Two psychology-based usability inspection techniques studied in a diary experiment

Kasper Hornbæk
Natural Sciences ICT Competence Center,
University of Copenhagen, Denmark
kash@diku.dk

Erik Frøkjær
Datalogisk Institut
Københavns Universitet, Denmark
erikf@diku.dk

ABSTRACT
Inspection techniques are widely used during systems design as a supplement to empirical evaluations of usability. Psychology-based inspection techniques could give important insights into how thinking shapes interaction, yet most inspection techniques do not explicitly consider users' thinking. We present an experiment comparing two psychology-based inspection techniques, cognitive walkthrough (CW) and metaphors of human thinking (MOT). Twenty participants evaluated web sites for e-commerce while keeping diaries of insights and problems experienced with the techniques. Using MOT, participants identified 30% more usability problems and in a reference collection of problems achieved a broader coverage. Participants preferred using the metaphors, finding them broader in scope. An analysis of the diaries shows that participants find it hard to understand MOT, while CW limits the scope of their search for usability problems. Participants identified problems in many ways, not only through the techniques, reflecting large differences in individual working styles.

Author Keywords
Usability evaluation techniques, inspection techniques, metaphors of human thinking, cognitive walkthrough, psychology, individual differences.

ACM Classification Keywords
H.5.2 [Information Interfaces and Presentation]: User Interfaces–Evaluation/methodology, User-centered design.

INTRODUCTION
A core activity in human-computer interaction studies for the past ten years has been to develop effective usability inspection techniques. Inspection techniques aim at uncovering potential usability problems by having evaluators inspect the user interface with a set of guidelines or questions [17]. Inspection techniques are widely used for early integration of evaluation into design and to supplement empirical evaluation techniques. Well-known inspection techniques include heuristic evaluation, which uses heuristics such as 'Be consistent' or 'Prevent errors' [18], p. 249; and cognitive walkthrough (CW) [12,30], where evaluators ask questions related to how users perceive the user interface and plan task-related actions.

However, most inspection techniques do not explicitly consider users' thinking. Of the first 37 guidelines in the classic collection by Smith and Mosier only 10 refer to users' thinking or psychological principles, and then only with superficial phrases such as 'Most users will forget to do it' or 'People cannot be relied upon to pay careful attention to such details' [26], p. 34. Heuristic evaluation [18] only mentions the user explicitly in two heuristics, and 'minimize users' memory load' is the only heuristic that comes close to considering users' thinking. Even in cognitive walkthrough, developed with a basis in psychological theories of exploratory learning [12], refinement has led to less emphasis on the psychological basis. In [30], the original list of nine questions (some with sub-questions) was reduced to four. Recently, in the so-called stream-lined cognitive walkthrough [28], only two questions are asked with no reference to psychological theory: 'Will the user know what to do at this step?' and 'If the user does the right thing, will they know that they did the right thing, and are making progress towards their goal?' [28], p. 355. So most inspection techniques consider users' thinking only vaguely, thus ignoring potential important insight into how thinking shapes interaction.

Taking up this challenge, we have proposed an inspection technique based on metaphors of human thinking (MOT) [4,6]. This technique builds upon introspective psychology as described by William James in the classical book *The Principles of Psychology* [9] and by Peter Naur in *Knowing and the Mystique of Logic and Rules* [16]. An experiment [7] comparing MOT to heuristic evaluation showed that MOT uncovers more of the usability problems that were assessed as being severe on users and complex to repair. However we do not know how MOT compares to cognitive walkthrough, nor do we have any detailed data on which problems evaluators experience when using MOT.

The aim of this study is to compare the effectiveness of inspection by metaphors of human thinking with cognitive

NordiCHI '04, October 23-27, 2004 Tampere, Finland

walkthrough. As a supplement to quantitative data from the evaluations, participants are required to keep a diary during the evaluation to shed light on problems and insights experienced when using the techniques. Data from the experiment will help improve MOT and CW, and identify strengths and weaknesses of the techniques.

First we briefly describe the evaluation techniques investigated, providing more detail on MOT as it is expected to be unknown to many readers. Next, we describe the procedure for this experiment including the diary writing. Finally, the results are presented and discussed.

DESCRIPTION OF INSPECTION TECHNIQUES

Cognitive walkthrough (CW)

Cognitive walkthrough focuses on evaluating how easy it is to learn an interface, especially by exploration [12,30]. The evaluation procedure consists of a preparation phase in which the evaluator chooses tasks to be analyzed, and characterizes potential users of the application. After the preparation, the evaluation itself consists of a walkthrough of the action sequences needed to complete sample tasks, during which four questions are considered [30, p. 112]:

1. Will the user try to achieve the right effect?
2. Will the user notice that the correct action is available?
3. Will the user associate the correct action with the effect trying to be achieved?
4. If the correct action is performed, will the user see that progress is being made toward solution of the task?

In walking through the actions, the evaluator crafts credible stories of successes or failures based on users' background knowledge and goals, thereby identifying usability problems with the interface.

Cognitive walkthrough has been extensively evaluated and modified, see [2,8,10,11,22,28].

Metaphors of human thinking (MOT)

We shall here outline MOT, for a comprehensive description of the technique see [4,6,7]. The basic idea of MOT is to walkthrough typical tasks with the interface while keeping in mind the five metaphors of human thinking described below. The use of metaphors as a descriptive device is intended to stimulate, generate insight, and break fixed conceptions.

Metaphor of Habit Formation

Habits shape most of our thought activity and behaviour, e.g. as physical habits, automaticity, all linguistic activity, and habits of reasoning. The metaphor is: Habit formation is like a landscape eroded by water. We propose this metaphor to indicate how a person's formation of habits leads to more efficient actions and less conscious effort, like a landscape through erosion adapts for a more efficient and smooth flow of water. Creeks and rivers will, depending on changes in water flow, find new ways or become arid and sand up, in the same way as a person's habits will adjust to new circumstances and, if unpracticed, vanish.

In usability evaluation, this metaphor calls for considering: Are existing habits supported? Can effective new habits, when necessary or appropriate, be developed? Can the user use common key combinations? Is it possible for the user to predict, a requisite for forming habits, the layout and functioning of the interface?

In design, there is an abundance of examples of user interfaces that violate human habits. One example is adaptive menus, used for example in Microsoft Office 2000. Adaptive menus change the layout of the menu according to how often menu items are used, for example by removing or changing the position of items seldom used. However, adaptive menus make it impossible to form habits in the selection of menu items, since their position may be different from when they were previously selected. A study by Somberg [27] showed the efficiency of constant position placement of menu items compared to menus that change based on use frequency. Somberg, however, did not explicitly link habit formation to the usefulness of constant placement of menu items.

Metaphor of the Stream of Thought

Human thinking is experienced as a stream of thought—in the continuity of our thinking, the richness and wholeness of a person's mental objects, of consciousness, and subjective life including experiences and feelings. The metaphor is: Thinking as a stream of thought. This metaphor was proposed by James to emphasize how consciousness does not appear to itself chopped up in bits: 'Such words as "chain" or "train" do not describe it fitly. It is nothing jointed; it flows'. Particular issues can be distinguished and retained in a person's stream of thought with a sense of sameness, as anchor points, which function as 'the keel and backbone of human thinking' [9, vol. I, p. 459].

In usability evaluation, this metaphor calls for considering: Is the flow in users' thought supported in the interface by recognizability, stability and continuity? Does the application make visible and easily accessible such interface elements that relate to the anchor points of users' thinking about their tasks? Does the application help users to resume interrupted tasks?

In design, a simple, yet effective, attempt to recreate part of the richness of the stream of thought when users return to resume interrupted work, is Raskin's [23] design of the Canon Cat. When the Canon Cat is started, the display immediately shows up as it was before work was suspended. Not only does this allow the user to start thinking about the task at hand while the system is booting. It also provides help in remembering and recreating the stream of thought as it was when work was interrupted.

Metaphor of the Dynamics of Thinking

Here the dynamics of human thinking are considered, i.e. awareness shaped through a focus of attention, the fringes of mental objects, association, and reasoning. The metaphor is: Awareness as a jumping octopus in a pile of rags. This metaphor was proposed by Naur [16] to indicate how the state of thought at any moment has a field of central awareness, that part of the rag pile in which the body of the octopus is located; but at the same time has a fringe of connections and emotions, illustrated by the arms of the octopus stretching out into other parts of the rag pile. The jumping about of the octopus indicates how the state of human thinking changes from one moment to the next.

In usability evaluation, this metaphor calls for considering: Are users' associations supported through flexible means of focusing within a stable context? Do users associate interface elements with the actions and objects they represent? Can words in the interface be expected to create useful associations for the user? Can the user switch flexibly between different parts of the interface?

In design, an example of a problematic solution is a use of modal dialog boxes that prevents the user from switching to potentially relevant information—in Microsoft Word 2002, for example, it is not possible to switch back to the document to look for a good file name once the 'save as ...' dialog has begun.

Metaphor of the Incompleteness of Utterances

Here the focus is on the ephemeral character of utterances and their incompleteness in relation to the underlying thinking. The metaphor is: A person's utterances relate to the person's insights as splashes over the waves to the rolling sea. This metaphor was proposed by Naur [16] to emphasize how utterances are incomplete expressions of the complexity of a person's current mental object, in the same way as the splashes tell little about the sea below.

In usability evaluation, this metaphor calls for considering: Does the application support changing and incomplete utterances? Are alternative ways of expressing the same information available? Are interpretations of users' input made clear? Does the application make a wider interpretation of input than users intend or are aware of?

For design, one implication of the metaphor of utterances as splashes over the waves is that we must expect users to describe the same objects and functions incompletely and in a variety of ways. Furnas et al. [5] investigated the diversity in words used for describing commands and everyday objects. On the average, two participants described the same command or object by the same term with less than 20% probability. The most popular name was chosen only in 15-35% of the cases. Furnas et al.'s suggestion for relieving this problem is called the unlimited alias approach, where terms unknown to the system may be interactively related to existing commands or object names. This proposal is coherent with the metaphor and uses interactivity to clarify the intentions of the user. However, it would partly go against the metaphor of habit formation.

Metaphor of Knowing

Human knowing is always under construction and incomplete. The metaphor is: Knowing as a building site in progress. This metaphor was proposed by Naur [16] and meant to indicate the mixture of order and inconsistency characterizing any person's insight. These insights group themselves in many ways, the groups being mutually dependent by many degrees, some closely, some slightly. As an incomplete building may be employed as shelter, so the insights had by a person in any particular field may be useful even if restricted in scope.

In usability evaluation, this metaphor calls for considering: Are users forced by the application to depend on complete or accurate knowledge? Is it required that users pay special attention to technical or configuration details before beginning to work? Do more complex tasks build on the knowledge users have acquired from simpler tasks? Are users supported in remembering and understanding information in the application?

In design, mental models have been extensively discussed. Consider as an example Norman's [20] description of the use of calculators. He argues that the use of calculators is characterized by users' incomplete understanding of the calculators, by the in-stability of the understanding, by superstitions about how calculators work, and by the lack of boundaries in the users' understanding of one calculator and another. These observations by Norman are perfectly coherent with the ideas expressed by the metaphor of knowing.

DIARY EXPERIMENT

The aim of the experiment is to compare how 20 participants evaluate and redesign web sites using MOT and CW. Data comprise problem lists, redesigns, diaries, and participants' preferences.

Participants

Twenty participants, 3 women and 17 men, participated in the experiment as part of a computer science graduate course in experimental design. On the average, participants were 27 years old and had studied computer science for 5.9 years. Three quarters of the students had previously attended courses on human-computer interaction; half had designed user interfaces in their part-time jobs.

Design and procedure

The experiment varies inspection technique (MOT vs. CW) and web site within participants. Participants were randomly assigned to one of two orders in which they use the inspection techniques; the order of the web sites was fixed.

Every participant used one week to complete an evaluation and a redesign for each web site; week 1 and week 2 used similar procedures. During the first half of each week, the participants first received a description of the inspection technique to be used and the web site to evaluate. Next, they had three days to evaluate the web site. When evaluating, participants knew that they later on had to redesign. During the second half of each week, participants were asked to redesign the three most problematic parts of the web site with respect to usability.

After having completed both redesigns, participants wrote a comparison of the techniques used. They also described which inspection technique they preferred and why. Throughout the evaluation and redesign activities participants kept a detailed diary. Figure 1 summarizes the design of the experiment.

Web sites evaluated

Each of the techniques was used to evaluate and redesign an e-commerce web site. The site evaluated in the first week was http://www.gevalia.com; the site in the second week was http://www.jcrew.com. Both sites are included in a large professional study of e-commerce sites [19], which offers insights into usability problems of e-commerce sites.

Inspection techniques and problem lists

The metaphors-of-human-thinking technique (MOT) was described to participants by a version of [6] that had the authors' names replaced by pseudonyms. One of the 20 students had an uncertain knowledge about the authors' involvement in MOT; the rest appeared convinced by the pseudonyms. As a description of cognitive walkthrough (CW) participants received [30], widely recognized as the classic presentation of the cognitive walkthrough technique. The descriptions of the inspection techniques were of comparable length (CW: 36 pages, ~14.000 words, MOT: 23 pages, ~10.000 words) and both in English.

To make comparisons between techniques easier, participants were suggested to use around two hours on the evaluation, disregarding any timing information mentioned in the description of the inspection technique.

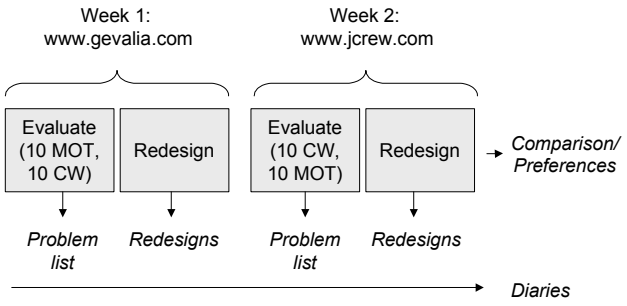

Figure 1. Experimental procedure. Text in italics refers to results of the evaluation and the redesign process that we analyze.

Each participant documented the evaluation of each week in a problem list with fields for characterizing the problems, for noting which metaphors/criteria had helped identify each problem, and for assigning a severity rating. The severity ratings used were based on a commonly used scale [13], p. 111: *Rate 1* is given to a critical problem; gives rise to frequent catastrophes; should be corrected before the system is put into use. This grade is for those few problems that are so serious that the user is better served by a delay in the delivery of the system; *Rate 2* is given to a serious problem; occasionally gives rise to catastrophes; should be corrected in the next version; and *Rate 3* is given to a cosmetic problem; should be corrected when an opportunity arises.

To investigate the quality of the identified problems, we compared them to a reference collection of important usability problems with e-commerce web sites [19]. This reference collection is based on think-aloud experiments with a range of e-commerce sites, including the two tested in the present experiment. Each problem found by the participants was compared to the 207 problems in the reference collection for either a full or a partial match; as this matching was done with relatively little information and no possibility for clarification with participants, we consider the combination of full and partial matches in the analysis below.

Redesign of web sites

As part of the experiment, participants were asked to redesign the web sites evaluated. The aim of this redesign was to force participants to choose the most important usability problems found. In addition, we wanted to investigate if participants change their perception of the usability problems when redesigning.

Each redesign of a part of the web site was described as each participant found fit, but was supposed to include a list of the problems that the redesign sets out to solve, a rationale for the redesign, and a detailed description of the redesign.

Diary writing

When evaluating and redesigning the web sites, participants were asked to keep a diary of their work. Diaries have previously been used to study for example information seeking [29], systems development [14], human-computer interaction [21,24,25], and the use of evaluation techniques [8,11].

The diaries used in this study are exemplified in Figure 2. The diary has a row for every half hour of the day (24 hours). The second field in each row allows participants to enter a description of the activity they are performing. In addition, the diary has for every half-hour interval fields for

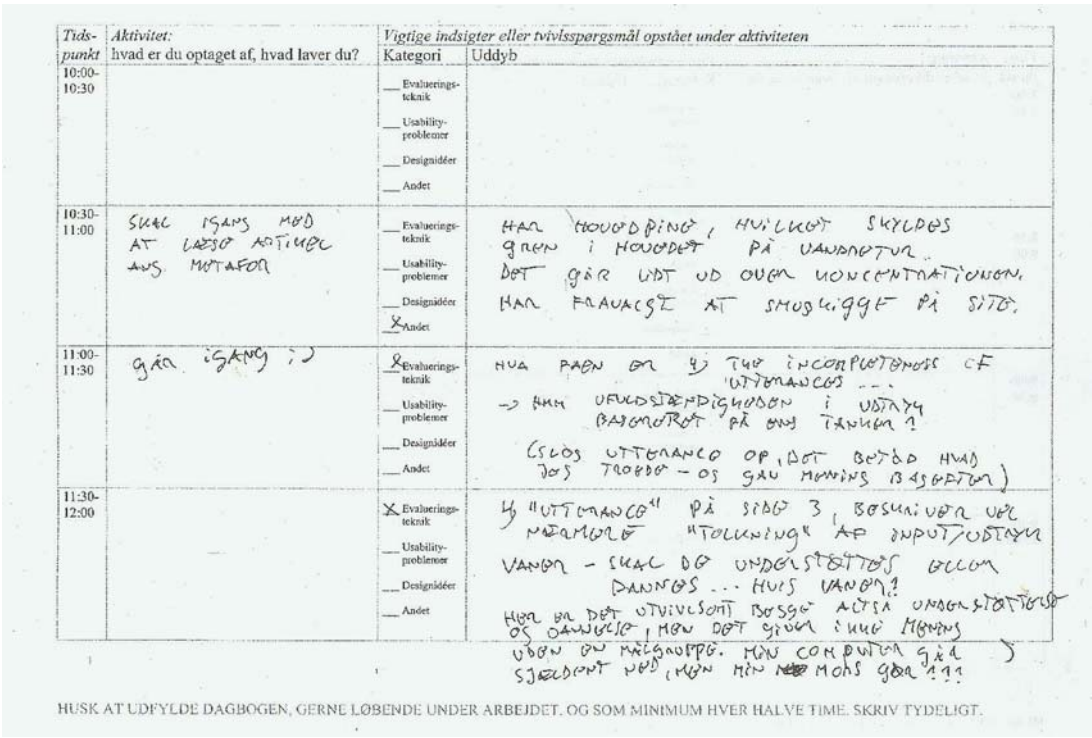

Figure 2. An example of the paper diaries used. The first column indicates half-hour intervals of the day and night. The second column lets participants describe their activities, and the third and forth column let participants categorize and describe questions and insights. The diary shown include an activity from 11:00-11:30, labeled 'starting :)', referring back to the previous half-hour activity, 'Is about to start reading the paper about the metaphors'. The description of the activity contains an example of a problem with understanding the evaluation technique: 'What is 4) The Incompleteness of utterances … →Hmm, incompleteness in expressions based on ones thoughts (looked up utterance, it meant what I thought and made sense afterwards)'. Also note that participants were instructed to and did in fact include several remarks regarding their more general activities, e.g. in the half-hour from 10:30-11:00 this participant writes 'Have a headache, caused by a tree branch hitting my head while hiking'.

entering specific insights or questions. Each time participants entered an insight or a question, they also categorized whether the insight or question related to the inspection technique, usability problems, design ideas, or something else.

A major problem in using diaries is that participants sometimes forget to fill them out [21]. We reminded participants by e-mail on the first day to fill out their diaries. In addition, the instructions guiding the experiment suggested participants work with their diaries placed in front of them.

RESULTS

Number and quality of problems identified

Analysis of variance show that participants identify significantly more problems using MOT compared to CW, $F(1,19)=8.68$, $p<0.001$. On average, participants identify 11.8 (SD=7.52) with MOT and 9.0 (SD=8.18) problems with CW, that is 31% more. In raw numbers, 13 participants find more problems with MOT, 3 identify the same number of problems, and 4 identify more with CW.

We find no difference in the severity ratings assigned by participants to the usability problems, $F(1, 19)=3.35$, $p>.05$.

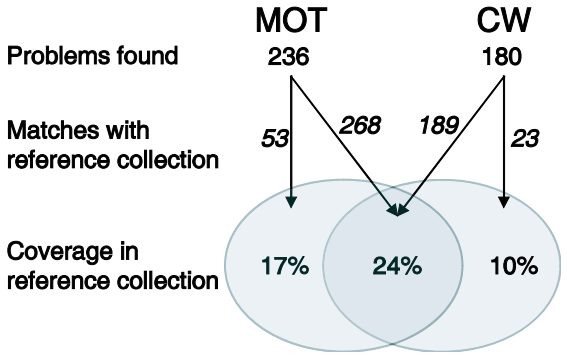

Figure 3. Relation between the problems found by participants and the reference collection of usability problems important for designers of e-commerce web sites [19]. ***Note*** **that one problem found by a participant may match more than one problem in the reference collection. On average, MOT achieves a better coverage of the reference collection than CW.**

On the average participants using MOT assess the severity of the problems as 2.31 (SD=.72); using CW average severity is 2.25 (SD=.69).

Figure 3 summarizes the relation between usability problems found by participants and the reference problems described in [19]. The figure shows that both techniques succeed in finding problems that hit the reference collection (9% of the problems identified could not be mapped to the reference collection; however the majority of those seemed relevant); and in combination the two techniques achieve 51% coverage of the collection (note that only two of the web sites studied in [19] were used here).

Using MOT, participants identify usability problems covering a broader group of problems in the reference collection, $F(1,19)=4.48$, $p<.05$. Among all evaluators, MOT identifies 36 problems (17%) in the reference collection that CW did not find; CW finds only 21 problems (10%) in the reference collection that MOT did not find.

Subjective preferences and comments

In the final comparison of techniques, 15 participants preferred using MOT for the usability evaluation; four preferred CW, and one participant presented arguments for preferring both. This difference is significant $\chi^2(1, N=19)=6.37$, $p<.05$. In explaining their preferences, seven participants argued that they found more and broader problems with MOT, for example 'I prefer the first technique (metaphor based evaluation) because it catches different kinds of problems'. In addition, some participants found evaluation with MOT to be faster, and two participants commented that they got better ideas for how to redesign the site.

The four participants preferring CW explained that they found the technique more easy to follow when evaluating, '[it is] easier to overview, seems like a recipe that you just have to follow'. Some participants who preferred MOT made similar comments:

> If you don't know how to evaluate a web site it is good that the technique [CW] gives you a systematic procedure for doing so.

It should be noted that at least three participants argued that their preferences depended on what web site they were going to evaluate.

ANALYSIS OF DIARIES

Below we present an analysis of the diaries aimed at capturing patterns across participants regarding MOT and CW. The analysis is based on an extraction from the diaries of 174 entries each containing one or more comments concerning the inspection techniques or redesigns. These comments fall in four major groups: (1) comments on the inspection technique being used, (2) comments comparing techniques, (3) comments on the procedure of evaluation, and (4) comments on redesign. To identify patterns among diary comments, both authors independently read all comments and grouped them according to similarities among comments. The resulting groupings were worked together and form the basis of the presentation below. Tables 1 and 2 contain a summary of insights and problems on inspection techniques reported by two or more participants. In addition, we present some findings relating more to the evaluation process, rather than to the specific techniques.

Insight/problem	*N*
Description of the metaphors challenging to understand, e.g. the octopus metaphor	7
Reflections and associations concerning the technique	5
Key questions and examples are useful when reading and when evaluating	4
Learning or changing of opinions about technique	6
Difficult to use metaphors in isolation when evaluating	5

Table 1. The main insights and problems with MOT. N is the number of different participants making similar comments.

Insights and problems experienced—MOT

Several participants make general, positive comments in their diaries about MOT, especially that the key questions and examples help the understanding and using of MOT. These participants write, for example, that 'Examples and key questions are very helpful' and '[I] use the table with key questions during the evaluation'.

Although there is no clear pattern among the four participants who describe general problems with MOT, several aspects of the MOT description are challenging to understand. One participant writes:

> The metaphors are more confusing than useful. It requires good faith to see the connection to the examples. Good points, however.

Other participants write that particular parts of the description are hard to understand. Three participants are somewhat unsure about the metaphor concerning the jumping octopus. Two participants mention that when reading about the metaphors on the stream of thought and the dynamics of thinking, they find it hard to understand what is meant by grouping tasks, e.g.

> [I] do not understand 'to group tasks' – I am only considering one task? Is it subtasks? Is it about getting an overview of the parts that have to be done?

For MOT we find a higher number of entries during reading that are best summarized as reflections and

associations. At least five participants had written one or more entries of this kind, for example:

> Support already existing habits and the development of new ones. Can this lead to some kind of conflict?

The diaries from at least six participants using MOT have comments on problems with understanding the technique that hours or days later are followed with a comment that the problem has been fully or partly resolved, i.e. it appears that participants learn and change their opinions about the technique. For example the participant quoted above on the confusion felt when reading the metaphors only half an hour later writes:

> The explanation of the metaphors makes sense, more logical now. Good and stimulating points concerning habits, especially the unintended effects of habits.

When doing the evaluation, participants noted that it is difficult to use the metaphors in isolation because they are interrelated; especially, participants appear to have had problems categorizing the problems to a metaphor, for example one participant writes:

> Have difficulty in using the metaphors separately, because the questions and the metaphors do not always appear to be related. Have difficulty categorizing the problems to a particular metaphor, because they are connected.

Other two participants noted that they mainly were considering the key questions, not so much the metaphors, during the evaluation.

Overall, the description of MOT seems fairly difficult to read for some participants; however, as the reading progresses and evaluation begins participants seem to face few problems in conducting the evaluation.

Insights and problems experienced—CW

Five participants comment in their diaries on various positive aspects of CW, including that it is 'well explained and exemplified' and that it is 'an exciting way of going through the users' tasks'. Overall, the technique appears to be easy to read and make sense of. However, participants also mention various general difficulties with CW, including that the description of CW is somewhat abstract. For example, 'I do not experience any questions of doubt, but I find the first part of the technique a little abstract.'

Participants report a number of specific problems related to understanding and using CW. For example, participants are unsure of how to handle tasks for which several sequences of actions could lead to the solution of the task. Among several possible action sequences, four participants raise questions of doubt how one sequence should be chosen for doing the walkthrough, wondering '[should] all possible sequences be listed?' The notion of correct action, used in the criteria for evaluating, three participants find hard to understand. Finally, three participants find it hard to put themselves in the place of potential users for whom the evaluation of the interface should be new, for example:

Insight/problem	*N*
Cognitive walkthrough is easy to read and makes sense	5
Problems in understanding the technique, for example how to handle many actions sequences and what is meant by 'correct action'	6
Limitations in the scope of problems that the technique help identify	8
During evaluation, it is hard to put yourself in user's place	3
General critique of cognitive walkthrough, especially that the focus is overly narrow	7

Table 2. The main insights and problems with CW. N is the number of different participants making similar comments.

> It is hard to ignore the hours used on looking through a web-site, a customer would use less than 5-10 minutes to learn to navigate.

Eight participants make various comments concerning the restricted scope of the technique. For example, four participants were concerned that none of the evaluation criteria help identify missing functions in the user interface. One participant writes:

> Cognitive walkthrough is not covering the possibility that the correct action is not available. For example, that it is impossible [on one of the web sites evaluated] to register an address when you are living in Denmark.

A related point is the criticism by some participants that CW does not help assess whether it makes sense to solve a task in a particular way. For example 'how to characterize a shortcoming? If the user can complete the process, but has to do it in a roundabout way'. Participants also commented that the focus of the technique was rather 'narrow-minded'.

Overall, participants seem able to easily read the description of CW. During the evaluation, however, participants felt that CW limits the scope of usability problems that can be identified.

The evaluation process

The analysis of the diaries with respect to the evaluation process yielded four findings. First, we looked at the overall patterns of activity between the two techniques by coding the activities reported in the diaries into phases where the main focus is on reading, orientation on the web-site, preparing evaluation, evaluation, documentation of evaluation, redesign, and other. Assessed from this coding, participants spend approximately similar time, around 3

hours, on reading the techniques (CW: M=194 min.; MOT: M=233 min.); performing the evaluation (CW: M=132 min.; MOT: M=129 min.); and making the redesigns (CW: M=393 min.; MOT: M=374 min.).

Second, the diaries show that usability problems are found in a variety of ways, and not just using the techniques as prescribed. At least ten participants identify problems already before reading the description of the inspection technique, or while initially orienting themselves on and gaining an overview of the web site. During her first visit on the web site before starting the evaluation procedure, one participant writes:

> Identification of immediate problems and some ideas for tasks. Especially the questionnaire [on the web site] is a disaster. The menu in the left side sometimes disappears. No systematic information on whether a word or a label is clickable…

That participant ends up reporting on her problem list three problems regarding the questionnaire.

Five participants also describe how they find problems during the evaluation that are not generated by specific metaphors or criteria. For example, one participant writes 'what if a problem can not be placed under a criterion?' and another writes:

> Should ALL usability problems be written down? Also those that are not found directly by cognitive walkthrough?

Indeed, even after finishing the evaluation procedure, participants continue to identify problems, as illustrated by the following remark in a diary:

> I am finding more usability problems as I fill out the problem lists. These are added to the problem list […]

Third, participants have difficulties assigning severity ratings. One person writes:

> [I] am in doubt about the structuring, but I think it is hard to assess severity ratings on the issues identified. It depends on who you are.

Interestingly, however, as reported in the diaries participants seem to have little trouble choosing problems to be addressed in their redesign.

Fourth, during the course of the evaluation participants change their opinion on what they consider a usability problem, e.g. some participants change their opinion about problems when redesigning. One participant writes that

> [I] have come to the conclusion that the buying procedure is really not so complicated that it will give errors for the user.

The same participant had on his problem list noted as a serious problem the cumbersome buying procedure. Conversely, at least five participants identify problems when redesigning which they had not previously been aware of, for example:

> Looking at a screen dump makes me aware of new usability problems. What am I to do with problems I have just discovered?

The diaries do not make it clear whether these problems arise from longer experience with the web site; or whether having to redesign the web site changes the evaluation focus.

DISCUSSION

In the direct comparison of techniques, MOT seems to outperform CW. MOT finds more problems and achieves a better coverage of the usability problems in the reference collection. Participants also prefer using MOT. In addition, the problems reported in the diaries about MOT mainly concern the process of understanding the technique; while the participants using CW often felt the scope of their search for usability problems restricted. The current experiment thus supports our previous study comparing MOT to heuristic evaluation [7], in showing how MOT performs better on several important measures.

Although significant and quite large differences between the two inspection techniques are found, two observations stand out from the diaries as striking. First, the dynamics of inspection, e.g. in finding problems upon initial orientation on the web sites or in changing ones mind about severity ratings and whether something is really a problem, suggests that static comparisons of evaluation techniques through lists of usability problems identified are an approach with limitations. In addition, reconsidering the problem lists knowing that redesigns should be proposed seems to change how evaluators assess problems; further work is needed to understand better the relation between evaluation and redesign.

Second, the participants' personal reading and working habits play a major role in shaping their interpretation and use of the inspection techniques. Of course the participants in this study are novices still learning to use the two techniques. But it is hardly to be expected that more competent usability evaluators or experts would be working in a more uniform or rule-based manner than these novices. While close studies of method usages and work processes of usability experts are still few, e.g. [8,11], theoretical and empirical studies of systems developers, for instance, have shown how even highly structured methods are used and understood quite differently [1,15]. More generally it has been argued that people in their learning processes from novice through to expert level develop a more and more personal and context dependent working style [3].

A consequence of these findings is that it is useful to develop an arsenal of usability evaluation techniques that are adequate and convenient for different combinations of

users, task domains, tools and interaction forms. It might be even more important to develop and maintain teaching and learning courses where people studying systems design and usability evaluation can achieve personal experiences with a range of effective tools and working styles.

CONCLUSION

Despite a promise from psychology-based inspection techniques to give important insights into how thinking shapes interaction, most inspection techniques do not explicitly consider users' thinking. We compared two techniques that do, namely cognitive walkthrough (CW) and the metaphors-of-human-thinking technique (MOT). Using MOT participants identified 30% more problems with the web sites inspected and achieved a broader coverage of a reference collection of usability problems. Participants also preferred using MOT. Judging from diaries written by participants while performing the inspections, MOT is challenging to understand, but few difficulties with the technique where reported during the inspection. Conversely, CW seems easier to understand, but during the inspection some participants feel that it limits the scope of their search for usability problems.

The analysis of the diaries suggests problems in many comparative studies of usability evaluation methods. Static lists of usability problems, for example, can be misleading because their interpretation, even for the individual evaluator, changes over time and across contexts. Also, we suggested how the role of techniques in usability inspection processes should be rethought, for example by putting less emphasis on rules guiding usability inspection and more on evaluators' intuition concerning what creates usability problems.

ACKNOWLEDGEMENTS

We are grateful to Aran Lunzer and Peter Naur for their comments on a draft of this paper.

REFERENCES

1. Bansler, J. & Bødker, K. A reappraisal of structured analysis: design in an organizational context. *ACM Transactions on Information Systems*, *11*, 2 (1993), 165-193.
2. Blackmon, M., Polson, P., Kitajima, M., & Lewis, C. Cognitive walkthrough for the web. *Proc. CHI 2002*, CHI Letters 4(1), 463-470.
3. Dreyfus, H. & Dreyfus, S. *Mind Over Machine*, The Free Press, New York, NY, 1986.
4. Frøkjær, E. & Hornbæk, K. Metaphors of human thinking in HCI: Habit, stream of thought, awareness, utterance, and knowing. *Proc. HF/OzCHI 2002*.
5. Furnas, G., Landauer, T., Gomez, L., & Dumais, S. The vocabulary problem in human-system communication. *Communications of the ACM*, *30*, 11 (1987), 964-971.
6. Hornbæk, K. & Frøkjær, E. Evaluating user interfaces with metaphors of human thinking. *Proc. User Interfaces for All*, Lecture Notes in Computer Science 2615, Springer-Verlag (2002), 486-507.
7. Hornbæk, K. & Frøkjær, E. Usability Inspection by Metaphors of Human Thinking Compared to Heuristic Evaluation. To appear in *International Journal of Human-Computer Interaction* (2004).
8. Jacobsen, N. & John, B. Two case studies in using cognitive walkthroughs for interface evaluation. Technical report *CMU-CS-00-132* (2000).
9. James, W. *The Principles of Psychology*. Henry Holt & Co., 1890.
10. Jefferies, R., Miller, J., Wharton, C. & Uyeda, K. User interface evaluation in the real world. *Proc. CHI'91*, ACM Press (1991), 119-124.
11. John, B. & Packer, H. Learning and using the cognitive walkthrough method: a case study approach. *Proc. CHI'95,* ACM Press (1995), 429-436.
12. Lewis, C., Polson, P., Wharton, C., & Rieman, J. Testing a walkthrough methodology for theory-based design of walk-up-and-use interfaces. *Proc. CHI'90*, ACM Press (1990), 235-242.
13. Molich, R. *Brugervenlige Edb-Systemer (in Danish)*. Teknisk Forlag, 1994.
14. Naur P. Program development studies based on diaries. Green, T., Payne, S., & van der Veer, G. *Psychology of Computer Use.* Academic Press, 1983, 159-170.
15. Naur P. Intuition in software development. Ehrig, H, Floyd, C., Nivat, M., & Thatcher, J. *Formal Methods and Software Development, Vol. 2,* Lecture Notes in Computer Science 186*,* Springer Verlag, 1985, 60-79.
16. Naur, P. *Knowing and the Mystique of Logic and Rules*. Kluwer Academic Publishers, Dordrecht, 1995.
17. Nielsen, J. & Mack, R. L. *Usability Inspection Methods*. Wiley and Sons Inc., 1994.
18. Nielsen, J. & Molich, R. Heuristic evaluation of user interfaces. *Proc. CHI'90*, ACM Press (1990), 249-256.
19. Nielsen, J., Molich, R., Snyder, C., & Farrell, S. *E-Commerce User Experience.* Nielsen Norman Group, 2001.
20. Norman D. Some observations on mental models. Gentner, D. & Stevens, A. *Mental Models,* Erlbaum, 1983, 7-14.
21. Palen, L. & Salzman, M. Voice-mail diary studies for naturalistic data capture under mobile conditions. *Proc. CSCW 2002,* CHI Letters 4(3), 87-95.

22. Pinelle, D. & Gutwin, C. Groupware walkthrough: adding context to groupware usability evaluation. *Proc. CHI 2002*, CHI Letters 4(1), 455-462.

23. Raskin, J. *The Humane Interface: New Directions for Designing Interactive Systems*. Addison-Wesley, 2000.

24. Rieman, J. A field study of exploratory learning strategies. *ACM Transactions on Computer-Human Interaction*, *3*, 3 (1996), 189-218.

25. Sellen, A. & Harper, R. Paper as an analytic resource for the design of new technologies, *Proc. CHI'97*, ACM Press (1997), 319-326.

26. Smith, S. L. & Mosier, J. N. Guidelines for designing user interface software, *ESD-TR-86-278* (1986).

27. Somberg, B. L. A comparison of rule-based and positionally constant arrangements of computer menu items. *Proc. CHI+GI'87*, ACM Press (1987), 255-260.

28. Spencer, R. The streamlined cognitive walkthrough method, working around social constraints encountered in a software development company. *Proc. CHI 2000*, CHI Letters 2(1), 353-359.

29. Toms, E. G. & Duff, W. "I spent 1 1/2 hours sifting through one large box. ...": diaries as information behavior of the archives user: lessons learned. *Journal of the American Society for Information Science and Technology*, *53*, 14 (2002), 1232-1238.

30. Wharton C., Rieman, J., Lewis, C. & Polson, P. The cognitive walkthrough method: a practitioner's guide. Nielsen, J. & Mack, R. L. *Usability Inspection Methods.* John Wiley & Sons, 1994, 105-140.