Available online at www.sciencedirect.com
SCIENCE DIRECT®

Interacting with Computers 16 (2004) 1153–1170

Interacting with Computers
www.elsevier.com/locate/intcom

Employing think-aloud protocols and constructive interaction to test the usability of online library catalogues: a methodological comparison

M.J. Van den Haak*, M.D.T de Jong, P.J. Schellens

Faculty of Behavioural Sciences, Department of Communication Studies, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands

Received 4 May 2004; revised 5 July 2004; accepted 9 July 2004
Available online 13 August 2004

Abstract

This paper describes a comparative study of three usability test approaches: concurrent think-aloud protocols, retrospective think-aloud protocols, and constructive interaction. These three methods were compared by means of an evaluation of an online library catalogue, which involved four points of comparison: number and type of usability problems detected; relevance of the problems detected; overall task performance; and participant experiences. The results of the study showed that there were only few significant differences between the usability test approaches, mainly with respect to manner of problem detecting, task performance and participant experience. For the most part, the usability methods proved very much comparable, revealing similar numbers and types of problems that were equally relevant. Taking some practical aspects into account, a case can be made for preferring the concurrent think-aloud protocols over the other two methods.

Keywords: Usability testing; Concurrent think-aloud protocols; Retrospective think-aloud protocols; Constructive interaction; Co-discovery; Validity

1. Introduction

Prior to reading this article, there may have been a stage at which you started looking for literature on usability testing. In doing so, you may have consulted an

* Tel.: +31-534892291; fax: +31-534894259.
E-mail address: m.j.vandenhaak@utwente.nl (M.J. Van den Haak).

0953-5438/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.intcom.2004.07.007

online library catalogue. Judging from previous evaluation studies of online catalogues (e.g. Battleson et al., 2001; Campbell, 2001; Norlin and Winters, 2002), this will probably have caused you to experience a number of usability problems. Some of these problems may have created real inconvenience; others may have gone unnoticed, but may still have had an adverse effect on your search for literature. In general, the use of online library catalogues is a potential source of user problems. As such it is not surprising that, as in other areas of human-computer interaction, usability testing is a common activity these days.

As a result of this interest in user evaluation, academic research into the validity and reliability of evaluation methods is also on the increase. The importance of this kind of methodological research is stressed, for instance, by the results of the comparative usability evaluation study conducted by Molich et al. (2004). They asked multiple organisations to evaluate the Hotmail website using several usability test approaches, and were surprised by the lack of agreement concerning the number and types of problems that were found. So far, various studies (for an overview, see De Jong and Schellens, 2000) have addressed issues such as the contribution of usability testing to the quality of artefacts, the similarities and differences in evaluation results of comparable methods (e.g. Smilowitz et al., 1994; Henderson et al., 1995; Allwood and Kalén, 1997; Sienot, 1997), the types of feedback given by different user groups, and the relationship between sample size and number of problems detected (Virzi, 1992; Nielsen, 1994; Lewis, 1994).

While the number of studies has been substantial, several major topics have not yet been explored. Nielsen (1993) and Dumas and Redish (1999) mention a few alternative data collection methods for usability testing, such as concurrent and retrospective think-aloud protocols and constructive interaction. Concurrent think-aloud protocols are clearly the most common; they involve participants verbalising their thoughts while performing tasks to evaluate an artefact. Retrospective think-aloud protocols are less frequently used: in this method, participants perform their tasks silently, and afterwards comment on their work on the basis of a recording of their performance. With constructive interaction, which is also known as co-discovery learning, two participants work together in performing their tasks, thereby verbalising their thoughts through interacting. As Miyake (1982) argues, such interaction will not only involve mutual problem-solving processes but also focus on the differences in knowledge domains of the participants.

While there is a substantial body of literature, especially relating to concurrent and retrospective think-aloud protocols, which describes the methods as a tool for uncovering cognitive processes (Ericsson and Simon, 1993), hardly any research has focused on comparing the methods as a tool for usability testing. Moreover, the little research that has been done on the methodology of usability testing is often of a doubtful quality: as Gray and Salzman (1998) pointed out, such research, for instance, fails to include a sufficient number of participants, to use rigorous experimental designs, and/or to perform adequate statistical testing. Moreover, hardly any of this research focused on a comparison of any of the three methods as described above.

This lack of comparative research deserves to be addressed as each of the methods seems to have its own benefits and drawbacks, which may make them more suitable for one purpose than another. For instance, constructive interaction would clearly seem the most suitable method for evaluating collaborative systems (Kahler, 2000). However, according to Nielsen (1993), it is also the most appropriate method for usability testing with children, as it allows them to work with an artefact as naturally as possible. Höysniemi et al. (2003) adopt a similar viewpoint, as they use peer tutoring with children to evaluate a computer game. On the other hand, for a usability test involving only few participants, Nielsen (1993) advocates using retrospective think-aloud protocols, as he assumes that these are most fruitful in terms of problems reported per participant. However, such assumptions are usually based on common sense rather than empirical comparative research.

It seems that only one study compared concurrent think-aloud protocols to constructive interaction. Hackman and Biers (1992) made a comparison between what they called 'team usability testing' and two think-aloud conditions, one with an observer present in the test lab and one where the participant worked entirely alone. They found that the team did not differ from the single user conditions with respect to performance or subjective evaluation of the artefact. However, the team condition revealed more high value feedback in a shorter period of time than the single user conditions. It should be noted, though, that the participants in the team condition were instructed 'to think-out-loud to one another as they work together as a team', which created a very unnatural communicative situation, basically causing the two participants to individually but simultaneously think aloud.

As for the comparison of concurrent and retrospective think-aloud protocols, two types of studies can be distinguished. The first type compares a concurrent think-aloud condition to a silent condition. Such a comparison facilitates analysing the effect of thinking aloud on task performance (reactivity), as is shown by, for instance Van Oostendorp and De Mul (1999). They found that thinking aloud had a positive effect on the learning ability of students exploring a computer system. Apparently, thinking aloud helped them to focus on their own behaviour, which eventually led to an improvement of their performance. It should be noted, however, that this study focuses on learnability, and not on usability, which may include but is not limited to an environment of learning. Another important observation to be made about studies involving a silent condition is that it is questionable whether they should be regarded as an adequate means of detecting usability problems, since these are not verbalised by the participants but can only be revealed by means of observation.

The second type of study compares a concurrent think-aloud condition to a more complete retrospective think-aloud condition (involving a silent session plus verbalisations in retrospect). Bowers and Snyder (1990) first made this comparison in a usability test involving the handling of multiple windows on a computer screen. They found no significant differences regarding task performance and task completion time, but concluded that the retrospective think-aloud condition did result in considerably fewer and different verbalisations than the concurrent think-aloud condition. Their study does have a serious drawback, however, in that it fails to report on the number and kinds of problems detected by the participants in the two think-aloud conditions. With problem detection typically being one of the most important functions of usability testing, Bowers and Snyder did not account for a crucial aspect in their comparison of the two methods.

Van den Haak, De Jong and Schellens (2003) compared retrospective and concurrent think-aloud protocols for the evaluation of an online library catalogue. They focused on the number and types of problems detected, the participants' task performance and the participants' experiences with the two respective methods. Results showed that there were no differences in terms of number of problems detected and experiences of the participants, who all felt positive about their participation. There was, however, a difference in task performance, as the participants in the retrospective condition performed better than the participants in the concurrent condition. This finding was also reflected in the way in which the usability problems came to light: the retrospective participants experienced fewer observable problems than the concurrent participants. Both findings can be explained by reactivity, i.e. the double cognitive workload in the concurrent condition.

Most of the above studies have been conducted on a small scale, and can be regarded as isolated cases, i.e. they are not part of a larger research project on the methodology of usability testing. Moreover, none of them have actually compared all three usability test approaches (concurrent and retrospective think-aloud protocols, constructive interaction). The study that is presented in this paper does form part of a large-scale research programme and it also focuses on a comparison of all three methods. It will address the following research questions:

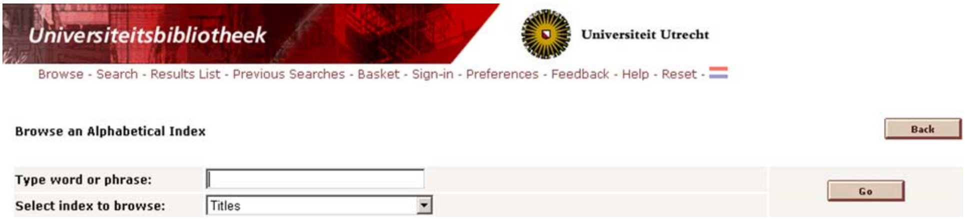
- Do the three methods differ in terms of numbers and types of usability problems detected?
- Do the three methods differ in terms of relevance of the usability problems detected?
- Do the three methods differ in terms of task performance?
- Do the three methods differ in terms of participant experiences?

2. Method

2.1. Test object

As in the previous study, in which we evaluated the UBVU catalogue of the Vrije Universiteit Amsterdam, our test object was an online library catalogue. This time we used the catalogue of Utrecht University (from now on referred to as UBU), at Utrecht, The Netherlands. By using a similar test object and observing the same test procedure (see below) we were able to compare the results of our current study to the results of the previous one.

As Fig. 1 shows, the UBU catalogue has a simple layout, consisting of a home page with a search engine in the middle, and ten links at the top of the page. These links represent various options that are standard to most online catalogues: conducting searches, saving search results, viewing search history, etc. As with


Universiteitsbibliotheek
Universiteit Utrecht
Browse - Search - Results List - Previous Searches - Basket - Sign-in - Preferences - Feedback - Help - Reset -
Browse an Alphabetical Index
Back
Go

Type word or phrase:

Select index to browse: Titles

Fig. 1. Screenshot of the home page of the UBU catalogue.

most catalogues, the UBU also has a help function with information on how to use the catalogue.

Although the catalogue is primarily intended for students and employees of Utrecht University, it can, except for some restricted areas as ‘loaning’ or ‘reserving’, also be accessed by guests, i.e. people outside the university. To accommodate for non-native users, all the information within the catalogue can be viewed in both Dutch and English.

2.2. Participants

Our study was conducted with a sample of eighty participants, all of whom were students at the University of Twente. The students were gathered by means of printed and e-mail announcements, and were given a small financial reward for their participation. Most of them (67 out of 80) were students of Communication Studies; the other 13 took different courses. At the time of the experiment, the majority of the participants (60 out of 80) were in their second or third year of education, which meant that they all had some experience in working with online library catalogues. However, as they were studying at a different university than the one hosting the UBU catalogue, none of the participants had worked with this particular catalogue before. Thus, being novice users as well as part of the target group, they were very suitable candidates for evaluating the UBU catalogue. Seventeen male and 63 female participants took part in the experiment, ranging in age from 19 to 25 (the average age was 21.7). The participants were evenly assigned to the three conditions in the experiment with no difference regarding their demographic details and experience in working with online library catalogues.

2.3. Tasks

To evaluate the UBU catalogue by means of the three usability test approaches, we formulated five search tasks that together covered the catalogue’s main search functions. All tasks were designed to be equally difficult, and could be carried out independently from one another, to prevent participants from getting stuck after one or two tasks. The entire set of tasks was as follows:

1. Find how many publications the UBU catalogue has on the topic ‘public relations’.
2. Find how many publications the UBU catalogue has on the topic ‘language’, excluding the language ‘English’.
3. Find how many publications the UBU catalogue has that are written by A. Hannay.
4. Find how many Dutch publications the UBU catalogue has on the topic ‘Shakespeare’ that were published from 1999 onwards.
5. Find how many publications the UBU catalogue has on the topic ‘web-’ (i.e. web site, web shop, web communication) within the context of the Internet.

These tasks do not represent real-life search tasks in the sense that people are rarely interested in finding how many publications a particular catalogue has on any given topic. Rather than that, users normally start out by searching for a range of publications, from which they then select those titles that are relevant to them. We have chosen not to include this final selection process in the formulation of our tasks, as it relates to the individual preferences of a user and not to the quality of the catalogue. Asking the participants to indicate the number of publications found had two advantages. For one, it offered us an immediate indication of whether a task was completed successfully or not. Secondly, it gave the participants the comfortable feeling that they had sufficiently finalised a particular search process.

Tasks 1–3 were designed to evaluate the catalogue’s ‘simple search’, ‘advanced search’ (using Boolean operators) and ‘sort results’ functions. Task 4 involved the narrowing down of search results (in terms of language and year of publication), and task 5 focused on evaluating the notion of truncation (a bibliographic term comparable to the more well-known wild card search option).

2.4. Questionnaires

Apart from carrying out the above tasks, participants also had to fill in two questionnaires. The first questionnaire, handed out to the participants as they entered the usability lab, contained questions about their demographic details, such as age, gender, and education. It also focused on the participants’ experience in working with online catalogues, containing questions like ‘Have you followed a course in using (online) library catalogues before?’, ‘Are you familiar with any of the following library functions (Boolean operators, truncation, …)?’, etc.

The second questionnaire aimed to measure how the participants had felt about participating in the experiment. It focused on three main aspects: (1) the participant experiences on having to think aloud (concurrent or retrospectively) or work together, (2) the participants’ estimation of their method of working on the five 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 had to rate their experiences on five-point scales based on semantic differentials. The questionnaire also offered space for additional comments.

Participants in the concurrent think-aloud condition (CTA) and the constructive interaction condition (TEAM) filled in the second questionnaire at the very end of the experiment, i.e. after completing their tasks. The participants in the retrospective think-aloud condition (RTA) received their second questionnaire in two parts at two occasions: the first part, with questions pertaining to their method of working, after finishing their tasks; the second part, with questions on how they had experienced thinking aloud, after finishing their retrospective session. To investigate whether the participants would think differently about their method of working after having seen a video recording of their performance, some of the questions that were posed in the first part of the questionnaire (like 'How many tasks do you think you completed successfully?') were repeated in the second part.

2.5. Experimental procedure

We conducted our study in 60 sessions, which we all held in the same usability lab. Twenty sessions were devoted to the 20 participants in the CTA condition; another 20 sessions were devoted to the 20 RTA participants; and a final 20 sessions were devoted to the 40 participants in the TEAM condition (they participated in the experiment in teams of two). During each session, we made video recordings of the computer screen and the participants' voices, and there was one facilitator who was present at all sessions to observe and take notes.

In the CTA condition, the experimental procedure was as follows. When the participant arrived, he or she filled in the first questionnaire on personal details and knowledge of online library catalogues. After completing this questionnaire, the participant received the UBU tasks as well as oral instructions on how to carry them out. These instructions, which the facilitator read out from paper for the sake of consistency, were the following: '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 talking aloud. Finally, remember that it is the catalogue, and not you, who is being tested'. Once the participant had finished his/her task performing, s/he received the second questionnaire to indicate how s/he had experienced her/his participation.

The experimental procedure in the TEAM condition was the following. As in the CTA condition, the two participants in the TEAM condition started out by filling in the first questionnaire. After completing these questionnaires, the participants were seated randomly at the computer, one of them sitting in front of it, the other next to it. They then received instructions which explicitly told them to *work together*, along the lines of 'even though only one of you can actually control the mouse, you have to perform the tasks as a team, by consulting each other continuously and making joint decisions'. As in the CTA condition, the two participants could not turn to the facilitator for assistance. Once the tasks were performed, the participants were each given the second questionnaire to indicate how they felt about participating.

In the RTA condition, the experimental procedure started, once again, with the questionnaire on personal details and prior knowledge. As in the other two conditions, the participants were then given the UBU 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 had to fill in the first part of the second questionnaire, containing questions on their method of working. They were then shown a recording of their performance on video and asked to comment on the process retrospectively. Finally, they were given the second part of the questionnaire, containing questions on how they had experienced thinking aloud retrospectively.

2.6. Processing of the data

Once the 60 sessions were completed, verbal transcripts were made of the team, concurrent and retrospective think-aloud comments, and all the participants' navigations through the catalogue were noted down. A subsequent step was to study the participants' navigation and other actions with a view to detecting usability problems that had arisen in the process of using the UBU. As a rule, a particular situation was marked as a problem when it deviated from the optimum working procedure for each task. As for the think-aloud protocols, these were scanned for verbal indicators of problems experienced, expressing, for instance, doubt, task difficulty, incomprehensibility, or annoyance relating to the use of the catalogue.

Our analysis of the data collected focused on three main issues. First, we examined the total number of usability problems that was detected in each condition. Then we classified all problems according to the way they 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, we also categorised all detected problems according to problem type, on the basis of the categorisation that we used in our previous study. We distinguished four problem types:

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 interact with the search system (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

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

To measure the relevance of the problems detected, five experts rated all individual problems on a five-point scale ranging from least to most relevant. By relevance we mean the degree to which solving a problem will enhance the usability of the UBU catalogue. We not only considered the relevance with respect to the three conditions, but also with respect to the manner of detecting, the type of problems detected, and the frequency of the problems.

To evaluate task performance in all three conditions, we used two indicators: tasks completed successfully and time required to complete the tasks. These indicators were applied both per task and for the overall performance of the five tasks.

3. Results

In the following section we will first present our results regarding the feedback (number and types of problems) collected with the three usability test approaches (Section 3.1). We will then discuss the relevance of the problems detected (Section 3.2) and the results in terms of task performance (Section 3.3). Following that we will describe how the participants experienced their participation in the study (Section 3.4).

3.1. Number and types of problems detected

After analysing the 60 recordings, we found a total number of 85 different problems. We will first discuss this output by comparing the mean number of problems and problem types detected per session in each condition. Following that, we will briefly consider the number of different problems detected in each condition and the overlap that exists between them.

Table 1 gives an overview of the mean number of problems detected per session. It classifies all problems according to the way in which they surfaced: (1) by observation, (2) by verbalisation, or (3) by a combination of observation and verbalisation. As the table shows, there was no significant difference in the total number of problems detected by the three usability test approaches ($F(2,57)=0.641$, $p=0.531$). On a global level, they were comparable in terms of their quantitative output.

As for the way in which this output came about, there were only significant differences between the CTA and RTA protocols. As is clear from Table 1, the participants in the CTA condition experienced significantly more observable problems than the participants in the RTA condition ($F(2,57)=7.125$, $p<0.005$; Bonferroni post hoc analysis $p<0.005$). This result is in line with the results of our previous study, which also showed a significantly larger number of observable problems in the CTA condition. Once again, this finding supports the notion of reactivity: as the CTA participants had a double workload (performing tasks and thinking aloud at the same time), it seems likely that the extra task of thinking aloud caused them to experience more visible problems.

Apart from differing in terms of observable problems, the CTA and RTA conditions also differed significantly with respect to the number of verbalised problems that was detected in each condition: the CTA participants verbalised fewer problems than the RTA participants ($F(2,57)=3.162$, $p<0.05$; Bonferroni post hoc analysis $p<0.05$). Compared to our previous experiment, where the CTA participants verbalised only 0.5 problems and the RTA participants as many as 4.5, the current

Table 1
Number of problems detected per session in the CTA, TEAM and RTA condition, classified according to the way in which they were detected

	CTA		TEAM		RTA	
	Mean	SD	Mean	SD	Mean	SD
Observed	5.5*	2.5	4.1	1.9	3.1*	1.7
Verbalised	1.7**	2.1	2.7	1.8	3.4**	2.3
Observed and verbalised	2.5	1.6	3.8	2.2	3.4	1.6
Total	9.7	2.2	10.5	2.5	9.9	2.5

$*p<0.005$; $**p<0.05$.

difference is smaller and only slightly significant, but both results point in the same direction and can be explained by the different workload of the participants in both conditions: the CTA participants, having to perform tasks and think aloud simultaneously, had less time to verbalise problems that were not directly task-related. The RTA participants, on the other hand, needed to perform only one task at a time, which gave them more opportunity to voice problems, both task-related and non-task-related.

The results of the TEAM condition interestingly show no significant differences, neither compared to the CTA condition nor compared to the RTA condition. The total number of problems that was detected in the TEAM condition was slightly larger than in the other two conditions, but both the number of observable problems and the number of verbalised problems in the TEAM condition were larger than the respective problems in the CTA condition but smaller than the respective problems in the RTA condition. As such, the participants in the TEAM condition seem to have performed rather averagely: the fact that they were working in teams did not cause them to detect a substantially larger or smaller number of problems.

One explanation for the fact that the TEAM condition did not result in a substantially *larger* number of problem detections could be that as there were two people involved in the TEAM condition, one of the two could in fact unintentionally solve a problem that the other might have before the latter was able to voice it. For instance, by casually describing a term during the conversation, a participant could help out his or her team mate, so that the latter need no longer verbally indicate that the term was problematic to him or her. In that way, the ‘problem’ of the second team mate would not be recorded as a problem in the TEAM condition, but might have been labelled as such in the other two conditions, where the team mate would have worked individually.

An explanation for the fact that the TEAM condition did not result in a substantially *smaller* number of problem detections could be formulated along the same lines: as there were two people involved in the TEAM condition, they could both suggest possible ways

Table 2
Examples of problem types detected in the usability test approaches

Layout	The participant cannot find the link to the help function The participant cannot find the button to go to the next page
Terminology	The participant does not know what the term ‘signature’ means The participant does not understand the description of the term ‘truncation’ in the help function
Data entry	The participant does not know how to use the Boolean operators The participant does not know how to use the filter function
Comprehensiveness	The participant feels that the help function is not sufficiently comprehensive The results pages do not mention which page is currently being viewed

of performing the five tasks, which gave them more opportunity to experience and voice problems during their task performance.

To investigate the types of problems that were detected in the three conditions, we labelled all problems according to the problem types that we described in Section 2.6. Table 2 shows a selection of problems as they occurred in the usability test approaches.

Table 3 shows the overall distribution of problem types in CTA, TEAM and RTA. As in our previous experiment, all participants clearly experienced most difficulties in entering data and understanding terminology. The results for the other problem types were quite similar across the three conditions too, with only two significant differences, one between CTA and RTA, and one between TEAM and RTA. The CTA and RTA conditions differed with respect to comprehensiveness ($F(2,57)=4.087$, $p<0.05$; Bonferroni post hoc analysis $p<0.05$), but this difference concerned only a very small category of problems (0 as opposed to 0.7). The TEAM and RTA condition differed with respect to terminology ($F(2,57)=3.338$, $p<0.05$; Bonferroni post hoc analysis $p<0.05$); this difference, however, was only slightly significant. As such, it seems that as in our previous experiment, the three conditions largely reveal similar types of problems in similar frequencies.

So far, we established a number of significant differences between the three usability test approaches. While the three methods did not differ in terms of the

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

	CTA		TEAM		RTA	
	Mean	SD	Mean	SD	Mean	SD
Layout	1.1	0.9	1.4	0.8	1.3	1.1
Terminology	3.8	1.3	4.2*	2.4	2.8*	1.5
Data entry	4.8	1.2	4.7	1.7	5.0	1.5
Comprehensiveness	0*	0	0.3	0.9	0.7*	1.0

*$p<0.05$.

number of problems detected and hardly differed in terms of type of problems detected, the CTA and RTA condition did differ with respect to the *manner* of problem detection: the CTA participants detected significantly more problems by means of observation and fewer problems by means of verbalization than the participants in the RTA condition. These results are in line with our previous study, and a logical explanation for them would be the double workload of the CTA participants.

It stands to reason that, in evaluating a different catalogue, the workload of the participants in the current study is different from the workload of the participants in the previous study. As such, it would be interesting to see what effect a different workload has on the task performance of the participants. This task performance will be discussed in Section 3.3; we will now briefly consider the number of different problems detected in each condition (i.e. the list of individual problems regardless of how many times they were detected) and the overlap between them.

In the CTA condition, 46 different usability problems were detected; the RTA revealed 59 different problems; in the TEAM condition, 64 different problems came to light. This means that with respect to the range of individual problems detected, the RTA and TEAM methods are more profitable than the CTA method. With respect to overlap in the three lists of usability problems, there were 33 problems (39%) that occurred in each of the three conditions. The overlap between two rather than three conditions was considerably higher, ranging from 49 to 56%. If we take the frequency of the problems into account, the degree of overlap is even more substantial: problems that were detected in one condition by at least five participants were in 92–100% of the cases also detected by at least one participant in one of the other conditions. This means that each of the three methods can clearly predict the main output of the other two methods.

3.2. Relevance of the problems detected

As was mentioned in Section 2.6 above, five experts evaluated all 85 individual problems in terms of relevance, which we defined as the degree to which solving a particular problem would enhance the usability of the site. The experts were asked to rate each problem on a scale of 1–5 (least relevant to most relevant). Their scores formed an adequately reliable scale (Cronbach's $\alpha = 0.63$). With respect to the relevance of the problems detected, an analysis involving 95% confidence intervals showed that there were no significant differences between the three methods. In other words, all usability test approaches proved equally useful in detecting relevant problems. With an average score of 3.43, the experts felt that the problems detected were clearly relevant. Also with regard to the relevance of the problems detected exclusively in each condition, the three methods did not differ significantly.

As for the manner in which the problems were detected, there were again no significant differences between the three conditions: the CTA condition revealed equally relevant problems by means of observation as the TEAM and RTA condition; equally relevant problems by means of verbalisation; and equally relevant problems by means of both observation and verbalisation.

With respect to the relevance of the various problem types that were distinguished, the three conditions once again showed no significant differences. However, the experts generally regarded the layout problems as significantly less relevant than the data entry, terminology or comprehensiveness problems.

A final consideration involved the correlation between relevance and frequency of the usability problems, which did not differ between the three conditions. As a whole, there was a moderate correlation between relevance and frequency ($r=0.46$, $p<0.001$), which means that generally speaking the more frequently a problem was detected, the more relevant it was. However, this relation between frequency and relevance was not nearly as strong as is often suggested in handbooks on usability testing.

In sum, it seems that the three usability approaches did not differ with respect to the relevance of the problems that they detected.

3.3. Task performance

To measure task performance, we used two indicators in this study: the time it took the participants to complete the five tasks, and the degree to which they were successful in doing so. Table 4 presents the results of both indicators.

With regard to the overall task completion time there was no significant difference between the CTA participants and the TEAM participants or the CTA participants and the RTA participants. Apparently, the CTA participants did not work more slowly than the other participants as a result of having to think aloud, a result which is in line with the findings of our previous study. Only one significant difference was found: the participants in the TEAM condition took significantly more time (30.4 min) to complete the tasks than the RTA participants (26.1 min) (ANOVA, $F(2,57)=7.139$, $p<0.005$, Bonferroni post hoc analysis, $p<0.005$). It would seem that they spent their extra time on tackling the tasks in different ways (reflecting the fact that they were working together, making use of the input of not one but two people).

With regard to the successful completion of the five tasks no significant differences were found between the three conditions. The CTA participants performed their tasks as successfully as the participants in the other two conditions, a result that deviates from the result of our previous study, which showed the CTA participants being *less*

Table 4
Task performance in the CTA, TEAM and RTA condition

	CTA		TEAM		RTA	
	Mean	SD	Mean	SD	Mean	SD
Overall task completion time in minutes	26.1	6.9	30.4*	8.8	21.5*	6.5
Number of tasks completed successfully	1.6	0.8	2.2	0.7	1.7	0.9

*$p<0.005$.

successful in performing their tasks than the participants in the other condition. Apparently, the participants in the present study were only affected by reactivity in terms of the number of observable problems, and not, as the participants in the previous study, in terms of successful task completion.

A likely explanation for this difference would seem that the current CTA participants simply had less difficulty in performing their tasks, resulting in a lighter workload which would make them less vulnerable to reactivity caused by thinking aloud while working. Such an explanation does, however, not seem valid here, as a comparison of the task performance of the participants in both studies shows that the current participants were *not* more successful in completing their tasks than the participants in the previous study. Indeed, while the participants in the previous study managed to complete 2.6 out of 7 tasks (37%), the participants in the present study completed only 1.6 out of 5 tasks (32%).

A more plausible explanation for the fact that the reactivity did only affect the number of observable problems and not the number of tasks completed successfully is a result of the weak correlation between these two indicators ($r = -0.27, p < 0.05$): high numbers of observable problems did not automatically correspond to unsuccessful task completions, which means that reactivity on the one indicator does not necessarily cause reactivity on the other.

3.4. Participant experiences

The questionnaire on participant experiences served to establish how the participants in the three conditions had felt about participating in the study. Questions involved three aspects of the experiment: (a) Experiences with having to think aloud (concurrently or retrospectively) or working together; (b) Method of working; (c) Presence of the facilitator and the recording equipment. Since the participants in the TEAM condition were working in pairs, each with a different role (actor/observer) that may have affected their experiences, they will be treated as separate subgroups in the following analyses. TEAM Actor will be representing the actors (those working behind the computer); TEAM Co-actor will be representing the co-actors (those sitting next to the person working behind the computer).

To start with, all participants were asked how they had felt about having to think aloud (concurrently or retrospectively) or working together by indicating, on a five-point scale, to which degree they thought this activity was difficult, unpleasant, tiring, unnatural, and time-consuming. Together, these five variables formed a reliable scale (Cronbach's $\alpha = 0.76$), so they were grouped together as a new variable measuring how the participants had felt about having to think aloud or work together. ANOVA testing and Bonferroni post hoc analyses showed that there were significant differences between the conditions ($F(3,76) = 15.271$, $p < 0.001$): both the TEAM Actor participants and the TEAM Co-actor participants felt significantly more positive about working together than the participants in the CTA condition ($p < 0.001$ in either case) and the participants in the RTA condition ($p < 0.001$ in either case) did about having to think aloud (retrospectively). The scores of the participants in the CTA and RTA condition were rather neutral (3.0 and 2.9, respectively); the scores of the

Table 5
Participants' method of working, compared to their usual working procedure

	CTA		TEAM actor		TEAM co-actor		RTA	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Faster–slower	3.5	0.5	3.4	0.8	3.4	0.8	3.3	0.5
More–less focused	2.5	0.6	2.9	0.8	3.0	0.7	2.8	0.5
More–less concentrated	2.6	0.6	2.9	0.7	2.9	0.7	2.9	0.7
More–less persevering	2.7	0.9	2.9	0.9	2.8	0.9	2.8	0.6
More–less successful	3.0	0.6	3.1	0.6	3.0	0.5	3.0	0.4
More–less pleasant*	3.5	0.8	3.5	0.7	2.9	0.6	2.9	0.3
More–less eye for mistakes	2.6	0.5	3.0	0.6	2.5	0.8	2.7	0.7
More–less stressful	3.5	0.7	3.1	0.4	3.1	0.9	3.0	0.4

*TEAM Co-actor and RTA differ significantly from CTA and TEAM actor ($p < 0.05$). Scores on a five-point scale (3 = no difference from usual).

participants in the TEAM Actor and TEAM Co-actor condition were clearly positive (both 1.95). Apparently, the fact that the TEAM participants were working together added much to their overall positive evaluation of their participation in the experiment.

Participants were next 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 normally do. Results, which are shown in Table 5, showed that the participants in all three conditions felt that they had not worked all that differently from usual: the scores for all items are rather neutral, ranking around the middle of the scale. Only the item 'pleasantness' showed significant differences between the conditions: compared to their normal working procedure, both the CTA participants and the TEAM Actor participants felt that they had worked less pleasantly than participants in the RTA and TEAM Co-actor conditions ($p < 0.05$). In other words, the conditions in which participants had to think aloud or work together while handling the computer were experienced as least pleasant. This might be explained by the fact that these two conditions involved a heavier workload than the other two: in CTA, participants had to perform tasks and think aloud, while in TEAM Actor, participants had to perform tasks and consult their team mate. The participants in TEAM Co-actor, on the other hand, only had to consult with their team mate, while the participants in the RTA condition could perform their tasks without any talking at all.

The final part of the questionnaire included questions about 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. As the items of both aspects together formed a reliable scale (Cronbach's $\alpha = 0.73$), they were grouped together as a new variable measuring the effect of the experimental setting on the participants. ANOVA testing then showed that there were no significant differences between the conditions. With all scores ranking around the middle of the five-point scale (ranging from 2.3 to 2.5), the participants clearly felt that they were hardly affected by the experimental setting.

In sum, while the three usability test approaches showed similar results with regard to the effect of the experimental setting and the participants' working procedure, the TEAM condition was clearly evaluated most positively by the participants. This would seem to suggest that, given the choice, participants would rather work together than individually.

4. Discussion

The results of our study show that there are both differences and similarities between the three usability test approaches. With regard to the output of each method, there was no significant difference in terms of quantity and relevance of the problems detected. All three methods revealed similar numbers of problems that were largely distributed over the same problem types and that were equally relevant. The CTA and RTA did, however, differ as to how this output came about: while the CTA method resulted in significantly more problems detected by means of observation only, the RTA method proved more fruitful in revealing problems that could only be detected by means of verbalisation. This difference corresponds with the results of our previous study and may be explained by reactivity: as the CTA participants had to think aloud and perform tasks at the same time, their double workload caused them to experience more observable problems and to have less time for verbalising non-task related problems than the RTA participants.

With respect to task performance, the three usability approaches also proved roughly similar. The participants in the TEAM condition took significantly more time to complete their tasks than the RTA participants, but there was no difference between the three methods in terms of successful task completion. An important implication here is that the participants in the CTA condition did not perform worse due to the fact that they had to think aloud and carry out tasks at the same time. Apparently, the notion of reactivity, which we just mentioned as an explanation for the larger number of observable problems in the CTA condition, had no effect on the participants' task performance. This result does not confirm the result of our previous study, which saw the CTA participants performing worse on tasks that were less difficult than the tasks in the present study. It seems that there is no apparent link between task difficulty and reactivity, but as we suggested in our previous study, more research is necessary to investigate this relationship.

With regard to the participant experiences, there was only one significant difference, with the constructive interaction method being evaluated more positively by the participants than concurrent or retrospective thinking aloud. Apparently, working together is regarded as an agreeable way of participating in a usability test. In the present study, most of the TEAM participants knew their partner to some extent, which might make it more enjoyable for them to work together. Nevertheless, it is easy to see why working together would be the preferred option for most participants, regardless of whether they know their team mate or not: participants can share their workload and they can talk to each other in a much more natural way than if they were required to think aloud concurrently or retrospectively.

From a practical point of view, it seems that, based on the results of this study, a strong case can be made for the concurrent think-aloud method. The output of this method is similar to the other two conditions in terms of quantity and relevance, but CTA has some practical advantages in terms of time and costs. The method is much less time-consuming than the retrospective think-aloud protocols, which require double the test time to allow the participants to watch and comment on their recorded performance. At the same time, CTA is much cheaper than the TEAM condition, which requires twice as many participants that (typically) need to be awarded for their participation.

All the same, it is too early to draw any firm conclusions about the value of the CTA method as yet. If we recall the numbers of individual problems detected as well as the number of unique problems per condition, it has to be said that the CTA method was less productive than the other two usability test approaches. Moreover, the method has one other potential drawback that needs to be further investigated. As we have seen, CTA revealed a significantly larger number of observable problems than the other two conditions. Naturally, these problems could all prove valid, in which case the CTA method could be regarded as particularly useful for testing artefacts that involve many visual steps. However, it is also possible that (part of) these observable problems in the CTA condition were caused by the method itself, with the task to think aloud causing the participants to experience problems that were in fact 'false alarms'. Such a possibility deserves to be the focus of future research.

References

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

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

Bowers, V.A., Snyder, H.L., 1990. Concurrent versus retrospective verbal protocols for comparing window usability, Proceedings of the Human Factors Society 34th Meeting, HFES, Santa Monica, CA, pp. 1270–1274.

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

De Jong, M., 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 (3), 242–260.

Dumas, J.S., Redish, J.C., 1999. A practical guide to usability testing (revised edition). Intellect, Exeter.

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

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

Hackman, G.S., Biers, D.W., 1992. Team usability testing: are two heads better than one? Proceedings of the Human Factors Society 36th Annual Meeting, HFES, Santa Monica, CA, pp. 1205–1209.

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

Höysniemi, J., Hämäläinen, P., Turkki, L., 2003. Using peer tutoring in evaluating the usability of a physically interactive computer game with children. Interacting with Computers 15, 203–225.

Kahler, H., 2000. Constructive interaction and collaborative work: introducing a method for testing collaborative systems. ACM Interactions 7 (3), 27–34.

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

Miyake, N., 1982. Constructive Interaction. Center for Human Information Processing. University of California, San Diego, CA.

Molich, R., Ede, M.R., Kaasgaard, K., Karyukin, B., 2004. Comparative usability evaluation. Behaviour and Information Technology 35 (1), 65–74.

Nielsen, J., 1993. Usability Engineering. Academic Press, Boston, MA.

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., Winters, C.M., 2002. Usability Testing for Library Websites: A Hands-on Guide. American Library Association, Chicago.

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

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

Van den Haak, M.J., De Jong, M.D.T., Schellens, P.J. 2003. Retrospective versus concurrent think-aloud protocols: Testing the usability of an online library catalogue. Behaviour & Information Technology, 22, 339–351.

Van Oostendorp, H., De Mul, S., 1999. Learning by exploration: thinking aloud while exploring an information system. Instructional Science 27, 269–284.

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