Metaphors of Human Thinking for Usability Inspection and Design

ERIK FRØKJÆR and KASPER HORNBÆK
University of Copenhagen

Usability inspection techniques are widely used, but few focus on users' thinking and many are appropriate only for particular devices and use contexts. We present a new technique (MOT) that guides inspection by metaphors of human thinking. The metaphors concern habit, the stream of thought, awareness and associations, the relation between utterances and thought, and knowing. The main novelty of MOT is its psychological basis combined with its use of metaphors to stimulate inspection. The first of three experiments shows that usability problems uncovered with MOT are more serious and more complex to repair than problems found with heuristic evaluation. Problems found with MOT are also judged more likely to persist for expert users. The second experiment shows that MOT finds more problems than cognitive walkthrough, and has a wider coverage of a reference collection of usability problems. Participants prefer using MOT over cognitive walkthrough; an important reason being the wider scope of MOT. The third experiment compares MOT, cognitive walkthrough, and think aloud testing, in the context of nontraditional user interfaces. Participants prefer using think aloud testing, but identify few problems with that technique that are not found also with MOT or cognitive walkthrough. MOT identifies more problems than the other techniques. Across experiments and measures of usability problems' utility in systems design, MOT performs better than existing inspection techniques and is comparable to think aloud testing.

Categories and Subject Descriptors: D.2.2 [**Software Engineering**]: Design Tools and Techniques—*User interfaces*; H.5.2 [**Information Interfaces and Presentation (e.g., HCI)**]: User Interfaces—*Evaluation/methodology*

General Terms: Experimentation, Human Factors, Measurement

Additional Key Words and Phrases: Usability evaluation, metaphors of human thinking, cognitive walkthrough, heuristic evaluation, think aloud testing, usability inspection

ACM Reference Format:

Frøkjær, E. and Hornbæk, K. 2008. Metaphors of human thinking for usability inspection and design. ACM Trans. Comput.-Hum. Interact. 14, 4, Article 20 (January 2008), 33 pages. DOI = 10.1145/1314683.1314688 http://doi.acm.org/10.1145/1314683.1314688

Authors' address: Department of Computing, University of Copenhagen, Universitetsparken 1, DK-2100 Copenhagen, Denmark; email: {erikf, kash}@diku.dk.

1. INTRODUCTION

A core activity in human–computer interaction studies over the past fifteen 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 [Nielsen and Mack 1994; Cockton et al. 2003]. Inspection techniques are widely used for early integration of evaluation into systems design and to supplement empirical evaluation techniques. Well-known inspection techniques include heuristic evaluation, which uses heuristics such as “Be consistent” or “Prevent errors” [Nielsen and Molich 1990, p. 249]; and cognitive walkthrough [Lewis et al. 1990; Wharton et al. 1994], where evaluators ask how users perceive the user interface and organize task-related actions. A large body of work has characterized the relative strengths of different inspection methods (e.g., Jeffries et al. [1991], Karat et al. [1992], and John and Packer [1995].

Existing inspection techniques suffer, in our view, from two shortcomings. First, they rarely consider users’ thinking explicitly. Heuristic evaluation [Nielsen and Molich 1990] only mentions the user explicitly in two heuristics, and “minimize users’ memory load” is the only heuristic that comes close to considering users’ thinking. Also the work of Bastien and Scapin [1995] on ergonomic criteria with its primary orientation towards the system and the interaction has little reference to users’ thinking. Even in cognitive walkthrough, developed with a basis in psychological theories of exploratory learning [Lewis et al. 1990], refinement has led to less explicit emphasis on the psychological basis. In Wharton et al. [1994], the original list of nine questions (some with subquestions) was reduced to four. In the so-called stream-lined cognitive walkthrough [Spencer 2000], only two questions are asked, neither with 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?” (p. 355). An exception to this general picture is the cognitive dimensions framework [Green and Petre 1996]. This technique is based upon a vocabulary of dimensions derived to capture the cognitively-relevant aspects of the structure of an artifact and how that structure determines the pattern of user activity. The dimensions were originally identified in studies of programming, but have shown a much broader applicability in evaluations of interactive artifacts. However, cognitive dimensions are not widely used (e.g., not mentioned in the survey by Rosenbaum et al. [2000]). Thus, many inspection techniques consider users’ thinking only vaguely, and do not make explicitly use of insight into how thinking shapes interaction.

A second shortcoming of most existing inspection techniques is that many of the guidelines or questions used are useful only for a particular device/interaction style (e.g., Windows Icons Menus Pointers-interfaces) or context of use (e.g., computing in front of the desktop). Take the classical collection of design guidelines by Smith and Mosier [1986] as an example. Although these guidelines were never intended to be used rule by rule in a usability inspection, it is striking how many guidelines were linked to a certain device/interaction style, for instance by treating issues like use of “reverse video” and “the ENTER key". Recently, in a discussion of heuristic evaluation, Preece et al. [2002] wrote "However, some of these core heuristics are too general for evaluating new products coming onto the market and there is a strong need for heuristics that are more closely tailored to specific products.", (p. 409). Pinelle et al. [2003] found too little focus on the work context in inspection techniques used for groupware and have tried to extend cognitive walkthrough to include such a focus. These are just a few examples of the limited transferability of inspection techniques to other use contexts than those originally intended for.

This article presents an inspection technique based on metaphors of human thinking, MOT. The development of MOT was spurred by our positive experiences with using metaphors of thinking when introducing computer science students, taking HCI classes, to the psychology of William James [1890] and Peter Naur [1988, 1995, 2000]. The use of metaphors as a communication device supports intuition and requires active interpretation; an effort orthogonal to developing inspection techniques that are more strictly formal and piecemeal analytical. Further, MOT attempts to address both shortcomings mentioned above; we thus conjectured that it could be an interesting supplement or alternative to existing inspection techniques.

This article seeks to empirically investigate this conjecture by (a) comparing MOT to widely used evaluation techniques, (b) studying the process of using metaphors for evaluation, and (c) comparing techniques in traditional and non-traditional use contexts across five different systems. More specifically, we report three experiments. The first experiment investigates if metaphors of thinking are useful for inspection by comparing MOT to the most widely used inspection technique, heuristic evaluation. The first experiment concludes that MOT helps novice evaluators produce problems that are seen as more serious, more complex, and more likely to persist for users. However, while novice evaluators are able to use MOT, they found it difficult to learn on their own. The second experiment compares MOT to the most widely used psychology-based inspection technique, cognitive walkthrough. While using the inspection techniques, evaluators wrote diaries, allowing us to study which problems evaluators face during inspection. The second experiment shows that MOT finds more problems than cognitive walkthrough and is preferred, and that most difficulties with using MOT are resolved as the evaluation progresses. Given these results, the third experiment seeks to challenge MOT by evaluating a speech interface and a mobile device, and by comparing to the "gold standard" of usability evaluation, think aloud testing [Landauer 1995]. The results of the third experiment are mixed, but MOT appears to perform as well as think aloud testing. Table I summarizes the conditions and main results of the three experiments.

The experiments aim to avoid the problems pointed out in recent criticisms of the validity of comparisons of evaluation techniques [Gray and Salzman 1998; Hartson et al. 2001; Wixon 2003]. The main techniques for doing so are using developers' assessment of usability problems, using a variety of data collection instruments that supplements quantitative measures, comparing techniques across different systems, and using large sample sizes. In the discussion we reflect upon the extent to which this methodological aim is reached.

Table I. Evidence from the Experiments

Exp.	Research Questions	Techniques	Participants	Evaluation Task	Classification of problems	Dependent Measures	Main Results
#1	How does MOT compare to the most widely used inspection technique, Heuristic Evaluation?	Heuristic evaluation (HE) Metaphors of human thinking (MOT)	87 computer science students (1*st* year multimedia course)	Evaluate a web site and document problems	Problems consolidated across similarity of types User action framework Problem persistence	Problem counts Participants' seriousness ratings Key manager and developer's ratings of severity and perceived solution-complexity	An equal number of problems found with MOT and HE MOT problems are more serious, more complex to repair, and more likely to persist for expert users MOT more difficult for novice evaluators to learn than HE
#2	How does MOT compare to Cognitive Walkthrough, a psychology-based inspection technique? What shapes the process of inspecting with metaphors of human thinking?	Cognitive walk-through (CW) Metaphors of human thinking (MOT)	20 computer science students (graduate course in design of empirical studies)	Evaluate and redesign two e-commerce web sites Record evaluation and redesign activities in a detailed diary	Problems compared to a reference collection of typical usability problems User action framework Problem persistence	Problem counts Participants' seriousness ratings Coverage of reference collection Analysis of redesign proposals' importance for usability, clarity, and coherence Preference	MOT finds more problems than CW MOT has a wider coverage of the reference collection than CW Evaluators prefer MOT to CW Difficulties with using MOT are resolved as the evaluation progresses
#3	Is MOT effective in usability evaluation of non-traditional interaction styles? How does MOT compare to Think Aloud Testing, the "gold standard" of usability evaluation?	Think Aloud testing (TA) Cognitive walk-through (CW) Metaphors of human thinking (MOT)	55 computer science students (third year course on systems design and HCI)	Evaluate a mobile phone and a natural-language UI Collaborate in groups to match problems All evaluators collaborate to create two goal lists using affinity diagramming	Matching of problems by groups and all participants User action framework Problem persistence	Problem counts Participants' seriousness ratings Developers' ratings of severity, complexity, and clarity Preference	Evaluators prefer using TA Evaluators identify few problems with TA not found also with MOT or CW More problems are identified with MOT than TA and CW, although small effects

Previous papers have described the technique [Frøkjær and Hornbæk 2002; Hornbæk and Frøkjær 2002] and parts of the experiments [Hornbæk and Frøkjær 2004a, 2004b]. New in this article are comparisons (a) with nontraditional user interfaces, (b) with think aloud testing, and (c) the first overall assessment of the effectiveness of MOT. The possibility of synthesizing across studies also allow us to more broadly and coherently describe the nature of the problems identified with MOT, the evaluation process with MOT, and developers' perception of the problems identified with MOT.

In the next section, an overview of metaphors of human thinking is given. Then, we present the three experiments that compare MOT to other evaluation techniques. Finally, we discuss important open questions of an inspection technique based on metaphors of human thinking.

2. OVERVIEW OF METAPHORS OF HUMAN THINKING

Metaphors of human thinking is an inspection technique based on the descriptions of human thinking made by William James [1890] and Peter Naur [1988, 1995, 2000]. Another inspiration was Jef Raskin's book *The Humane Interface*. [Raskin 2000], with its focus on the role of habits in HCI. Several of the aspects of human thinking described in these works are of critical importance to successful design of human-computer interaction: (1) the role of habit in most of our thought activity and behaviour—physical habits, automaticity, all linguistic activity, habits of reasoning; (2) the human experience of a stream of thought—the continuity of our thinking, the richness and wholeness of a person's mental objects, the dynamics of thought; (3) our awareness—shaped through a focus of attention, the fringes of mental objects, association, and reasoning; (4) the incompleteness of utterances in relation to the thinking underlying them and the ephemeral nature of those utterances; and (5) knowing—human knowing is always under construction and incomplete. The main part of our current description of the MOT technique [Hornbæk and Frøkjær 2002] consists of descriptions of these aspects of thinking by quotations from James and Naur.

The MOT technique summarizes each of these aspects by a metaphor. Metaphors in the HCI literature have been used in describing certain styles of interfaces, such as the desktop metaphor [Johnson et al. 1989], and as a vehicle for representing and developing designs of interfaces (e.g., Erickson [1990], Madsen [1994]). We use the term differently, in that the metaphors are not in any way intended as interface metaphors, nor are the metaphors imagined to form part of designs. Rather, the aim of the metaphors is to support the evaluator/systems designer in a focused study of how well certain important aspects of human thinking are taken into account in the user interface under inspection. The metaphors are intended to stimulate critical thinking, generate insight, and break fixed conceptions. Such use of metaphors has been thoroughly studied in the literature on creative thinking [Gardner 1982; Kogan 1983] and illustratively applied by Sfard [1998] in the educational domain.

In the next section, we summarize each of the metaphors and give examples of what to consider during evaluation; we also briefly describe the suggested procedure for a MOT evaluation. A full presentation of the various aspects of

MOT is beyond the scope of this paper. A description of the technique that can be used to conduct usability evaluations are presented in Hornbæk and Frøkjær [2002]; Frøkjær and Hornbæk [2002] contains examples of how to understand human-computer interaction design issues through the metaphors and the aspects of thinking that they highlight.

2.1 Metaphor M1: Habit Formation is Like a Landscape Eroded by Water

Habits shape most of human thought activity and behavior (e.g., as physical habits, automaticity, all linguistic activity, and habits of reasoning). This metaphor should 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.

2.2 Metaphor M2: Thinking as a Stream of Thought

Human thinking is experienced as a stream of thought, for example in the continuity of our thinking, and in the richness and wholeness of a person's mental objects, of consciousness, of emotions and subjective life. This metaphor was proposed by William James [1890, vol. I, p. 239] 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, acquaintance objects, 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" [James 1890, vol. I, p. 459].

2.3 Metaphor M3: Awareness as a Jumping Octopus in a Pile of Rags

Here the dynamics of human thinking are considered, that is the awareness shaped through a focus of attention, the fringes of mental objects, association, and reasoning. This metaphor was proposed by Peter Naur [1995, pp. 214–215] 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 vague and shifting connections and feelings, 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, while any center of attention remains in focus for the duration of the *specious present*, that is, about 30 seconds [Naur 2007].

2.4 Metaphor M4: Utterances as Splashes over Water

Here the focus is on the incompleteness of utterances in relation to the thinking underlying them and the ephemeral character of those utterances. This metaphor was proposed by Naur [1995, pp. 214–215] to emphasize how utterances are incomplete expressions of the complexity of a person's current mental object, in the same way as the splashes over the waves tell little about the rolling sea below.

2.5 Metaphor M5: Knowing as a Building Site in Progress

Human knowing is always under construction and incomplete. Also this metaphor was proposed by Naur [1995, pp. 214–215] 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.

2.6 Procedure for a MOT Evaluation

In addition to the five aspects of thinking and their corresponding metaphors, MOT comprises key questions to consider in a usability inspection (see Table II). The procedure of doing a MOT evaluation is similar to that of doing a heuristic evaluation. The evaluator should select a few representative tasks, walk through the interface with those tasks and MOT in mind, and note any usability problems identified. Hornbæk and Frøkjær [2002] suggests a slightly more elaborate procedure, but the basic idea remains as above.

3. EXPERIMENT #1

Experiment #1 compares MOT to heuristic evaluation (HE) by having each participant use one of the techniques to inspect a web application; the problems found were consolidated to a common list; and the key manager/developer of the application assessed the problems. The main goal of experiment #1 is to investigate if MOT is useful at all and whether it can perform as well as heuristic evaluation.

3.1 Participants, Application and Evaluation Techniques

As a compulsory part of a first-year university course in multimedia technology, 87 computer science students used either HE or MOT to evaluate the web application. The web application inspected was a portal for students at the University of Copenhagen to course administration, e-mail, information on grades, university news, etc. (see http://punkt.ku.dk). Participation was anonymous. The participants were free to choose whether their data could be included in the analysis, and were unaware of the authors' special interest in the MOT technique.

Forty-four participants received as a description of MOT a pseudonymized version of Hornbæk and Frøkjær [2002]; forty-three participants received as a description of HE the pages 19–20 and 115–163 from Nielsen [1993]. Each participant individually performed the evaluation supported by scenarios made available by the developers of the web application. Participants received no instruction specific to the techniques. The participants were instructed to write for each usability problem identified (a) a brief title, (b) a detailed description,

Table II. Summary of the MOT-Technique. The Five Metaphors, Their Implications for User Interfaces, and Examples of Questions to be Asked During Usability Inspection

Metaphor of Human Thinking	Implications for User Interfaces	Key Questions/Examples
Habit formation is like a landscape eroded by water.	Support of existing habits and, when necessary, development of new ones.	Are existing habits supported? Can effective new habits be developed? Is the interface predictable?
Thinking as a stream of thought.	Users' thinking should be supported by recognizability, stability and continuity.	Do the system make visible and easily accessible the important task objects and actions? Does the user interface make the system transparent or is attention drawn to non-task related information? Does the system help users to resume interrupted tasks? Is the appearance and content of the system similar to the situation when it was last used?
Awareness as a jumping octopus.	Support users' associations with effective 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? Is the graphical layout and organization helping the user to group tasks?
Utterances as splashes over water.	Support changing and incomplete utterances.	Are alternative ways of expressing the same information available? Are system interpretations of user input made clear? Does the system make a wider interpretation of user input than the user intends or is aware of?
Knowing as a site of buildings.	Users should not have to rely on complete or accurate knowledge—design for incompleteness.	Can the system be used without knowing every detail of it? Do more complex tasks build on the knowledge users have acquired from simpler tasks? Is feedback given to ensure correct interpretations?

(c) an identification of the metaphors or heuristics that helped uncover the problem, and (d) a seriousness rating.

Participants chose seriousness ratings from a commonly used scale [Molich 1994, p. 111]: *Rating 1* is given to a *critical problem* that gives rise to frequent catastrophes which 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; *Rating 2* is given to a *serious problem* that occasionally gives rise to catastrophes which should be corrected in the next version; and *Rating 3* is given to a *cosmetic problem* that should be corrected sometime when an opportunity arises.

3.2 Consolidation of Problems

In order to find problems that were similar to each other, a consolidation of the problems was undertaken. In this consolidation, the two authors grouped together problems perceived to be alike. The consolidation was done over a five-day period, with at least two passes over each problem. The consolidation was done blind to what technique had produced the problems and resulted in a list of 341 consolidated problems. Each consolidated problem consisted of one or more predicted problems.

To test the reliability of the consolidation, an independent rater tried to consolidate a random subset of the problems. The rater received 53 problems together with the list of the consolidated problems from which these 53 problems had been deleted. For each problem, the rater either grouped together that problem with a consolidated problem, or noted that the problem was not similar to any of the consolidated problems. Using Cohen's kappa, the interrater reliability between ratings was $\kappa = .77$, suggesting an excellent agreement beyond chance [Fleiss 1981].

3.3 The Client's Assessment of Problems

In practical usability work, developers and managers often have an important influence on how problems are taken up and possibly addressed. A crucial part of this experiment was therefore to have the consolidated problems assessed by persons who were developing the web application, here called the *client*. In this experiment, the person who managed the development of the web application and was responsible for developing the design represented the client.

For each consolidated problem the client was asked to assess aspects considered likely to influence how to prioritize and revise the design. Two are of relevance here:

Severity of the Problem. The severity of the problem related to users' ability to do their tasks was judged as 1 (very critical problem), 2 (serious problem), 3 (cosmetic problem), or % (not a problem). Note that this grading is different from the participants' seriousness ratings in that only the nature of the problem is being assessed, not when the problem should be corrected, which is contingent upon resources within the development organization. We also included the possibility of assigning something as not being a problem, for instance, if the client felt that the problem missed the goal of the product or conflicted with other essential requirements of the application. The rationale for this metric is that the perceived severity is a likely factor in whether a problem gets solved. In addition, it serves as an indicator of which usability problems are likely to pose actual difficulties to users.

The Perceived Complexity of Solving the Problem. The client also assessed how complex it would be to reach a clear and coherent proposal for how to change the web application in order to remove the problem. The client used a four-point rating scale to judge perceived solution-complexity: (1) very complex solution: will take several weeks to make a new design, possibly involving outside expert assistance; (2) complex solution: a suggestion may be arrived at by experts in the development group in a few weeks; (3) moderately complex solution: new design can be devised in a few days; (4) simple solution: while the actual implementation may take long, a solution to the problem can be found in a few hours. The rationale behind measuring perceived solution-complexity is that this metric captures some aspect of whether a problem is trivial or not. Though problems that can easily be solved are of utility, we argue that problems requiring complex solutions are more important in moving usability evaluation beyond identifying surface-level problems and towards proposing new functionality or task organizations. The latter kind of problems inevitably requires more consideration and redesign effort. Note that this measure considers mainly the complexity of the redesign, not the actual implementation effort required.

See Hornbæk and Frøkjær [2004a] for a description of all aspects of the client's assessment. The assessment was done from a list, which for each consolidated problem showed all the problems that it was consolidated from. The client performed the rating blind to what technique had produced the problems, and he was not aware of what techniques were studied.

3.4 Classification of Usability Problems

In addition to describing the number and overlap of problems found with each technique, we wanted in more detail to describe the differences between the problems found. We did this in two ways. First, we classified usability problems with the User Action Framework, UAF [Andre et al. 2001]. The UAF is a taxonomy of the problems users may experience based on the notion of an interaction cycle, related to Norman's [1986] model of stages of action. The UAF separates in its top categories issues of planning, translation, physical actions, outcome and system functionality, assessment, and problems independent of the interaction cycle. For each of these issues, usability problems may be classified into subcategories that aim at describing more closely the nature of the problem. The version of UAF used consisted of a total of 382 categories organized in two to six levels. Below, we only discuss the broad patterns in the classification, as reflected in the top categories. The purpose of using the UAF is simply to describe what kinds of problems they identify; we are not suggesting that one distribution of problems over the UAF categories would be better than another. However, we are suggesting that knowing what kinds of problems a technique is most likely to help identify is useful to understand that technique, and thereby improve the understanding of its strengths and weaknesses. Thus, we are reporting the relative proportion of problems in a certain UAF category.

Second, we looked at whether problems would typically be experienced by novices only or also by persons with experience in using the system. Nielsen [1993] suggested that this question was important to assess. Operationally, we consider a problem persistent if users experienced with the system (i.e., have used it more than 10 times within a week) would experience the problem more than half of the times when in a situation similar to that where the problem occurred. The rationale in looking at persistence is an assumption that problems that persist for experts are more valuable as an evaluation result, based on the simple assumption that they will affect two user groups rather than just one (and by definition continue to affect those groups). Further, initial data

Table III. The Client's Assessment of Usability Problems Found by Participants in Experiment #1

	HE (43 participants)			MOT (44 participants)		
	M	*SD*	%	*M*	*SD*	%
Number of problems	11.3	6.2	—	9.6	5.7	—
Severity (avg.)***	2.4	0.9	—	2.2	0.7	—
Very Critical (1)	0.8	1.1	7	1.2	1.1	12
Serious (2)	4.8	3.0	42	5.0	3.6	52
Cosmetic (3)	5.6	4.2	49	3.2	2.8	33
Not a problem ()	0.1	0.4	1	0.3	0.5	3
Complexity (avg.)***	3.2	1.0	—	3.00	0.8	—
Very complex (1)	0.1	0.3	1	0.02	0.2	0
Complex (2)	2.7	1.9	24	3.3	2.5	34
Moderate comp. (3)	2.8	2.0	24	2.3	1.9	23
Simple (4)	5.2	3.7	46	3.3	2.6	35
Not graded ()	0.5	0.8	5	0.7	0.9	7

Note: *** = $p < .001$; averages are weighted by the number of problems; HE = heuristic evaluation; MOT = evaluation by metaphors of thinking. Due to rounding errors percentages may not add up.

Table IV. Overlap between Techniques and Participants Regarding Problems Found in Experiment #1

	HE ($N = 43$)			MOT ($N = 44$)		
	M	*SD*	%	*M*	*SD*	%
Number of problems	11.3	6.2	—	9.6	5.7	—
Found by both tech.	6.9	3.6	61	7.2	4.3	74
Found with one tech.						
Many participants	1.3	1.4	11	0.7	1.2	7
One participant*	3.2	3.0	28	1.8	1.8	19

Note: * = p < .05; HE = heuristic evaluation; MOT = evaluation by metaphors of thinking.

suggest that developers find persistent problems more useful in their work compared to novice-only problems [Hornbæk and Frøkjær 2006]. Note, however, that persistence is assessed, not measured from actual behavior.

Both of the classifications above were done blind to which technique had produced a problem and to the client's assessment of that problem. In order to assess the reliability of our classification, we had an independent rater classify a random selection of half of the problems classified in UAF, that is 38 problems. The interrater reliability at the top level was $\kappa = .78$, suggesting an excellent agreement [Fleiss 1981]. The rater also classified whether those problems would persist as users would gain experience with the application; that classification also suggesting an excellent agreement ($\kappa = .89$).

4. RESULTS OF EXPERIMENT #1

Table III summarizes the differences in problems between techniques; Table IV shows the overlap between problems as determined by the consolidation of problems. Because we find an overall difference between techniques (Wilks's lambda = .715, $p < .001$), we below analyze the data from the two tables with individual analyses of variance.

4.1 Number of Problems and Participants' Seriousness Rating

There was no significant difference between the number of problems participants identified with the two techniques, $F(1, 85) = 1.76$, $p > .1$. Between participants, large differences exist in the number of problems uncovered; for example, one participant finds only 2 problems, another finds 28.

Participants' ratings of the seriousness of the problems found differed only marginally between techniques, $F(1, 85) = 2.98$, $p = .09$. Problems found by participants using MOT (Mean, $M = 2.14$; standard deviation, $SD = 1.31$) were reported marginally more serious than were problems found by HE ($M = 2.28$; $SD = 1.05$).

4.2 Client's Assessment

Analyzing the client's assessment of the severity of problems, a significant difference between techniques was found, $F(1, 85) = 15.51$, $p < .001$. The client assessed problems identified with MOT as more severe ($M = 2.21$; $SD = 0.73$) than problems found by HE ($M = 2.42$; $SD = 0.87$). As can be seen from Table III, 49% of the problems identified with HE were assessed cosmetic problems by the client; only 33% of the problems found with MOT were assessed cosmetic. The number of problems that the client did not perceive as usability problems was surprisingly small, between 1% and 3%.

The complexity of the problems identified was significantly different between techniques, $F(1, 85) = 12.94$, $p < .001$. The client assessed problems found with MOT as more complex to solve ($M = 3.00$, $SD = 0.80$) compared to those found by HE ($M = 3.21$, $SD = 0.96$). As shown in Table III, approximately 20% more problems considered "complex" were found with MOT compared to HE; around 60% more problems considered "simple" were found with HE compared to MOT.

4.3 Overlap between Evaluators and Techniques

One use of the consolidation of problems is to describe the overlap between participants using the same technique and the overlap between techniques, see Table IV.

Between techniques, a significant difference was found in the number of problems identified by only one participant, $F(1, 85) = 6.58$, $p < .05$. On the average, participants using HE found 78% more one-participant problems compared to participants using MOT. Incidentally, the one-participant problems found by MOT and those found by HE have comparably low (2.72) average severity (MOT: $SD = 0.80$, HE: $SD = 0.62$). Participants using MOT found problems that were more generally agreed upon among the participants as usability problems. Using a measure from research on the evaluator effect [Hertzum and Jacobsen 2001], the average overlap in problems found by two evaluators using MOT—the so-called any-two agreement measure—was 9.2%, while the any-two agreement for HE was 7.2%.

HE found 74% of the problems found by MOT; MOT found 61% of the problems found by HE. The large number of one-participant problems found by HE resulted in the total number of different problems found being larger for HE

Table V. Classification of Usability Problems Identified in Experiment #1

	HE	MOT
Single evaluator problems (sample of 20)	20	20
UAF->planning[a]	3 (15%)	4 (20%)
UAF->translation	7 (35%)	11 (55%)
UAF->physical action	1 (5%)	0
UAF->outcome & system functionality	2 (10%)	2 (10%)
UAF->assessment	3 (15%)	3 (15%)
UAF->independent	4 (20%)	0
Expert problems[b]	3 (15%)	6 (30%)
Novice problems	17 (85%)	14 (70%)
Multi-evaluator problems	23	13
UAF->planning[a]	2 (9%)	1 (8%)
UAF->translation	10 (43%)	7 (54%)
UAF->physical action	2 (9%)	1 (8%)
UAF->outcome & system functionality	6 (26%)	3 (23%)
UAF->assessment	1 (4%)	1 (8%)
UAF->independent	2 (9%)	0
Expert problems[b]	5 (22%)	5 (38%)
Novice problems	18 (78%)	8 (62%)

Notes: [a]UAF refers to the User Action Framework [Andre et al. 2001], a system for classifying what aspect of interaction a usability problem is associated with.
[b]Novice and expert problems are based on a classification scheme explained in the text.

(249), compared to MOT (181). Thus in some sense, HE resulted in a broader class of problems.

4.4 Differences between Problems Found

Table V shows the results of applying the two classification schemes to a sample of problems unique to either HE or MOT. We chose for the sample (a) all problems found by more than one evaluator and just one technique, and (b) a sample of 20 one-participant problems from each technique. Two observations may be made from the table. It seems that MOT helps evaluators find more problems of the translation type, that is problems about finding out how to do something with the interface, than HE (MOT: 54–55%, HE: 35–43%). Thus, evaluators using MOT would often report problems like (a) “Addresses cannot directly be associated with ones address book. The word ‘address’ in the e-mail menu should be exchanged with ‘address-book’ or the like” and (b) “The form with new e-mail address is difficult to understand. Hard to know what is meant by ‘alias’ and the explanation ‘must be unique’ is not of much help. It is not clear either what this additional information is to be used for.”

The second observation of interest in Table V is that significantly more MOT problems are classified as persistent (30–38%) than HE problems (15–22%), $\chi^2(1, N = 124) = 4.1$, $p < .05$. Thus, MOT seems to uncover problems that are less likely to go away as users gain experience, including problems concerning missing features, inefficient ways of completing tasks, and low quality of the functionality. An example of a problem that was classified as being also significant for expert users is the following:

> Interpretation of e-mail address is not clear. If you for example are not writing out the entire e-mail address, for example just writing “foo,” and mailing it,

> then it is not clear what "foo"-address it is being sent to. If you are giving an incomplete e-mail address it either has to be expanded by the system, so that you can see how the system is interpreting the address, or be rejected, so that the user is forced to write the entire address.

4.5 Summary of Experiment #1

The experiment demonstrated that MOT was useful as an inspection technique for novice evaluators. We were surprised to see how MOT performed at the level of heuristic evaluation, or even better on important measures. The evaluators found an equal number of problems with the two techniques, but problems found with MOT are more serious, more complex to repair, and more likely to persist for expert users. However, understanding MOT as a technique for evaluating interfaces seemed difficult. Instead of trying to improve the description of MOT, we chose to study in more depth how people learn and make use of MOT in evaluation and redesign, leading to experiment #2.

5. EXPERIMENT #2

Experiment #2 compares how 20 participants evaluate and redesign web sites using MOT and cognitive walkthrough (CW). The aim of the experiment was to corroborate the results of experiment #1, this time comparing with the most widely used psychology-based inspection technique. In addition, we wanted to gain some insights into the evaluation process, in particular into the difficulties with understanding MOT that experiment #1 had shown. Consequently, participants were required to keep diaries.

5.1 Participants, Applications and Evaluation Techniques

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. Again, participants were unaware of the authors' involvement in MOT, except one who had an uncertain knowledge about this from a previous course.

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; in the second week, http://www.jcrew.com. Both sites were included in a large professional study of e-commerce sites [Nielsen et al. 2001], which offers insights into usability problems of e-commerce sites. Nielsen et al. [2001] also illustrate some of the differences between the two web sites.

As in experiment #1, MOT was described to participants by a version of Hornbæk and Frøkjær [2002] that had the authors' names replaced by pseudonyms. As a description of cognitive walkthrough (CW) participants received Wharton et al. [1994], widely recognized as the classic presentation of the cognitive walkthrough technique. Though the descriptions varied in length (CW: 36 pages, about 14.000 words; MOT: 23 pages, about 10.000 words) participants used indistinguishable amounts of time reading them, as indicated by their diaries (CW: $M = 194$ min, $SD = 72$; MOT: $M = 233$ min, $SD = 143$; $t(19) = 1.267$, $p > .2$). Note that the participants received no further instruction in the techniques than these documents.

5.2 Procedure for Participants' Inspection

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. Participants spend one week evaluating each web site. Throughout the evaluation and redesign activities participants kept a detailed diary. 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 seriousness rating. The scale for seriousness ratings was the same as in experiment #1.

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. 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.

5.3 Procedure for Redesign of Web Sites

The aim of this redesign activity was to uncover whether qualities of the redesigns varied between evaluation techniques. In addition, we wanted to investigate if participants change their perception of the usability problems when redesigning. We did not attempt to control how participants redesigned the web sites or what resources they used in their redesign activities.

Each redesign of a part of the web site was described as each participant found fit, but should 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.

5.4 Procedure for Diary Writing

The diaries used in the experiment contained a row for every half hour of the day (24 hours). The second field in each row allowed participants to enter a description of the activity they were performing. In addition, the diary had for every half-hour interval fields for 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.

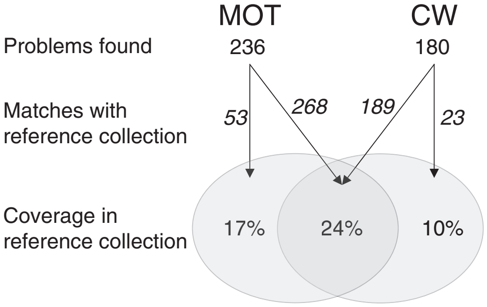
MOT
CW
Problems found
236
180
Matches with reference collection
53
268
189
23
Coverage in reference collection
17%
24%
10%

Fig. 1. Relation between the problems found by participants and the reference collection of usability problems important for designers of e-commerce web sites [Nielsen et al. 2001]. 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.

5.5 Relation of Problems to a Reference Collection

To investigate the quality of the identified problems, we compared them to a reference collection of important usability problems with e-commerce web sites [Nielsen et al. 2001]. This 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 to determine whether it matched or not.

6. RESULTS OF EXPERIMENT #2

Overall, we find a significant difference between techniques (Wilks's lambda = .633, $p < .05$). We therefore proceed to analyze each dependent measure in turn below.

6.1 Number and Nature 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, $F(1, 19) = 3.35$, $p > .05$. On the average participants using MOT assess the severity of the problems as 2.31 ($SD = 0.72$); using CW average severity is 2.25 ($SD = 0.69$).

Figure 1 summarizes the relation between usability problems found by participants and the reference problems described in Nielsen et al. [2001]. 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.

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.

6.2 Subjective Preferences and Comments

In the final comparison of techniques, 15 participants preferred using metaphors of human thinking 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.

6.3 Analysis of Diaries

The analysis of diaries is based on an extraction from the diaries of 224 comments concerning the inspection techniques or redesigns. Here, we only report a few core findings about MOT and CW, the main results of the analysis of diaries are reported in Hornbæk and Frøkjær [2004b].

6.3.1 *Insights and Problems Experienced—MOT*. Several participants make general, positive comments in their diaries about MOT, especially that the key questions and examples help their understanding and use 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." However, 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.

For MOT, we find a higher number of diary 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 in understanding the technique that are hours or days later followed with a comment that the problem is more clear now; that is, 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.

6.3.2 *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. Participants were also unsure 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 about how one sequence should be chosen for doing the walkthrough, wondering “[should] all possible sequences be listed?” and three participants find the notion of correct action, used in the criteria for evaluating, hard to understand.

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 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.

6.4 Differences in Problems Found

Using the procedure from experiment #1, we classified a random sample of half of the problems identified (207 problems). An independent rater classified a random selection of 25% of those problems (52 problems). The interrater reliability at the top level of UAF was $\kappa = .74$. The interrater reliability for the classification of persistence was $\kappa = .79$. These kappas indicate good to excellent agreement [Fleiss 1981].

Table VI summarizes the classification of usability problems. Two trends are clear. First, the UAF classification differs between techniques. In particular, MOT identifies more problems classified as concerning functionality and fewer problems classified as concerning translation than CW. Second, MOT identifies more problems classified as persistent for expert users (38%) than CW (13%), $\chi^2(1, N = 207) = 16.48$, $p < .001$. The focus of cognitive walkthrough on exploratory learning of an interface is perhaps the reason for the latter observation.

Table VI. Classification of Usability Problems Found in Experiment #2

	CW	MOT
Number of problems	92	115[c]
UAF->planning[a]	12 (13%)	17 (15%)
UAF->translation	50 (54%)	45 (39%)
UAF->physical action	6 (7%)	9 (8%)
UAF->outcome & system functionality	9 (10%)	22 (19%)
UAF->assessment	13 (14%)	11 (5%)
UAF->independent	2 (2%)	11 (10%)
Expert problems[b]	12 (13%)	44 (38%)
Novice problems	80 (87%)	71 (62%)

Notes: [a]UAF is referring to the User Action Framework [Andre et al. 2001].
[b]Novice and expert problems are based on a classification scheme explained in the text.
[c]One problem found with MOT was too unclear to allow classification.

6.5 Analysis of Redesign Proposals

To further analyze the differences between techniques, we assessed the redesign proposals that participants handed in. A total of 113 redesign proposals were analyzed: each subject handed in up to three redesign proposals for each of the two evaluation techniques. The assessment was based on four criteria: (a) what the redesign proposal aimed to correct, (b) why the redesign was important, (c) the details of the redesign proposal, and (d) an overall assessment of the proposal, emphasizing its importance for usability, clarity, and coherence. Each of these criteria was assessed by the first author of this paper on a scale from 1 (poor) to 5 (very good or outstanding); redesign proposals were assessed in a random order and blind to which technique/participant had produced them.

An analysis of the overall assessment showed no difference between techniques in the assessments, $F(1,19) = 1.43$, $p > .25$. Ten participants had their redesigns assessed higher with MOT, 8 participants had their redesigns assessed higher with CW, and 2 participants achieved similar assessments. We therefore discontinued the analysis of the redesign proposals and simply note that even though the analysis of proposals was conducted within subjects, large variations between redesign proposals still exist. In part, this might be a result of our choice not to control redesign procedure or resources. Consequently, any effect associated to evaluation techniques would be difficult to detect.

The diaries shed slightly more light on the participants' redesign activities. Counting the diary entries shows that participants used comparable amounts of time on creating the redesigns (CW: $M = 6.6$ hours, $SD = 3.3$; MOT: $M = 6.2$, $SD = 3.1$; $t(19) = 0.46$, $p > .5$). So the inability to find differences in the assessment of redesign results is not related to some technique-specific difference in how long participants used to redesign.

The comments in the diaries during redesign activities fall in two major groups. One of these concerns problems that participants change their mind about. One participant, for example, wrote 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. This change in opinion may systematically cause some problems to be ignored, problems that for example could be among the more complex type. Another group of comments suggests that at least five participants identify problems during redesign that they were previously unaware of Should they have chosen to redesign such problems, problems that formed the basis of the redesigns could have been of another nature from those found during evaluation. Such changes during the redesign work are possible reasons why the quantitative differences regarding usability problems do not show themselves in the analysis of the redesigns.

6.6 Summary of Experiment #2

The experiment showed that MOT performed significantly better on a number of important measures than cognitive walkthrough in evaluating two e-commerce web sites. The evaluators found more problems with MOT and these problems had a wider coverage of a reference collection describing important and typical problems with e-commerce web sites. As found in experiment #1, the evaluators had a harder time reading and understanding MOT compared to cognitive walkthrough. But detailed diaries covering both the initial reading and learning activities as well as the actual evaluations show how most difficulties are eventually resolved by the evaluators themselves. Despite these difficulties a clear majority of the evaluators ends up preferring MOT to cognitive walkthrough. Further, the experiment shed some light over differences in the kind of problems identified with the two techniques. Differences concerning especially the translation and the system functionality category were large, which indicate that MOT and cognitive walkthrough might be good supplementing techniques. Finally, MOT seems to find more problems likely to persist for expert users while cognitive walkthrough finds more novice user problems, a possible effect of cognitive walkthrough being a technique related to exploratory learning and problem solving. Having again found good performance relative to a frequently used inspection technique, we wanted to challenge MOT by comparing it to the "gold standard" of usability testing, think aloud testing, leading to experiment 3.

7. EXPERIMENT #3

Experiment #3 compares how participants evaluate two non-traditional user interfaces using think aloud user testing, cognitive walkthrough, and metaphors of human thinking. Participants used two of three evaluation techniques and evaluated both interfaces. The aim of the experiment was to investigate our conjecture that MOT would be effective for evaluating nontraditional user interfaces. In addition, we wanted to compare the performance of MOT to an empirical evaluation technique, in this case think aloud user testing.

7.1 Participants, Applications and Evaluation Techniques

As part of a course on systems design and human-computer interaction, 58 participants evaluated non-traditional user interfaces. The participants had a mean age of 24 years, and had studied computer science for, on the average, 2.8 years. Approximately one half of the participants had previously taken courses on user interfaces; one-quarter had used the think aloud technique, one-sixth had used metaphors of human thinking, and one tenth had used cognitive walkthrough. With respect to the interfaces being evaluated, approximately three-quarters of the students owned a mobile phone; among those three-quarters owned the particular brand being evaluated.

For the experiment, we chose two nontraditional interfaces. The natural language interface (NLI) gave access to information on holiday allowances through a spoken-language-only telephone dialog [Dybkjær and Dybkjær 2002]. The mobile phone interface (MP) comprised the phonebook application of the mobile phone Nokia 7210, see http://www.nokia.com/phones/7210.

The description of think aloud (TA) user testing given to participants was Molich [2003]. CW and MOT were described as in experiment #2. Participants received no additional instruction in the techniques.

7.2 Procedure

In the first week of the experiment, participants evaluated the natural language interface using one of the three evaluation techniques determined at random. In addition to the description of the evaluation technique, participants received scenarios describing imagined use of the interface and a description of the user group they should consider when evaluating. This evaluation resulted in an *individual list* of usability problems. That list used the same format as in experiments #1 and #2 and the same scale for rating seriousness (see Section 3.3). A total of 55 individual lists for the NLI interface were generated.

After having completed the individual lists, participants who had used the same technique were asked to work together in groups of two or three persons to produce a *group list* of problems. In producing this list, participants had to group similar problems. For each problem on the group list, participants were asked to give a title, a description of the problem, a list of individual problems that the group problem was based on, and an explanation of why those individual problems were merged to a group problem. Also, the group list specified problems that for some reason the group did not—when grouping problems—consider usability problems. A total of 20 group lists for the NLI interface was generated.

Week two of the study followed the procedure for week one, except that the mobile phone was evaluated and that participants used a new evaluation technique determined at random between the two techniques that any participant had not yet used. This resulted in 55 individual lists and 20 group lists being generated for the mobile phone.

One week after participants had finished the evaluation, the participants met to create two *goal lists*; one for the NLI interface and one for the MP interface. This list was made using a bottom-up procedure, in which a representative of each group placed one problem at a time. If the problem were similar to an already placed problem, as judged by the participants, they were merged. The procedure was based on that by Beyer and Holzblatt [1998] for creating affinity diagrams.

Finally, clients, that is, one person from each of the two organizations responsible for developing the interfaces being evaluated, judged the problems

Table VII. Characteristics of Usability Problems from Experiment #3

	TA		CW		MOT	
	M	*SD*	*M*	*SD*	*M*	*SD*
NLI						
No. problems*	5.70	2.45	4.71	1.86	6.00	3.12
Seriousness*	2.11	0.68	2.29	0.66	1.95	0.66
Severity	2.93	0.61	3.14	0.67	2.99	0.63
Complexity*	1.12	0.51	1.34	0.83	1.09	0.42
Clarity	1.41	0.78	1.38	0.76	1.45	0.80
Not a problem*	14%	35	29%	46	18%	39
MP						
No. problems*	6.78	2.13	6.47	3.30	7.15	4.92
Seriousness*	2.06	0.71	2.24	0.77	2.30	0.69
Severity	2.48	0.66	2.55	0.62	2.51	0.58
Complexity	3.03	0.58	3.01	0.56	3.02	0.61
Clarity	1.99	0.39	2.05	0.46	2.08	0.49
Not a problem	3%	16	4%	20	2%	13

Notes: Seriousness (1 = critical, 2 = serious, 3 = cosmetic), severity (1 = very critical, 2 = serious, 3 = cosmetic), complexity (1 = very complex, 2 = complex, 3 = moderate complexity, 4 = simple) and clarity (1 = very clear, 2 = clear, 3 = unsure) are weighted by number of problems. Each cell contains data from between 17 and 20 participants. * = $p < .05$.

found. Clients received a randomly ordered list of all problems from the individual lists concerning their interface, and the problems on group lists concerning their interface that were based on more than one individual problem. We asked the clients to judge severity and complexity of the problems using the rating procedure of experiment #1. In addition we asked the client to grade the *clarity of the problem*, that is whether the description of the problem was clear and understandable. We used a three-point scale to judge clarity: (1) very clear description which gives the reader a sure and non-ambiguous understanding of what the author of the problems intends to point out as a problem; (2) clear description, which gives the reader a sure or relatively sure understanding of what is intended; (3) unclear description, which leaves the reader with an unsure understanding of what is intended.

Below, we analyze only the individual lists and the clients' grading of problems on these; the analysis of performance at the level of the group and goal lists will be reported elsewhere. As the experiment used an incomplete block design (meaning that participants did not use all three evaluation techniques, but rather only two of the techniques) we used intrablock analysis of the data to compare participants' performance [John 1971, 219 ff]. Data from the clients' judgment are not comparable between the two clients, so we analyzed each interface in turn with analyses of variance. As three participants failed to complete, both evaluations and finding at least one problem with each interface, we removed the data from those participants before analyzing the experiment.

8. RESULTS OF EXPERIMENT #3

Table VII summarizes the participants' performance.

8.1 Number of Problems and Participants' Seriousness Rating

We find significant differences between the numbers of problems identified. Linear contrasts show that MOT identified more problems than CW, F(1, 52) = 5.15, p < .05, and than TA, F(1, 52) = 7.72, p < .05. These effects, however, are small: 19 participants found the most problems with MOT; 15 found the most problems with TA; and 14 found the most problems with CW. Seven participants found the same number of problems with the techniques they used.

The average of the seriousness ratings assigned by participants to problems differs between techniques. For the NLI interface, participants perceived the problems they identify as more serious with MOT ($M = 1.95$, $SD = 0.66$) compared to the CW technique ($M = 2.29$, $SD = 0.66$), $F(1, 52) = 5.12$, $p < .05$. As an illustration, 11% of the problems found with CW were categorized by participants as very critical problems (grade 1); the corresponding percentage for MOT was 24%. For the MP interface, however, participants considered problems found with MOT ($M = 2.30$, $SD = 0.69$) less serious than problems found with TA ($M = 2.06$, $SD = 0.71$), $F(1, 52) = 5.41$, $p < .05$. Here the percentages of very critical problems were TA: 22% and MOT: 13%.

8.2 Clients' Grading

First, we consider the case of the NLI interface. We find no difference in average severity between techniques, $F(2, 214) = 0.14$, $p > .05$. However, a significant difference between techniques exists with respect to the proportion of participants' problems that the client did not consider to be usability problems, $F(2, 52) = 3.34$, $p < .05$. CW seems more likely than the other techniques to identify problems that the client did not consider a problem. Of the problems that were identified with CW and assessed on severity, 29% were *not* considered problems by the client (compare to MOT 18% and TA 14%).

The client's grading of the complexity differs between techniques, $F(2, 239) = 3.92$, $p < .05$. Linear contrasts show that the client assessed problems found with CW less complex ($M = 1.34$, $SD = 0.62$) than problems found with either TA ($M = 1.12$, $SD = 0.26$; $F(1, 239) = 6.85$, $p < .05$) or MOT ($M = 1.09$, $SD = 0.20$; $F(1, 239) = 5.57$, $p < .05$). Supporting this finding is the larger proportion of problems found by CW considered to be of moderately or simple complexity (grade 3 or 4; 10%), compared to the proportions for the other techniques (TA 3%; MOT 2%).

No differences in clarity between techniques could be found, $F(2, 282) = 0.16$, $p > .5$.

With the MP interface, there appear to be few differences between techniques in the client's grading. None of the dependent measures derived from the client's grading were significant, and we consequently discontinued our analysis of this portion of the data.

8.3 Participants' Preference and Comments

Thirty-six participants used TA (21 in combination with MOT and 15 in combination with CW). Eighteen participants preferred TA over MOT and 3 participants preferred MOT over TA; 9 participants preferred TA over CW and 6 participants preferred CW over TA. Among the 19 participants who used

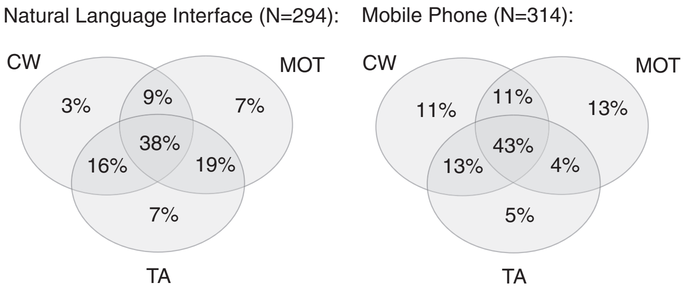

Fig. 2. Distribution of usability problems found in Experiment #3.

MOT and CW in combination, 11 preferred MOT and 6 preferred CW, supporting the finding of experiment #2.

Two reasons for preferring TA stand out. At least six participants commented that the result of think aloud tests are easier to interpret. A similar number of participants remarked that they prefer TA because they may identify problems they could not have found on their own.

As in experiment #2, some participants remarked that MOT is challenging to understand. One participant, for example, commented that "Some of the metaphors were a bit too abstract and perhaps did not fit reality." CW, conversely, is seen as easy and intuitive to understand. However, MOT was seen by six participants as requiring less knowledge than CW about the system under evaluation. MOT was also considered to give the evaluator more freedom in identifying problems; CW, on the other hand, was seen by eight participants as finding only a limited kind of problems, for example "the technique [CW] has a very limited scope."

Note also the comments from a couple of participants that it was hard to adapt the evaluation techniques to a nontraditional user interfaces. Especially, TA was hard to use in the context of the natural language interface.

8.4 Difference in Problems between Techniques

From the participants' construction of a goal list using affinity diagramming, we can extract group problems that are unique to a particular evaluation technique. Figure 2 shows the distribution between the three techniques of problems from evaluators' individual lists, based on the overlap between group problems at the goal list. This figure shows two important results. First, TA identified relatively few problems that the participants judged as not identified by other techniques (from 5% to 7% of the total number of problems on the goal list). This number is surprisingly low, if one expected TA to help identify a distinct set of problems that were different from the problems that usability inspectors might find. In addition, some of these problems were not accepted by the clients as usability problems. Using TA, 14% of the problems found with the NLI and 3% found the MP interface were not considered problems by the clients.

Second, the overlap between techniques differs between the two interfaces evaluated. For the NLI interface, MOT and CW has a relatively small degree of overlap, while for the MP interface, it appears to be the overlap between MOT and TA that are low. We do not have any explanation for these differences.

Table VIII. Classification of Usability Problems Found in Experiment #3

	TA	CW	MOT
NLI problems	21	10	21
UAF->planning [a]	3 (14%)	1 (10%)	2 (10%)
UAF->translation	7 (33%)	4 (40%)	8 (38%)
UAF->physical action	0	0	0
UAF->outcome & system functionality	9 (43%)	5 (50%)	7 (52%)
UAF->assessment	1 (5%)	0	0
UAF->independent	1 (5%)	0	0
Expert problems [b]	10 (48%)	2 (20%)	9 (43%)
Novice problems	11 (52%)	8 (80%)	12 (57%)
MP problems	17	33[c]	42
UAF->planning	4 (24%)	4 (12%)	2 (5%)
UAF->translation [a]	8 (47%)	23 (70%)	26 (62%)
UAF->physical action	0	2 (6%)	2 (5%)
UAF->outcome & system functionality	2 (12%)	1 (3%)	7 (17%)
UAF->assessment	1 (6%)	3 (9%)	5 (12%)
UAF->independent	2 (12%)	0	0
Expert problems[b]	5 (30%)	7 (21%)	13 (31%)
Novice problems	12 (71%)	26 (79%)	29 (69%)

Notes: [a] UAF is referring to the User Action Framework [Andre et al. 2001].
[b] Novice and expert problems are based on a classification scheme explained in the text.
[c] One problem found with CW was too unclear to allow classification.

To investigate the differences in problems identified, we tried to characterize the kinds of problems typical of the three techniques. Using the procedure from experiment #1, we categorized problems in the User Action Framework; the interrater reliability of classifying 50% of the problems was $\kappa = .72$ for both the UAF and persistence classifications, showing good agreement. Table VIII shows that the distribution of problems across the top-level categories of the UAF is relatively similar. The main point of difference between techniques is that TA identified fewer problems (33 and 47%) in the translation category than MOT (38 and 62%) and CW (40 and 70%). One possible interpretation of this finding is that translation problems are easy to find with the analytical techniques; in our data a common kind of problem found with these techniques is of the form "It is not clear how" followed by some aspect of the interface where users have difficulty in identifying the user interface object related to the task they are trying to accomplish.

As in experiment #1, we also classified problems according to whether they persist as users gain experience with the system. Table VIII shows the classification, which differs significantly between evaluation techniques, $\chi^2(1, N = 184) = 6.53$, $p < .05$. Compared to CW, it seems that MOT identified more expert problems (MOT: 43 and 31%; CW: 20 and 21%). This result supports the findings from experiment #2. Compared to TA, however, MOT identified a similar amount of expert problems (TA: 48 and 30%).

A particular point of this experiment was to compare MOT to an empirical technique, TA. One interesting observation is that with the analytical techniques CW and MOT, evaluators seemed to identify problems that they run into when using the system, what we call *evaluator-as-user* problems, rather than problems they found when using the evaluation techniques. Thus, the problem lists contain problems that are not found as a result of evaluating the interface, but rather as a result of using it. This observation complements those from experiment #2. Another difference seems to be that with TA more concrete formulations of the problems were given. Example problems are (1) "Gets a nonusable answer and quits. The user chooses a 'wrong' category. The system asks if the user have any more questions; the user says no; and the system says goodbye. The system should have asked if the user got the information needed." and "Doesn't understand word. The user has to say the word 'e-mail' several times (louder and louder) before it is correctly interpreted. [· · ·]". Sometimes problems found with MOT (and CW) can be somewhat abstract, for example "Quantity. Too many options in menus and submenus" or "The order of items in [the menu] Other Possibilities. It doesn't seem like there is any ordering in how different topics under Other Possibilities are presented". Note that these differences did not make the client rate the clarity of MOT (or CW) problems lower. Therefore, it seems that both abstract and concrete statements of problems have utility to the client.

8.5 Summary of Experiment #3

All three techniques showed to be applicable in this context of evaluating nontraditional interfaces, but none came out as the overall most effective. A clear majority of the evaluators prefer using think aloud testing to any of the two inspection techniques; MOT was preferred to cognitive walkthrough, a result coherent with experiment #2. More problems are identified with MOT than with think aloud testing and with cognitive walkthrough, but the size of these effects is small. Think aloud identifies relatively few problems that the participants judge as not identified by other techniques; but the overlap between techniques differs between the two interfaces, a result that we can not explain. In the UAF classification, the main point of difference found was that think aloud identifies fewer problems in the translation category than MOT and cognitive walkthrough.

9. DISCUSSION

The experiments generate optimism regarding the conjecture on the effectiveness of MOT presented in the Introduction; Table I summarizes the results across experiments. As for the more specific research questions, evaluators using MOT performed well in comparison with those using HE, by predicting problems that are more severe, more complex to repair, and more likely to persist. In comparison with CW, MOT supports evaluators in identifying more problems compared to CW, while constraining the scope of evaluation less than CW. The process of inspection with MOT studied in experiment #2 suggests that understanding MOT is challenging, even though participants often are able to revise their misunderstandings or dispel their doubts. The research questions on whether MOT is effective for evaluating nontraditional interaction styles has no clear-cut answer. All three techniques used in experiment #3 appeared to be hard for participants to use efficiently with the mobile phone and the natural language interface. In particular, users of CW and MOT remarked that the examples of how to apply the techniques given in the descriptions of the techniques may not be pertinent to the non-traditional applications. Finally, compared to TA, MOT performs comparably with respect to seriousness, perceived complexity, and persistence. Though MOT finds slightly more problems, TA is preferred by evaluators.

9.1 Relative Merits of the Techniques

A few comments pertaining to each of the techniques that were compared to MOT are relevant. The main difference between MOT and heuristic evaluation is that the latter strives for simplicity. From the inception of the technique, Nielsen and Molich have tried to simplify the formulation of the heuristics. From a certain point of view, we are doing the opposite in MOT by trying to capture human thinking by simplified but still rich and association-creating metaphors. In experiment #1 similar numbers of problems are found, but the difference in intent of the technique may be behind the quite different kinds of problems found, for example in terms of severity, complexity, and persistence. An experiment by Fu and Salvendy [2002] has suggested that “heuristic evaluation is more effective in predicting usability problems that more advanced users will experience” (p. 142). Conversely, “user testing is more effective in discovering usability problems that novice users encounter” (p. 141). However, our data indicate the opposite pattern: in experiment #1 MOT finds more usability problems that would persist with practice than did HE; in experiment #3, MOT and TA find similar numbers of persistent problems. More research is needed to interpret these seemingly dissimilar results.

Cognitive walkthrough provides detailed and according to some participants overly strict procedures for walking through tasks; in MOT, there are less such support. Cockton et al. [2003] suggested that methods like cognitive walkthrough, which they call procedural inspection techniques, are “better placed to consider interaction beyond brief encounters with individual system features” (p. 1121). We have no evidence that the problems predicted with the use of CW differ from those predicted by MOT in this way. Rather, our experiments suggest that the procedure of CW restricts the scope of the evaluation and that problems predicted with MOT are likely to be more persistent.

Regarding think aloud testing, experiment #3 suggests that problems identified with TA to a large extent overlap the problems identified with the other evaluation techniques. In addition, the problems unique to TA are not assessed as particularly serious by either client. Rather, substantial percentages are not considered usability problems at all. These findings are in contrast to common expectations in the literature that think aloud testing provides a “gold standard” and that TA usually perform very well compared to other evaluation techniques (e.g., Cockton et al. [2003]). However, the findings are supported by the study by Hornbæk and Frøkjær [2005]. They found that the utility of problems identified with TA and with MOT were not assessed differently by developers in terms of the problems’ utility as input to the development process. One explanation is that differences between techniques stems from the experience of using TA. Preference data from experiment #3 appear to support this explanation. In that experiment, participants strongly preferred using TA over MOT. They also commented on the inspiration of seeing someone else interact with the system. Indeed, the experience of experiencing TA has long been claimed a hallmark of that method [Helms Jørgensen 1990]. An alternative explanation is that such differences stem from the description of the problems, perhaps because problems produced with the aid of TA are clearer and always cast in terms of the users' problems. In the study by Hornbæk and Frøkjær [2005], at least one developer suggested that exactly the concrete descriptions of user actions were valued; however in a follow-up analysis of the problems from that study, no impact of mentioning of concrete user difficulties in problem reports affected developers' ratings of the utility [Hornbæk and Frøkjær 2006]. A final explanation might be that user testing finds more problems not classified as translation problems in the User Action Framework, covering the interaction circle more completely than MOT or HE.

9.2 Validity of the Experiments

In designing the experiments reported in this article, we have tried to meet the concerns of a number of authors in usability research about valid experimental comparisons of evaluation techniques. In particular, we have paid attention to the five validity issues listed in Gray and Salzman's review [1998]. Their issues concern statistical conclusion validity, internal validity, construct validity, external validity, and conclusion validity. Below we discuss each of these in relation to the experiments presented in this paper.

The main threat to statistical conclusion validity is random heterogeneity of participants. All three experiments have 17 or more evaluators per technique, which reduce the influence of participants' heterogeneity, and experiments #2 and #3 are within-subjects experiments, allowing subjects to serve as their own controls. Another issue is doing too many statistical tests, thereby inflating the overall likelihood of reaching significant results. Experiments #1 and #2 use one overall significance test for the main dependent measures, thereby protecting the experiment-wide probability of finding significant results where none exist; in Experiment #3, due to the complex nature of the experimental design, this was not possible. Finally, many of the significant tests identify effects that are only small in size. This is the case with respect to counts of problems in experiment #3, for example. Other effects, such as the difference in counts of problems in experiment #2 (with a size of $\text{eta}^2 = .314$), are medium according to Cohen [1992]. In practical reality—it could be argued—these small to medium effects do not matter much. It is true that many factors would probably impact evaluation more (evaluator training, evaluation in multidisciplinary teams, better techniques for prioritizing problems). However, we believe that these effects are substantial enough to warrant attention. As examples, MOT finds 30% more problems than CW (experiment #2) and would help a typical team of three evaluators find one or two serious/critical problems more than heuristic evaluation (experiment #1).

Internal validity concerns an experiment's ability to establish causal relationships between independent and dependent variables. We tried to avoid the difficulties listed by Gray and Salzman [1998] by using similar classification and reporting schemes for all evaluation techniques and by using random assignments of participants to evaluation techniques. We did not, however, enforce strict time limits on evaluation. In experiment #2 we gave participants an expected duration for the evaluation and in experiment #3 participants used comparable time on their evaluations. In experiment #1, however, participants used more time on heuristic evaluation compared to MOT (see Hornbæk and Frøkjær [2004a]). We cannot say how this may have affected the evaluation results, but it seems unlikely that longer evaluations would result in poorer evaluations.

Construct validity concerns whether the experimenters are "manipulating what they claim to be manipulating (the causal construct) and [· · ·] measuring what they claim to be measuring (the effect construct)" [Gray and Salzman 1998, p. 213]. Regarding the causal construct, defining the evaluation techniques used are crucial. We used authoritative descriptions of cognitive walkthrough and think aloud testing. However, experiment #2 did not support evaluators with task descriptions, as did the other two experiments. This might have affected cognitive walkthrough adversely, because task descriptions are particularly crucial for the performance with that technique (e.g., Sears and Hess [1998]). Confounding of treatments is another causal construct concern (e.g., by subjects applying previously used evaluation techniques in addition to the one they are meant to use). By using random assignment and similar instructions for participants we have largely avoided this issue. The diaries from experiment #2, however, suggests that participants in the second week of that experiment cannot help also consider the evaluation technique used in the first week. While this effect should be symmetric across techniques, it suggests that the increased control over participant performance in a within-subjects design comes at a price.

Validity in relation to effect constructs is more difficult to assess. Some of the dependent measures of the experiments (ratings of severity and perceived solution-complexity) have lead to a greater reliance upon the judgments of the developers. We consider this a consequence of moving closer towards understanding how the results of usability evaluations are taken up in design, and, effectively, of downplaying the role of isolated use of for instance problem counts. However, these measures are affected by a developer effect [Law 2006], which we have not attempted to protect against in our experiments. Only a detailed content analysis can reveal if developers react to particular ways of describing problems. While we believe that severity and complexity are indicators of developers' appreciation of usability problems, it must be noted that these are opinions, not behavioral measures. Similarly, while the persistence classification has good inter-rater reliability, it only represents an expert assessment of likely persistence, not an actual behavioral measure. Understanding the validity of these measures in relation to actual behavior would be valuable. Finally, evaluators' preference and comments were widely treated in this article. Even though these measures are probably as biased as traditional subjective satisfaction measures from usability research, they appear useful as one indicator of how evaluators actually go about evaluating.

Another validity concern concerns the possibility of generalizing the findings across settings and persons. The experiments reported here utilize only novice evaluators, that is, computer science students. We consider this is an important target group for making the best of usability in industrial systems design and development, and a group that will give us much useful insight into differences and similarities between evaluation techniques. Yet, the move towards figuring out what is relevant evaluation results in design work necessarily suggests emphasizing expert evaluators in future studies.

The final validity concern from Gray and Salzman [1998]—conclusion validity—can only be ensured by careful writing; this we will not discuss.

9.3 Future Work

The experiments have identified several possible improvements to the current description of MOT. As mentioned above, illustrations of the technique should be more diverse with reference to different interaction styles and use contexts. With such illustrations, usability inspectors may be better supported in learning and applying the technique in new fields. Second, while all of the metaphors may certainly be extended, we feel in particular that aspects of metaphor 2—thinking as a stream of thought—could be formulated in a sixth metaphor. The aspects we are thinking of concern the observation that 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" [James 1890, vol. I, p. 459]. Currently, we consider this aspect crucial to doing a solid evaluation with MOT. Such emphasis on anchor points in the stream of thought is related to Shneiderman's Object-Actions Interface model [see Shneiderman and Plaisant 2005, p. 95 ff] and to the goal of identifying objects central in the users' tasks in object-oriented modeling. Similarly, the tight correspondence between concepts in the user's understanding of the application domain and the concepts pertained in the user interface is the main target of another evaluation technique called Concept-based Analysis for Surface and Structural Misfits, CASSM [Connell et al. 2004].

Another area calling for a more specific treatment in MOT is the role of "feelings" and "emotions" in human-computer interaction. Feelings and emotions are integrated in the James-Naur descriptions of human thinking, so we do not need a new theory to describe these aspects. As briefly mentioned in Section 2.3, any mental object embraces feelings, and feelings and emotions are highly important in shaping habits, stream of thought, acquaintance objects, and utterances of any individual.

None of the three experiments has in a deep manner investigated the impact of the metaphors on the inspection process, that is, whether MOT actually affects the inspection process in terms of stimulating thinking and breaking fixed conceptions, as argued earlier in this article. We are eager to investigate further the specific contribution of the metaphorical description form and its influence on evaluator's thinking.

Finally, the attempt to investigate whether redesigns reflect differences in evaluation techniques was unsuccessful. However, we think that looking at redesigns are a natural consequence of focusing on downstream utility [John and

Marks 1997], and we thus suggest that future work might attempt to control redesign procedure and redesign resources so as to come closer to any existing effects of evaluation techniques.

10. CONCLUSION

We have described the usability evaluation technique metaphors of human thinking (MOT). The goal of the technique is to provide effective usability inspection by focusing on users' thinking and thus being applicably across different devices and contexts of use. In a series of experiments, we have compared MOT to three common usability evaluation techniques: heuristic evaluation, cognitive walkthrough, and think aloud testing. The experiments show that MOT performs better on important measures than the inspection techniques cognitive walkthrough and heuristic evaluation. In two experiments MOT finds more problems, and the problems found with MOT are typically more complex and more likely to persist for expert users. MOT performs comparably to think aloud testing in terms of the number of problems found; however, think aloud testing appears to find usability problems of more diverse types. Evaluators prefer think aloud testing to the inspection techniques, but MOT was preferred over cognitive walkthrough and heuristic evaluation. Overall, our data suggest that evaluation by metaphors of human thinking is an effective evaluation technique.

ACKNOWLEDGMENTS

We wish to thank Peter Naur for his continuing support in our work. Hans Dybkjær at Prolog Development Center, Claus Damgaard at Copenhagen University, and Niels Jacobsen at Nokia generously participated in a demanding and very time consuming activity of assessing problems, and additionally provided invaluable feedback. We are indebted to the students participating in the experiments who made this work possible. Rex Hartson and Jon Howarth were helpful in all matters relating to the User Action Framework. Finally, the anonymous reviewers generously offered their insights and time in helping us clarify the article.

REFERENCES

ANDRE, T. S., HARTSON, H. R., BELZ, S. M., AND MCCREARY, F. A. 2001. The user action framework: A reliable foundation for usability engineering support tools. *Int. J. Human-Comput. Stud. 54*, 1, 107–136.

BASTIEN, C. AND SCAPIN, D. 1995. Evaluating a user interface with ergonomic criteria. *Int. J. Human-Comput. Stud. 7*, 2, 105–121.

BEYER, H. AND HOLTZBLATT, K. 1998. *Contextual Design*. Morgan-Kaufman, San Francisco, CA.

COCKTON, G., LAVERY, D., AND WOOLRYCH, A. 2003. Inspection-based evaluations. In *The Human-Computer Interaction Handbook*. J. A. Jacko, and A. Sears, Eds. Lawrence Erlbaum Associates, Mahwah, NJ, 1118–1138.

COHEN, J. 1992. A power primer. *Psych. Bull. 112*, 1, 155–159.

CONNELL, I., BLANDFORD, A., AND GREEN, T. 2004. CASSM and cognitive walkthrough: Usability issues with ticket vending machines. *Behav. Inf. Tech. 23*, 5, 307–320.

DYBKJÆR, H. AND DYBKJÆR, L. 2002. Experiences from a Danish spoken dialogue system. In *2nd Danish Human-Computer Interaction Research Symposium*, E. Frøkjær and K. Hornbæk Eds. DIKU tech. rep. 02/19, 15–19.

Erickson, T. D. 1990. Working with interface metaphors. In *The Art of Human Computer Interface Design*. B. Laurel, Eds. Addison-Wesley, Reading, MA, 65–73.

Fleiss, J. L. 1981. *Statistical Methods for Rates and Proportions*. Wiley, New York.

Frøkjær, E. and Hornbæk, K. 2002. Metaphors of human thinking in HCI: Habit, stream of thought, awareness, utterance, and knowing. In *Proceedings of HF2002 / OzCHI 2002* (Melbourne, Australia, Nov. 25–27).

Fu, L. and Salvendy, G. 2002. Effectiveness of user-testing and heuristic evaluation as a function of performance classification. *Behav. Inf. Tech. 21*, 2, 137–143.

Gardner, H. 1982. *Art, Mind and Brain: A Cognitive Approach to Creativity*. Basic Books, New York.

Gray, W. D. and Salzman, M. C. 1998. Damaged merchandise? A review of experiments that compare usability evaluation methods. *Human-Comput. Interact. 13*, 3, 203–261.

Green, T. R. G. and Petre, M. 1996. Usability analysis of visual programming environments: a 'cognitive dimensions' framework. *J. Vis. Lang. Comput. 7*, 131–174.

Hartson, H. R., Andre, T. S., and Williges, R. C. 2001. Criteria for evaluating usability evaluation methods. *Int. J. Human-Comput. Interact. 13*, 4, 373–410.

Helms Jørgensen, A. 1990. Thinking-aloud in user interface design: A method promoting cognitive ergonomics. *Ergonomics 33*, 4, 501–507.

Hertzum, M. and Jacobsen, N. E. 2001. The evaluator effect: A chilling fact about usability evaluation methods. *Internat. J. Human-Comput. Interact. 13*, 421–443.

Hornbæk, K. and Frøkjær, E. 2002. Evaluating user interfaces with metaphors of human thinking. In *Proceedings of 7th ERCIM Workshop "User Interfaces for All,"* (ERCIM Workshop on User Interfaces for All), Chantilly, France, Oct. 24-25). N. Carbonell and C. Stephanidis, Eds. Lecture Notes in Computer Science, vol. 2615, Springer-Verlag, Berlin, Germany, 486–507.

Hornbæk, K. and Frøkjær, E. 2004a. Usability inspection by metaphors of human thinking compared to heuristic evaluation. *Int. J. Human-Comput. Interact. 17*, 3, 357–374.

Hornbæk, K. and Frøkjær, E. 2004b. Two psychology-based usability inspection techniques studied in a diary experiment. In *Proceedings of the Nordic Conference on Human-Computer Interaction* (Nordichi 2004), 2–12.

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

Hornbæk, K. and Frøkjær, E. 2006. What kind of usability-problem description are useful for developers? In *Proceedings of the Annual Meeting of Human Factors and Ergonomics Society* (HFES), 2523–2527.

James, W. 1890. *Principles of Psychology*. Henry Holt & Co.

Jeffries, R., Miller, J., Wharton, C., and Uyeda, K. 1991. User interface evaluation in the real world: A comparison of four techniques. In *Proceedings of the ACM Conference on Human Factors in Computing*. ACM, New York, pp. 119–124.

John, P. W. M. 1971. *Statistical Design and Analysis of Experiments*. Macmillan, New York.

John, B. E. and Packer, H. 1995. Learning and using the cognitive walkthrough method: a case study approach. In *Proceedings of the ACM Conference on Human Factors in Computing Systems* (CHI'95) (Denver, CO, May 7–11). ACM, New York, 429–436.

John B. and Marks, S. 1997. Tracking the effectiveness of usability evaluation methods. *Behav. Inf. Tech. 16*, 4/5, 188–202.

Johnson, J., Roberts, T., Verplank, W., Smith, D., Irby, C., Bear, M., and Mackey, K. 1989. The Xerox star: A retrospective. *IEEE Comput. 22*, 9, 11–29.

Karat, C.-M., Campbell, R., and Fiegel, T. 1992. Comparison of empirical testing and walkthrough methods in usability interface evaluation. In *Proceedings of ACM Conference on Human Factors in Computing Systems*. ACM, New York, pp. 397–404.

Kogan, N. 1983. Stylistic variation in childhood and adolescence: Creativity, metaphor, and cognitive styles. In *Handbook of Child Psychology: Vol. 3*. Cognitive Development, J. H. Flavell and E. M. Markman, Eds. Wiley, New York, 630–705.

Landauer, T. 1995. *The Trouble with Computer*. The MIT Press, Cambridge, MA.

Law, E. 2006. Evaluating the downstream utility of user tests and examining the developer effect: A case study. *Int. J. Human-Comput. Interact. 21*, 2, 147–172.

Lewis, C., Polson, P., Wharton, C., and Rieman, J. 1990. Testing a walkthrough methodology for theory-based design of walk-up-and-use interfaces. In *Proceedings of the ACM Conference on Human Factors in Computing Systems* (CHI 1990) (Seattle, WA, Apr. 1–5). ACM, New York, 235–242.

Madsen, K. H. 1994. A guide to metaphorical design. *Commun. ACM 37*, 12, 57–62.

Mankoff, J., Dey, A. K., Hsieh, G., Kientz, J., Ames, M., and Lederer, S. 2003. Heuristic evaluation of ambient displays. In *Proceedings of the ACM Conference on Human Factors in Computing Systems* (CHI 2003). ACM, New York, 169–176.

Molich, R. 1994. *Brugervenlige edb-systemer* (in Danish). Teknisk Forlag.

Molich, R. 2003. *User Testing, Discount User Testing*. www.dialogdesign.dk

Naur, P. 1988. Human knowing, language, and discrete structures. In *Computing: A Human Activity*. ACM Press/Addison Wesley, New York, 518–535.

Naur, P. 1995. *Knowing and the Mystique of Logic and Rules*. Kluwer Academic Publishers, Dordrecht, the Netherlands.

Naur, P. 2000. CHI and human thinking. In *Proceedings of the 1st Nordic Conference on Computer-Human Interaction* (NordiCHI 2000) (Stockholm, Sweden Oct. 23–25). ACM, New York.

Naur, P. 2007. Computing versus human thinking. *Commun. ACM 50*, 1, 85–94.

Nielsen, J. 1993. *Usability Engineering*. Academic Press, San Diego, CA.

Nielsen, J. and Mack, R. L. 1994. *Usability Inspection Methods*. Wiley, New York.

Nielsen, J. and Molich, R. 1990. Heuristic evaluation of user interfaces. In *Proceedings of the ACM Conference on Human Factors in Computing Systems* (CHI'90) (Seattle, WA, Apr. 1–5). ACM, New York, 249–256.

Nielsen, J., Molich, R., Snyder, C., and Farrell, S. 2001. *E-commerce User Experience*. Nielsen Norman Group, Fremont, CA.

Norman, D. A. 1986. Cognitive engineering. In *User Centered System Design*. Erlbaum, Hillsdale, NJ. 31–61.

Pinelle, D., Gutwin, C., and Greenberg, S. 2003. Task analysis for groupware usability evaluation: Modeling shared-workspace tasks with the mechanics of collaboration. *ACM Trans. Computer-Hum. Interact. 10*, 4, 281–311.

Preece, J., Rogers, Y. and Sharp, H. 2002. *Interaction Design*. Wiley, New York.

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

Rosenbaum, S., Rohn, J. A., and Humburg, J. 2000. A toolkit for strategic usability: Results from workshops, panels, and surveys. In *Proceedings of ACM Conference on Human Factors in Computing Systems* (CHI 2000). ACM, New York, 337–344.

Sears, A. and Hess, D. 1998. The effect of task description detail on evaluator performance with cognitive walkthroughs. In *Conference Summary on Human Factors in Computing Systems* (CHI'98). ACM, New York, 259–260.

Sfard, A. 1998. On two metaphors for learning and on the dangers of choosing just one. *Educat. Research. 27*, 2, 4–13.

Shneiderman, B. and Plaisant, C. 2005. *Designing the User Interface*. 4th Edition. Addison-Wesley, Reading, MA.

Smith, S. L. and Mosier, J. N. 1986. Guidelines for designing user interface software. Mitre Tech. Rep. ESD-TR-86-278.

Spencer, R. 2000. The streamlined cognitive walkthrough method, working around social constraints encountered in a software development company. In *Proceedings of the ACM Conference on Human Factors in Computing Systems* (CHI'2000) (The Hague, Netherlands). ACM, New York, 353–359.

Wharton, C., Rieman, J., Lewis, C., and Polson, P. 1994. The cognitive walkthrough method: A practitioner's guide. In *Usability Inspection Methods*, J. Nielsen and R. L. Mack, Eds. Wiley, New York, 105–140.

Wixon, D. 2003. Evaluating usability methods: Why the current literature fails the practitioner. *Interactions 10*, 4, 29–34.

Received July 2005; revised September 2006, March 2007, and April 2007; accepted April 2007 by Robin Jeffries