Jeffries, R., Miller, J., Wharton, C., & Uyeda, K. User Interface Evaluation in the Real World: A Comparison of Four Techniques., *Proc. CHI'91*, (1991), 119-124.
22. John, B. Beyond the UI: Product, Process and Passion, *Proc. Nordichi 2004*, ACM Press (2004), 285-286.
23. John, B. E. & Mashyna, M. M. Evaluating a Multimedia Authoring Tool, *Journal of the American Society of Information Science*, *48*, 9 (1997), 1004-1022.
24. John, B. E. & Packer, H. Learning and Using the Cognitive Walkthrough Method: a Case Study Approach, *Proc. CHI'95*, ACM Press (1995), 429-436.
25. Karat, C.-M., Campbell, R., & Fiegel, T. Comparison of Empirical Testing and Walkthrough Methods in Usability Interface Evaluation, *Proc. CHI'92*, ACM Press (1992), 397-404.
26. Molich, Rolf, User testing, Discount user testing, 2003, www.dialogdesign.dk.
27. Molich, R., Ede, M. R., Kaasgaard, K., & Karyukin, B. Comparative Usability Evaluation, *Behaviour & Information Technology*, *23*, 1 (2004), 65-74.
28. Nielsen, J. *Usability Engineering*, Morgan Kaufmann Publishers, San Francisco, CA, 1993.
29. Nielsen, J. Finding Usability Problems Through Heuristic Evaluation, *Proc. CHI'92*, ACM Press (1992), 373-380.
30. Pace, S. A Grounded Theory of the Flow Experiences of Web Users, *International Journal of Human-Computer Studies*, *60* (2004), 347-363.
31. Sawyer, P., Flanders, A., & Wixon, D. Making a Difference - The Impact of Inspections, *Proc. CHI'96*, ACM Press (1996), 376-382.
32. Spencer, R. The Streamlined Cognitive Walkthrough Method, Working Around Social Constraints Encountered in a Software Development Company, *Proc. CHI'2000*, (2000), 353-359.
33. Strauss, A. & Corbin, J. *Basics of Qualitative Research - Techniques and Procedures for Developing Grounded Theory*, Sage Publications, California, (1998).
34. Szczur, M. Usability Testing - on a Budget: a NASA Usability Test Case Study, *Behaviour & Information Technology*, *13* (1994), 106-118.
35. Vredenburg, K., Mao, J.-Y., Smith, P. W., & Carey, T. A Survey of User-Centered Design Practice, *Proc. CHI 2002*, ACM Press (2002), 472-478.
36. Wilson, S., Bekker, M., Johnson, P., & Johnson, H. Helping and Hindering User Involvement - a Tale of Everyday Design, *Proc. CHI'97*, ACM Press (1997), 178-185.
37. Wixon, D. Evaluating Usability Methods: Why the Current Literature Fails the Practitioner, *interactions*, *10*, 4 (2003), 29-34.
38. Zirkler, D. & Ballman, D. R. Usability Testing in a Competive Market: Lessons Learned, *Behaviour and Information Technology*, *13*, 1&2 (1994), 191-197.