Taylor & Francis
Taylor & Francis Group

Retrospective vs. concurrent think-aloud protocols: testing the usability of an online library catalogue

MAAIKE J. VAN DEN HAAK, MENNO D. T. DE JONG and PETER JAN SCHELLENS

University of Twente, Institute for Behavioural Research, Department of Communication Studies, P.O. Box 217, 7500 AE Enschede, The Netherlands; E-mail: m.j.vandenhaak@utwente.nl

Abstract. Think-aloud protocols are a dominant method in usability testing. There is, however, only little empirical evidence on the actual validity of the method. This paper describes an experiment that compares concurrent and retrospective think-aloud protocols for a usability test of an online library catalogue. There were three points of comparison: usability problems detected, overall task performance, and participant experiences. Results show that concurrent and retrospective think-aloud protocols reveal comparable sets of usability problems, but that these problems come to light in different ways. In retrospective think-aloud protocols, more problems were detected by means of verbalisation, while in concurrent think-aloud protocols, more problems were detected by means of observation. Moreover, in the concurrent think-aloud protocols, the requirement to think aloud while working had a negative effect on the task performance. This raises questions about the reactivity of concurrent think-aloud protocols, especially in the case of high task complexity.

1. Introduction

Think-aloud protocols are a widely used method for the usability testing of software, interfaces, websites, and (instructional) documents. The basic principle of this method is that potential users are asked to complete a set of tasks with the artefact tested, and to constantly verbalise their thoughts while working on the tasks. The method has high face validity, since the data obtained reflect the actual use of an artefact, and not the participants' judgements about its usability. The method is embedded in a well-respected research paradigm focusing on people's cognitive processes during the execution of a wide range of tasks — e.g. writing texts, reading, playing chess, and choosing between alternative options — with the monograph by Ericsson and Simon (1993) as a methodological milestone. Over the years, various textbooks have been published providing detailed instructions on how to conduct a think-aloud usability test (e.g. Nielsen 1993, Rubin 1994, Dumas and Redish 1999, Barnum 2002).

However, the advice that is offered in such text books is hardly supported by methodological research. In their overview of validation research available in the fields of HCI and Document Design, De Jong and Schellens (2000) distinguish between studies focusing on predictive validity (investigating the usefulness of the feedback collected with a particular method), congruent validity (comparing the amount and types of feedback collected with several methods), reliability, sample composition, and the value of evaluation findings in a subsequent revision phase. With regard to predictive validity, only few studies have systematically explored the usefulness of feedback collected with think-aloud protocols. The little research that has been done has addressed the usefulness of think-aloud results in combination with other feedback sources, such as expert evaluation (Jansen and Steehouder 1992, Schriver 1997) or user edits (Allwood and Kalén 1997), but the results of these studies cannot be used to judge the contribution of think-aloud protocols per se. This lack of research might be due to the high face validity of think-aloud usability testing: there seems to be little doubt whether problems revealed in a usability test are real user problems. Various research contributions, however, have pointed out that the requirement to think aloud could result in reactivity, i.e. that it may affect the way participants handle tasks, the time it takes them to carry out tasks, and their eventual success in task completion (see Ericsson and Simon 1993 for an overview).

Behaviour & Information Technology
ISSN 0144-929X print/ISSN 1362-3001 online © 2003 Taylor & Francis Ltd
http://www.tandf.co.uk/journals
DOI: 10.1080/0044929031000

The research on congruent validity focuses rather strongly on the question as to whether usability experts are able to predict the results of a usability test (e.g. Dieli 1986, John and Marks 1997, Schriver 1997). In general, this does not seem to be the case. Experts evaluating an interface, website, or document may give important suggestions for improvement, but they tend to highlight different problems than a sample of users in a usability test. The use of tools such as heuristics or a cognitive walkthrough procedure also fails to consistently improve the experts' ability to predict the results of a usability test. In addition, there are only few studies comparing think-aloud protocols with other evaluation approaches (e.g. Smilowitz *et al.* 1994, Henderson *et al.* 1995, Allwood and Kalén 1997, Sienot 1997), and those studies that are available have a design and results that are too scattered to offer univocal conclusions about think-aloud protocols as a useful method of usability testing.

The research on reliability, sample composition, and revision on the basis of think-aloud protocols is even more limited. With regard to reliability, two studies suggest that a small sample of five or six participants may already produce more or less stable results (Virzi 1992, Nielsen 1994), but a study by Lewis (1994) led to considerably less optimistic conclusions. Caulton (2001) claims that a heterogeneous sample of participants affects the relationship between sample size and stability and exhaustiveness of the problems detected.

With regard to sample composition, only one recent contribution may be mentioned: Hall *et al.* (forthcoming) investigated whether participants from collectivistic and individualistic cultures differ in the feedback that they produce during a usability test. This appeared to be the case in two respects: (1) individualistic participants formulated their feedback in a more direct way than collectivistic participants, and (2) individualistic participants were more inclined to provide comments that were not directly related to the tasks executed. This result indicates that participant characteristics can have an effect on the feedback collected in a usability test.

With regard to the phase of detecting, diagnosing and revising, some studies have addressed the problem of severity ratings and shown that it appears to be very hard for usability professionals to provide a reliable estimation of the severity of usability problems detected (e.g. Hassenzahl 2000). Finally, Bolton (1993) addresses the issue of detecting problems in think-aloud protocols: she describes and evaluates a procedure she developed for automatically detecting user problems in think-aloud protocols, based on lists of verbal and non-verbal expressions people may use to signal surprise, disapproval, doubts, etcetera.

Being of poor quantity, the quality of the existing research is also questionable, as is shown in a review by Gray and Salzman (1998). Discussing five influential studies which compare usability evaluation methods, they argue that each of these studies has so many flaws that it is not possible to draw firm conclusions from them, let alone that they could guide decisions in adopting a usability test approach. Likewise, Lavery *et al.* (1997) addressed the problem of comparing the results of various evaluation methods, proposed a solution for it in the form of structured usability problem reports, but first and foremost drew attention to the problem of investigating the congruent validity of methods.

In a recent contribution, Boren and Ramey (2000) cast doubt on the methodological foundations of think-aloud usability testing. They observed that the strict guidelines prescribed by Ericsson and Simon (1993), with a facilitator who remains in the background and only reminds participants to 'keep thinking aloud' whenever they stop doing so, are hardly complied with in practice. Therefore, they propose a 'speech communication' paradigm that allows the facilitator more freedom to interact with participants. This is motivated by a review of the differences in purpose between research into cognitive processes and research into usability testing. Years earlier, Wright and Monk (1991) had already come to similar conclusions in an experimental comparison of a strict to a more liberal interpretation of think-aloud research, which failed because none of the facilitators in the 'strict' condition behaved according to the guidelines prescribed.

All in all, there are considerably more uncertainties regarding the value and the optimal design of think-aloud usability testing than are suggested in the numerous textbooks available. Many aspects of think-aloud usability testing deserve serious and systematic research attention. The current paper is part of a larger research project that focuses on the merits and restrictions of variations of think-aloud protocols for usability testing. It describes a first experiment, comparing concurrent and retrospective think-aloud protocols for the evaluation of an online library catalogue. Retrospective think-aloud protocols, also known as 'retrospective testing' (Nielsen 1993) or 'aided subsequent verbal protocol' (Henderson *et al.* 1995), differ in one respect from concurrent think-aloud protocols: rather than thinking aloud while working, participants initially carry out their tasks working silently, and only verbalise their thoughts afterwards on the basis of a video recording of their task performance.

Theoretically, there are both benefits and drawbacks to using retrospective think-aloud protocols instead of concurrent think-aloud protocols. One benefit involves a possible decrease in reactivity: participants are fully enabled to execute a task in their own manner and pace, and are therefore not likely to perform better or worse than usual. Concurrent thinking aloud, on the other hand, is more prone to reactivity: participants may perform better than usual as a result of a more structured working process, or they may perform worse as a result of their double workload (Russo *et al.* 1989). A second benefit concerns the recording of working times per task, which is possible in the case of retrospective think-aloud protocols, but which would not be useful in the case of concurrent think-aloud protocols, since the requirement to think aloud is thought to slow down the process of task execution in variable degrees. A third advantage would be that participants have the possibility to reflect on their process of using the artefact, which might cause them to highlight higher-level causes for individual usability problems. Finally, with regard to usability testing which is carried out across cultures involving multiple languages, retrospective thinking aloud may be an appealing alternative to traditional think-aloud tests, since it is probably less difficult for participants to verbalise their thoughts in a foreign language after their task performance than while working.

Apart from benefits, using retrospective think-aloud protocols instead of concurrent think-aloud protocols also has some drawbacks. One drawback relates to the duration of the participant sessions, which is considerably longer for retrospective think-aloud protocols, since the participants not only perform their tasks but also watch these in retrospect. Another, more important drawback concerns the fact that participants may produce biased accounts of the thoughts they had while performing the tasks. They may, for instance, forget specific things that occurred during a task. Ericsson and Simon (1993) emphasise that vital information may be lost in the case of retrospective research, which is confirmed by several studies (e.g. Russo *et al.* 1989, Teague *et al.* 2001). Much depends, however, on the stimuli participants get to help them recall their thoughts. In the case of retrospective thinking-aloud, participants are immediately exposed to a recording of the entire process they went through, which places the method more or less in an intermediate position between concurrent and retrospective research, and makes it less vulnerable to criticism. Bias may also arise as a result of participants deciding to conceal thoughts they had, invent thoughts they did not have, or modify their thoughts, for reasons of self-presentation or social desirability. While participants in the concurrent think-aloud method may make similar decisions, the participants in the retrospective think-aloud method have more opportunity to do so as they are reflecting on their work only after they performed it. Nevertheless, they are at all times bound to the events that are recorded, and hence are considerably less free to edit their thoughts than in the case of unaided retrospective methods.

The literature on usability testing tends to describe concurrent and retrospective think-aloud protocols as equal alternatives (e.g. Nielsen 1993). However, there is very little empirical evidence to support this standpoint. Several studies claim to compare concurrent and retrospective verbal protocols, while in fact they describe the kind of research which was previously referred to as retrospective research, i.e. research which fails to include stimuli to recall the task performance (Branch 2000, Kuusela and Paul 2000, Taylor and Dionne 2000).

So far, only two studies have indeed compared actual retrospective and concurrent think-aloud protocols. Hoc and Leplat (1983) used the two types of think-aloud protocols to investigate a problem-solving process of participants (they had to order a set of letters on a computer screen using a limited set of commands). In the retrospective condition, participants were first asked to give an unaided account of their process, and after that had to think aloud while watching all the steps in the process, which had been recorded in a computer log file. They conclude that unaided retrospective accounts should be avoided, because of the distortions and gaps in the protocols, but that the retrospective and concurrent think-aloud protocols produce similar results. It should be noted, however, that both the task given to the participants (which more or less resembled a logical puzzle) and the analysis of the results (focusing more on strategies than on problems encountered) do not correspond to the situation of usability testing.

Bowers and Snyder (1990) compared the two think-aloud variations in a usability test focusing on the handling of multiple windows on a computer screen. They found no significant differences regarding task performance and task completion time, but the retrospective think-aloud condition resulted in considerably fewer verbalisations, and these were often of a different type than the concurrent verbalisations, focusing more on explanations and less on procedures. While these results are interesting, the study has a serious drawback in that it does not report on the number and kinds of problems detected by the participants in the two think-aloud conditions. As problem detection is typically one of the most important functions of usability testing, this meant that a crucial aspect was not included in the comparison of the two methods.

This paper addresses the lack of literature on concurrent vs. retrospective think-aloud protocols by comparing the two think-aloud variations for the purpose of usability testing. Three research questions will be addressed:

- Do concurrent and retrospective think-aloud protocols differ in terms of numbers and types of usability problems detected?
- Do concurrent and retrospective think-aloud protocols differ in terms of task performance?
- Do concurrent and retrospective think-aloud protocols differ in terms of participant experiences?

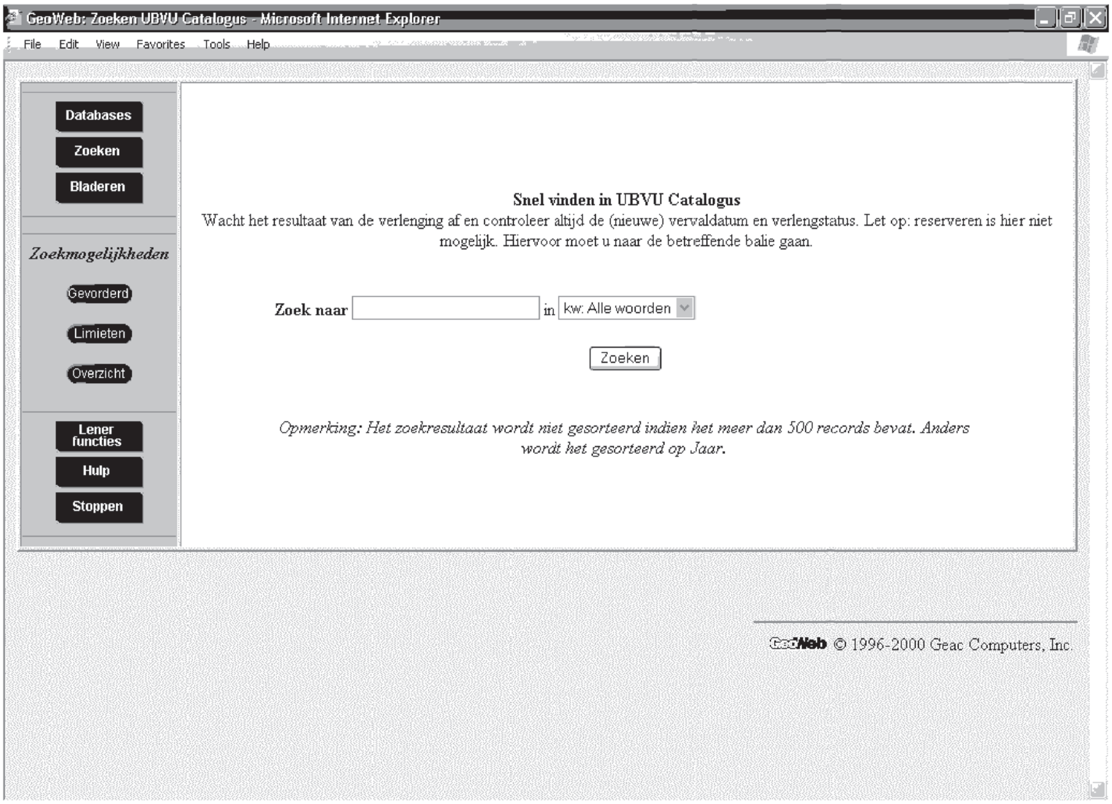
2. Method

2.1. *Test object*

The object of this study was the online library catalogue (UBVU) of the Vrije Universiteit, Amsterdam, the Netherlands. The reason for choosing this particular object lies in the fact that online catalogues combine the characteristics of a search engine with the online features of a website: they are task-focused, they require substantial use of navigation, and they are often complex, especially for novice users. Given these features, they are obvious candidates for usability testing. This is increasingly expressed in the literature on library and information science, which contains a fair number of publications on the usability testing of online catalogues (e.g. Campbell 2001, Battleson *et al.* 2001, Norlin and Winters 2002).

The UBVU catalogue was set up some years ago and has not been subjected to change ever since. As figure 1 below shows, the catalogue has a very simple layout, consisting of a homepage with a search engine positioned in the middle, and nine buttons to the left. These buttons represent search options that are standard to most online catalogues, allowing the user to conduct simple or advanced searches and to sort or browse through results. As with most catalogues, the UBVU also features a help function with information on how to use the catalogue.

While the catalogue is primarily intended for students and employees of the university, it can, with the exception of some restricted areas like 'loaning' or 'reserving', also be accessed by people outside the university. All the information within the catalogue can be viewed in both Dutch and English, except for the help function, which is offered only in English.



Snel vinden in UBVU Catalogus

Wacht het resultaat van de verlenging af en controleer altijd de (nieuwe) vervaldatum en verlengstatus. Let op: reserveren is hier niet mogelijk. Hiervoor moet u naar de betreffende balie gaan.

Opmerking: Het zoekresultaat wordt niet gesorteerd indien het meer dan 500 records bevat. Anders wordt het gesorteerd op Jaar.

GeoWeb © 1996-2000 Geac Computers, Inc.

Figure 1. Homepage of the UBVU web catalogue.

2.2. *Participants*

The research was conducted with a sample of 40 participants, all of whom were students of Communication Studies at the University of Twente. At the time of the study, all participants were in their second or third year of education, which generally meant that they had some knowledge of online library catalogues. As they attended a different university than the one hosting the UBVU catalogue, none of them had worked with this particular catalogue before. As such, the participants were in a good position to evaluate the UBVU catalogue: they were novice users of this particular catalogue *and* they belonged to the main target group.

The participants were gathered by means of printed and e-mail announcements, asking them to participate in the experiment in return for a financial reward. Participants were selected on a 'first come, first serve' basis: other than the requirement that they were second or third year students of Communication Studies, there were no participation criteria for sex, age, etc. In the end, five male and 35 female participants took part in the experiment, ranging in age from 18 to 24. The participants were evenly assigned to the two conditions in the experiment with no difference in gender, age, and prior knowledge of online catalogues.

2.3. *Tasks*

In order to evaluate the UBVU catalogue by means of the concurrent and retrospective think-aloud protocols, seven search tasks were formulated that together cover the catalogue's main search functions. All tasks were designed to be equally difficult, and could be carried out independently from one another, in order to prevent participants getting stuck after one or two tasks. The entire set of tasks was as follows:

(1) Find how many publications the UBVU catalogue has on the topic 'communication';
(2) Find how many publications the UBVU catalogue has on the topic 'language or interaction';
(3) Find how many publications the UBVU catalogue has that are written by A. Hannay;
(4) Find which author within the UBVU catalogue has written most books on the topic 'pop music';
(5) Find how many Dutch publications the UBVU catalogue has on the topic 'Shakespeare';
(6) Find how many publications the UBVU catalogue has on the topic 'telecommunication', that were published from 1999 onwards;
(7) Find how many publications the UBVU catalogue has on the topic 'web-' (i.e. web site, web shop, web communication) within the context of the Internet.

Tasks 1 to 4 were designed to evaluate the catalogue's 'simple search', 'advanced search' and 'sort results' functions. Tasks 5 and 6 focused on the narrowing down of search results (in terms of language and year of publication), and task 7 was designed to evaluate the notion of truncation (a bibliographic term similar to the more well-known wild card search option).

2.4. *Questionnaires*

Apart from the seven tasks, the study also included two questionnaires, designed to be filled in by all participants in both conditions. The first questionnaire, which was handed to the participants at the start of the experiment, contained questions on the demographic details of the participants, such as age, gender, and education. It also enquired after the participants' experience in working with online catalogues, with questions like 'Have you ever followed a course in using (online) library catalogues?', 'Are you familiar with the following library functions (boolean operators, truncation, . . .)?', etc.

The second questionnaire, which was given to the participants at the end of the experiment, was designed in order to measure how the participants had felt about their participation in the experiment. It contained questions on three main aspects: (1) the participants' experiences on having to think aloud (concurrent or retrospectively); (2) the participants' estimation of their method of working on the seven tasks (e.g. more vs. less structured, faster vs. slower than normal); and (3) the participants' judgments about the presence of the facilitator and the recording equipment. For each of these three aspects, participants were asked to rate their experiences on five-point scales based on semantic differentials. In addition, the questionnaire offered extra space for additional comments.

2.5. *Experimental procedure*

The experiment was carried out in 40 individual sessions, which were all held in the same usability lab. During each session, video recordings were made of the computer screen and the participant's voice, while the facilitator was also present to observe and take notes.

In the concurrent think-aloud condition (CTA), the experimental procedure was as follows. Upon arriving, each participant was asked to fill in the first questionnaire on personal details and knowledge of online library catalogues. After completing this questionnaire, the participant was given the UBVU tasks and oral instructions on how to carry them out. These instructions, which were read out from paper to ensure consistency, told the participant to: 'think aloud while performing your tasks, and pretend as if the facilitator is not there. Do not turn to her for assistance. If you fall silent for a while, the facilitator will remind to keep thinking aloud. Finally, remember that it is the catalogue, and not you, who is being tested'. Once the participant had finished the tasks according to these instructions, s/he was given the second questionnaire to indicate how s/he had experienced her/his participation.

In the retrospective think-aloud condition (RTA), the experimental procedure started, again, with the questionnaire on personal details and prior knowledge. As in the first condition, the participants were then given the UBVU tasks and oral instructions, but here they were instructed to simply carry out the tasks in silence, again without seeking assistance from the facilitator. Having done that, they were asked to watch their recorded performance on video and comment on the process retrospectively. Finally, they were given the second questionnaire with questions on how they had experienced their participation in the experiment.

2.6. *Processing of the data*

Once the 40 sessions were completed, verbal transcripts were made of the concurrent and retrospective think-aloud comments, and all the participants' navigations through the catalogue were noted down. The participants' navigation and other actions were studied in order to detect usability problems in the process of using the UBVU. As a rule, a particular situation was marked as a problem when it deviated from the optimum working procedure for each task. The think-aloud protocols were scanned for verbal indicators of problems experienced, referring, for instance, to doubt, task difficulty, incomprehensibility, or annoyance regarding the use of the catalogue.

The analysis of the think-aloud data focused on three main issues. First, the total number of usability problems detected in each condition was examined. After that, a distinction was made according to the way the usability problems had surfaced in the data: (1) through observation of the behavioural data; (2) through verbalisation by the participant; or (3) through a combination of observation and verbalisation. Finally, a categorisation of types of problems was made. For the specific combination of think-aloud data and online catalogues, there was no standard list of possible problem types available. Based on a decomposition of the search process and a review of the data collected, the following five problem types were distinguished:

Layout problems: The participant fails to spot a particular element within a screen of the catalogue;
Terminology problems: The participant does not comprehend part(s) of the terminology used in the catalogue;
Data entry problems: The participant does not know how to conduct a search (i.e. enter a search term, use dropdown windows, or start the actual searching);
Comprehensiveness problems: The catalogue lacks information necessary to use it effectively;
Feedback problems: The catalogue fails to give relevant feedback on searches conducted.

Apart from these five types of problems, participants also occasionally experienced technology problems, such as trouble with the network connection, the browser, or the computer used. These problems were excluded from the analyses.

With regard to task performance, two indicators were used: tasks completed successfully and time required to complete the tasks. These indicators were applied both per task and for the overall performance of the seven tasks.

3. Results

Section 3.1 presents the results regarding the feedback collected with the two kinds of think-aloud protocols. Section 3.2 describes the results in terms of task performance. Section 3.3 addresses the participants' experiences during the usability tests, as measured by the second questionnaire.

3.1. *Number and types of problems detected*

After analysing the 40 recordings, a total number of 72 different problems were found. While some of the problems were detected by almost all (30 to 35) participants, more than half of the total number of different problems were detected by only five or fewer of the 40 participants. This indicates that there were quite a few individual problems: problems which were found by some participants, but which were unproblematic for most other participants.

Table 1 gives an overview of the mean number of problems detected per participant. In the table, a distinction is made according to the way the problems

Table 1. Number of problems detected per participant in the CTA and RTA condition, sorted by the way the problems surfaced in the test.

	Concurrent think-aloud protocols		Retrospective think-aloud protocols		
	Mean	SD	Mean	SD	Significance
Observed	6.7	2.2	4.0	2.0	$p < 0.001$
Verbalised	0.5	0.7	4.5	3.4	$p < 0.001$
Observed and verbalised	6.7	4.0	5.1	2.2	n.s.
Total	13.9	3.3	13.6	4.1	n.s.

had surfaced: (1) by observation; (2) by verbalisation; or (3) by a combination of observation and verbalisation. There was no significant difference in the total number of problems detected by the two think-aloud variations. On a global level, concurrent and retrospective think-aloud protocols were comparable in terms of their quantitative output.

The two methods did, however, differ significantly as to how this output came about. With regard to the manner of problem-detecting, the RTA condition clearly revealed more problems by means of verbalisations only (*t*-test, $t = 5.168$, $df = 38$, $p < 0.001$, Cohen's $d = 1.29$). While the RTA participants on average verbalised 4.5 problems that were not otherwise observable, the verbal protocols of the CTA participants resulted in a meagre 0.5 problems per person. This notable difference may be explained by the fact that the RTA participants simply had more time to verbalise problems. Unlike the CTA participants, the RTA participants commented on the catalogue only after finishing their tasks, which meant that they could fully concentrate on evaluating the catalogue. This gave them more opportunity to not only verbalise the problems they had experienced while working, but also comment on additional problems. The CTA participants, on the other hand, had to verbalise and work at the same time, which gave them less time to comment on problems that were not acute, i.e. that did not directly arise from their task performance. As they first and foremost focused on their tasks, they mainly verbalised their actions and the problems that arose as a result of these actions. This is also reflected in the number of problems that were detected both by a combination of observation and verbalisation: 93% of all comments made by CTA participants corresponded to an observable problem in their task execution, compared to 54% of the comments of the RTA participants.

Another significant difference between the two think-aloud conditions lies in the number of problems detected by means of non-verbal indicators, i.e. by observation only (*t*-test, $t = 4.083$, $df = 38$, $p < 0.001$, Cohen's $d = 1.63$). As table 1 shows, the CTA condition resulted in considerably more observable problems (6.7) than the RTA condition (4.0). Apparently, the participants in the CTA condition experienced more observable difficulties while performing their tasks than their RTA colleagues. This difference could again be attributed to the different workload in both conditions: while the RTA participants had only their tasks to perform, the CTA participants were asked to perform tasks *and* think aloud. It is conceivable that this extra burden had a negative influence on the task performance of the CTA participants, causing them to experience additional problems while working.

To investigate the types of problems detected in both conditions, all problems were labelled according to the problem types that were described in section 2.6. Table 2 shows a selection of problems as they occurred in the think-aloud protocols.

Table 3 shows the overall distribution of problem types in CTA and RTA. There were no significant differences between the types of problems detected in the two conditions. Both the CTA and the RTA condition brought to light all five problem types in similar frequencies. Terminology and data entry clearly presented most problems to the participants in both conditions.

The analyses discussed so far have focused on the overall trends in the results, and have not yet looked into the individual problems detected. A comparison of the lists of problems detected in both conditions offers a first impression of the degree of overlap between CTA and RTA problems. Of the 72 problems that were detected, 47% were reported in both conditions, 31% were detected exclusively in the CTA condition, and another 22% were detected exclusively in the RTA condition. There is more overlap when the frequency of the problems is taken into account. Table 4 shows that 89% of all the problem detections involved problems that were experienced by participants in both conditions.

In all, the overall picture that arises is one in which the CTA and RTA are comparable in terms of number and types of problems detected. The two methods differ,

Table 2. Examples of problem types detected in the think-aloud protocols.

Layout	The participant has trouble finding the advanced search button on the catalogue's homepage
	The participant cannot locate the names of co-authors in the catalogue's result list
Terminology	The participant does not understand the meaning of the term 'limits'
	The participant does not understand the meaning of the term 'truncation'
Data entry	The participant has trouble using the boolean operators
	The participant does not know how to enter dates in the 'year' box
Comprehensiveness	Authors' names are missing in the result list
	The help function offers information only in English, not in Dutch
Feedback	The catalogue fails to provide an error notice when the participant makes a mistake
	The catalogue fails to indicate how its results are sorted (by year, author, etc.)

Table 3. Types of problems detected per participant in the CTA and RTA condition.

	Concurrent think-aloud protocols		Retrospective think-aloud protocols		
	Mean	SD	Mean	SD	Significance
Layout	2.9	1.2	2.6	1.3	n.s.
Terminology	4.1	1.5	4.1	2.0	n.s.
Data entry	4.9	1.2	4.9	1.2	n.s.
Comprehensiveness	1.1	0.9	1.2	0.6	n.s.
Feedback	1.0	1.0	0.9	0.6	n.s.

Table 4. Percentage of problem detections unique to either condition.

	Unique to CTA	Unique to RTA	Detected in both
Layout	10	12	78
Terminology	1	6	93
Data entry	6	2	92
Comprehensiveness	11	4	84
Feedback	8	2	91
Total	6	5	89

however, with regard to the manner of detecting: while the CTA method reveals more problems that can be observed during task performance, the RTA method depends more on the participants' verbalisations. These verbalisations play a significantly less substantial role in the CTA method. This result is remarkable, because the rationale of thinking aloud as a usability test approach is that the verbal protocols result in the detection of problems. Apparently, the verbal protocols in this study do not so much serve to reveal problems but rather to verbally support the problems that are otherwise observable. The fact that these observable problems are significantly more substantial in the CTA method might, as suggested before, be explained by the double workload of the CTA participants. For this reason, it would be interesting to investigate whether this double workload has had an effect on the participants' task performance.

3.2. *Task performance*

Two indicators of task performance were used in this study: the successful completion of the seven tasks, and the time it took the participants to complete them. Table 5 presents the results of both indicators. Both with regard to the overall task completion time and the time per task, no significant differences were found. Apparently, concurrent thinking aloud did not slow down the process of task performing. However, the participants' double workload did have an effect on the overall completion of tasks, in that the CTA participants were significantly less successful in completing their tasks than the RTA participants (*t*-test, $t = 2.252$, $df = 38$, $p < 0.05$, Cohen's $d = 0.71$). There were no significant differences with regard to individual tasks. This result is in line with the conclusion previously drawn that the CTA protocols contained more observable problems

Table 5. Task performance in the CTA and RTA condition.

	Concurrent think-aloud protocols		Retrospective think-aloud protocols		
	Mean	SD	Mean	SD	Significance
Number of tasks completed successfully	2.6	1.0	3.3	1.0	$p < 0.05$
Overall task completion time in min	21.1	5.7	19.6	5.0	n.s.

Table 6. Participant experiences on having to think aloud.

	Concurrent think-aloud protocols		Retrospective think-aloud protocols		
	Mean	SD	Mean	SD	Significance
Difficult – easy	2.4	0.8	2.7	1.2	n.s.
Unpleasant – pleasant	2.7	0.8	2.9	1.0	n.s.
Tiring – not tiring	3.4	1.0	3.8	1.4	n.s.
Unnatural – natural	3.4	0.9	3.0	1.5	n.s.
Time-consuming – not time-consuming	3.2	1.2	3.2	1.1	n.s.

Note: Scores on a five-point scale (1 = negative, 5 = positive).

than the RTA protocols, which would be an indication of reactivity. It should be pointed out, however, that the participants in general had difficulty in performing the tasks: on average, only 40% of the tasks were completed successfully. In the CTA condition, the average successful completion amounted to 2.6 tasks (SD 1.0, range 1 to 4 tasks); in the RTA condition, the average successful completion amounted to 3.3 tasks (SD 1.0, range 2 to 5 tasks). The most difficult task (task 7) was completed successfully by only one of the 40 participants; the easiest task (task 4) by as many as 38 out of 40 participants. This finding will be elaborated on in the discussion.

3.3. *Participant experiences*

The questionnaire on participant experiences served to establish how the participants in both conditions had felt about participating in the study. Questions involved three aspects of the experiment: (a) experiences with concurrent or retrospective thinking aloud; (b) method of working; (c) presence of the facilitator and the recording equipment.

Participants were asked, first of all, how they had felt about having to think aloud concurrently or retrospectively by indicating, on a five-point scale, to which degree they thought this activity was difficult, unpleasant, tiring, unnatural, and time-consuming. Together, these variables failed to form a reliable scale, so each variable was analysed individually. These individual analyses (see table 6) showed that there were no significant differences as to how the participants in both conditions experienced the concurrent or retrospective thinking aloud. On average, the participants rated their experiences with thinking aloud rather neutrally, with scores ranking around the middle of the five-point scale. For the CTA condition, this meant that the notion of reactivity, which was described in section 3.2 as a possible negative influence on CTA participants, is not experienced as such by the participants themselves.

Participants were also asked to estimate in what respect(s) their working procedure differed from usual, by marking, on a five-point scale, how much faster or slower, more focused or less focused, etc. they had worked than they would usually do. Results, which are shown in table 7, showed that there were no significant differences between CTA and RTA. In both conditions, the participants estimated that their behaviour differed only slightly from their normal working procedure. After recoding the variables to investigate any deviation (to either side of the scale) from the regular working procedure, the eight variables formed a reliable scale (Cronbach's alpha = 0.84), which showed that the participants in the RTA condition had, in their view, worked significantly more differently during the experiment than the participants in the CTA condition (with a mean deviation of 0.33 vs. 0.29; *t*-test, $t = 2.242$, $df = 38$, $p < 0.05$, Cohen's $d = 0.72$). So, in contrast to the conclusions regarding problems detected and task performance, the participants in the RTA condition experienced more reactivity of the test situation than the

Table 7. Participants' method of working, compared to their usual working procedure.

	Concurrent think-aloud protocols		Retrospective think-aloud protocols		
	Mean	SD	Mean	SD	Significance
Faster – slower	2.7	0.7	2.3	0.8	n.s.
More – less focused	2.6	0.6	2.1	0.9	n.s.
More – less concentrated	3.3	0.6	3.5	0.9	n.s.
More – less persevering	2.6	0.9	2.7	0.9	n.s.
More – less successful	3.0	0.5	2.9	0.7	n.s.
More – less pleasant	3.2	0.5	3.4	0.6	n.s.
More – less eye for mistakes	2.6	0.7	2.2	0.7	n.s.
More relaxed – more stressful	3.4	0.6	3.7	0.5	n.s.

Note: Scores on a five-point scale (3 = no difference from usual).

Table 8. Participants' experiences of the test situation: presence of facilitator and recording equipment.

	Concurrent think-aloud protocols		Retrospective think-aloud protocols		
	Mean	SD	Mean	SD	Significance
Unpleasant	2.8	0.3	2.7	0.8	n.s.
Unnatural	2.9	0.7	3.1	1.3	n.s.
Disturbing	4.3	0.6	3.7	0.9	$p < 0.05$

Note: Scores on a five-point scale (1 = negative, 5 = positive).

participants in the CTA condition. This finding might be due, however, to the moment of filling in the questionnaire in the RTA condition, which was after watching the video recording and verbalising. It is well imaginable that the artificial task of verbalising afterwards and hence the participants' reflection on their working method have affected the judgments given in the questionnaire.

The final part of the questionnaire included questions on the presence of the facilitator and the use of recording equipment. Participants were first asked to indicate, once again on a five-point scale, to which degree they found it unpleasant, unnatural or disturbing to have the facilitator present during the experiment. They were then asked the same question with regard to the use of the recording equipment. For all three qualifications of the test situation, a sufficiently reliable two-item scale could be formed (Cronbach's alpha = 0.66 for 'unpleasant', 0.81 for 'unnatural', and 0.62 for 'disturbing'). The results are presented in table 8. The scores regarding pleasantness and naturalness are neither negative nor positive, and do not differ significantly between the two conditions. The scores regarding a disturbing test situation are rather positive in both conditions, but the CTA participants found the test situation less disturbing than the RTA participants (*t*-test, $t = 2.368$, $df = 33.4$, $p < 0.05$, Cohen's $d = 0.75$). This difference between the two conditions may again be explained by the time at which the RTA participants filled in the questionnaire. Another explanation would be that the presence of the facilitator during the first part of the RTA test (silent task performance) is less functional than in a CTA design, and that it may be confronting for participants to see their actions back on video. A last possible explanation would be the workload of the participants. The CTA participants had to actively perform tasks *and* think aloud, which considerably reduced the amount of attention they could spare for noticing the facilitator and the recording equipment. The RTA participants, on the other hand, were only performing one task at a time, which gave them more opportunity to pay attention to the facilitator and the recording equipment.

All in all, the participant experiences in the usability tests provide additional support for the usefulness of both the CTA and the RTA method. All measures included in the questionnaire yielded neutral to positive judgments for the two evaluation methods. There were some differences between the CTA and the RTA condition, though, that were not in line with the data about problems detected and task performance reported in sections 3.1 and 3.2. The participants in the RTA condition reported more reactivity as a result of the test situation, and found the test situation more disturbing than the CTA participants. This may reflect a real difference between the two methods, but it is also likely that it is caused by the research procedure chosen (with RTA participants filling in the questionnaire not immediately after performing the seven tasks, but after the second round of watching the video recording and verbalising).

4. Discussion

The present study shows that there are both similarities and significant differences between concurrent and retrospective think-aloud protocols. The differences that were found between the two think-aloud variations provide new insights into the validity of think-aloud protocols for usability testing. While both methods were comparable in terms of quantitative output, they differed significantly as to how this output was established. The CTA method resulted in significantly more problems detected by means of observation only. The RTA method, on the other hand, proved significantly more fruitful in revealing problems that were not observable, but could only be detected by means of verbalisation. These results indicate that the CTA method is a more faithful representative of a strictly task-oriented usability test, while the RTA method is likely to yield a broader gamut of user reactions. This is in line with the earlier comparison of CTA and RTA by Bowers and Snyder (1990), who found that RTA participants tended to give explanations and suggestions, while CTA participants more often limited themselves to giving descriptions of their actions. To investigate the usefulness of the feedback collected with both methods, further research into the predictive validity of concurrent and retrospective think-aloud protocols is essential: how important are the problems reported? Are there many false alarms, particularly in the observable CTA problems and the verbalised RTA problems?

Regarding the use of concurrent think-aloud protocols, the results of this study highlight two important issues. The first is the very limited contribution of the participants' verbalisations to the outcome (in terms of user problems detected) of the usability test. The participants' verbalisations only marginally resulted in the detection of problems, but served predominantly to emphasise or explain the problems that could also be observed in the participants' actions. Naturally, this may still be an important contribution, especially for the subsequent steps of diagnosing the user problems and estimating their severity. Nevertheless, the concurrent verbalisations played a less substantial role in the present study than is usually suggested in handbooks on usability testing.

A second, more important observation is that the CTA method caused reactivity in the usability test. This corresponds to earlier findings by Russo *et al.* (1989) who studied the validity of think-aloud protocols for investigating a variety of cognitive tasks, and found that thinking aloud could both enhance and impede task performance. But it contradicts the results by Bowers and Snyder (1990), who found no differences in task performance between CTA and RTA participants. In the present study, thinking aloud had a consistent and plausible negative effect on task performance. The task of concurrently verbalising thoughts caused the participants to make more errors in the process of task performing and to be less successful in completing the seven tasks. This finding casts doubt on using task outcome in a CTA evaluation as an overall indication of the usability of an artefact, and on the implicit assumption that the problems found in a think-aloud usability test are by definition real user problems. Research into the predictive validity, as defined by De Jong and Schellens (2000), of think-aloud usability data is not a superfluous effort to establish what is already known, but an important step to further explore the method's reactivity. There is always a possibility that a problem detected in a CTA usability test is (partly) caused by the method used. In this study, for that matter, the task to concurrently think aloud caused more extra (observed) problems than it revealed in the participants' verbalisations. Whether this is harmful or not is as yet open to discussion. Most usability tests aim at identifying and diagnosing user problems in an artefact, and it could be argued that it is helpful that such problems come to light easily in a CTA test, provided that they reflect the problems real users have in normal situations.

The most plausible explanation for the two observations regarding the CTA method lies in the participants' workload: the difficulty of the tasks given to the participants may have been a crucial factor in this study. The data on task performance show that the seven tasks given to the participants were very difficult for them. The cognitive load of the tasks combined with the extra task of thinking aloud appears to have had a negative effect on both the participants' verbalisations *and* their task performance. The gaps in verbalisations are supported by Ericsson and Simon (1993: 91), who claim that participants may stop verbalising when they are under a high cognitive load. The negative effect on task performance, however, is not univocally explained by the existing literature (Russo *et al.* 1989, Ericsson and Simon 1993). Indeed, some studies even show that concurrent thinking aloud has a positive effect on task performance (Loxterman *et al.* 1994). It would therefore be interesting to further investigate the three-way relationship between task difficulty, degree of verbalisations, and task performance in CTA participants.

A final remark concerns the generalisability of the present study. Readers should note that this is only a first comparative study, which involved only one artefact. An important characteristic of the UBVU catalogue and the tasks used in this study is that there was much to observe in the way people interact with the computer. The task performance of participants could easily be segmented into steps and analysed without verbalisations. It would be interesting to investigate whether the same results will also be found in applications with a less overt usage process. A replication of this study using documentation, websites, or interfaces with a more open task domain could be an interesting follow-up to further explore the CTA and RTA methods.

All in all, the results of this study indicate that concurrent and retrospective think-aloud protocols can be regarded as equivalent, but clearly different evaluation methods. A strong, and new argument in favour of RTA protocols is that they may be less susceptible to the influence of task difficulty, both in terms of reactivity and in terms of completeness of the verbalisations. Directions offered for think-aloud research often state that the researcher should formulate tasks with a moderate difficulty, so that participants are not inclined to follow an automated working process, but will also not be burdened with a cognitive load that is too high. In usability testing, however, this guideline is not always practical. After all, neither the quality of the artefact tested nor the selection of realistic tasks are within the control of the usability test team.

References

Allwood, C. M. and Kalén, T. 1997, Evaluating and improving the usability of a user manual. *Behaviour & Information Technology*, **16,** 43 – 57.

Barnum, C. M. 2002, *Usability Testing and Research* (New York: Longman).

Battleson, B., Booth, A. and Weintrop, J. 2001, Usability testing of an academic library web site: a case study. *Journal of Academic Librarianship*, **237,** 188 – 198.

Bolton, R. L. 1993, Pretesting questionnaires: content analyses of respondents' concurrent verbal protocols. *Marketing Science*, **12,** 280 – 303.

Boren, M. T. and Ramey, J. 2000, Thinking aloud: reconciling theory and practice. *IEEE Transactions on Professional Communication*, **43,** 261 – 278.

Bowers, V. A. and Snyder, H. L. 1990, Concurrent versus retrospective verbal protocols for comparing window usability. Human Factors Society 34th Meeting, 8 – 12 October 1990 (Santa Monica: HFES), pp. 1270 – 1274.

Branch, J. L. 2000, Investigating the information-seeking processes of adolescents: The value of using think alouds and think afters. *Library & Information Science Research*, **22,** 371 – 392.

Campbell, N. (ed.) 2001, *Usability Assessment of Library-Related Web Sites: Methods and Case Studies*. (Chicago: LITA).

Caulton, D. A. 2001, Relaxing the homogeneity assumption in usability testing. *Behaviour & Information Technology*, **20,** 1 – 7.

De Jong, M. and Schellens, P. J. 2000, Toward a document evaluation methodology: what does research tell us about the validity and reliability of methods? *IEEE Transactions on Professional Communication*, **43,** 242 – 260.

Dieli, M. 1986, *Designing Successful Documents: An Investigation of Document Evaluation Methods* (Dissertation Carnegie Mellon University, Pittsburgh, PA).

Dumas, J. S. and Redish, J. C. 1999, *A Practical Guide to Usability Testing,* Revised edition (Exeter: Intellect).

Ericsson, K. A. and Simon, H. A. 1993, *Protocol Analysis: Verbal Reports as Data* (Cambridge, MA: MIT Press).

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

Hall, M., De Jong, M. and Steehouder, M. forthcoming, Cultural differences and usability evaluation; individualistic and collectivistic participants compared.

Hassenzahl, M. 2000, Prioritizing usability problems: data-driven and judgement-driven severity estimates. *Behaviour & Information Technology*, **19,** 29 – 42.

Henderson, R. D., Smith, M. C., Podd, J. and Varela-Alvarez, H. 1995, A comparison of the four prominent user-based methods for evaluating the usability of computer software. *Ergonomics*, **38,** 2030 – 2044.

Hoc, J. M. and Leplat, J. 1983, Evaluation of different modalities of verbalization in a sorting task. *International Journal of Man-Machine Studies*, **18,** 283 – 306.

Jansen, C. and Steehouder, M. 1992, Forms as a source of communication problems. *Journal of Technical Writing and Communication*, **22,** 179 – 194.

John, B. E. and Marks, S. J. 1997, Tracking the effectiveness of usability evaluation methods. *Behaviour & Information Technology*, **16,** 188 – 202.

Kuusela, H. and Paul, P. 2000, A comparison of concurrent and retrospective verbal protocol analysis. *American Journal of Psychology*, **113,** 387 – 404.

Lavery, D., Cockton, G. and Atkinson, M.P. 1997, Comparison of evaluation methods using structured usability problem reports. *Behaviour & Information Technology*, **16,** 246 – 266.

Lewis, J.R. 1994, Sample sizes for usability studies: additional considerations. *Human Factors*, **36,** 369 – 378.

Loxterman, J. A., Beck, I. L. and McKeown, M. G. 1994, The effects of thinking aloud during reading on students' comprehension of more or less coherent text. *Reading Research Quarterly*, **29,** 353 – 367.

Nielsen, J. 1993, *Usability Engineering* (Boston, MA: Academic Press).

Nielsen, J. 1994, Estimating the number of subjects needed for a thinking aloud test. *International Journal of Human-Computer Studies*, **41,** 385 – 397.

Norlin, E. and Winters, C. M. I. 2002, *Usability testing for library websites: a hands-on guide* (Chicago: American Library Association).

Rubin, J. 1994, *Handbook of Usability Testing: How to Plan, Design, and Conduct Effective Tests* (New York: Wiley).

Russo, J. E., Johnson, E. J. and Stephens, D. L. 1989, The validity of verbal protocols. *Memory & Cognition*, **17,** 759 – 769.

Schriver, K.A. 1997, *Dynamics in Document Design; Creating Text for Readers* (New York: Wiley).

Sienot, M. 1997, Pretesting web sites; a comparison between the plus-minus method and the think-aloud method for the World Wide Web. *Journal of Business and Technical Communication*, **11,** 469 – 482.

Smilowitz, E. D., Darnell, M. J. and Benson, A. E. 1994, Are we overlooking some usability testing methods? A comparison of lab, beta, and forum tests. *Behaviour & Information Technology*, **13,** 183 – 190.

Taylor, K. L. and Dionne, J. P. 2000, Accessing problem-solving strategy knowledge: The complementary use of concurrent verbal protocols and retrospective debriefing. *Journal of Educational Psychology*, **29,** 413 – 425.

Teague, R., De Jesus, K. and Nunes-Ueno, M. 2001, Concurrent vs. post-task usability test ratings. Conference on Human Factors and Computing Systems, 31 March – 5 April 2001 (Seattle, WA: ACM SIGCHI), pp. 289 – 290.

Virzi, R. A. 1992, Refining the test phase of usability evaluation: how many subjects is enough? *Human Factors*, **34,** 457 – 468.

Wright, P. C. and Monk, A. F. 1991, A cost-effective evaluation method for use by designers. *International Journal of Man-Machine Studies*, **36,** 544 – 565.